Valid Measurement of Laboratory Learning Experience Quality from the Student Perspective

Samuel James Priest
Department of Chemistry, School of Physical Sciences

July 2016

THE UNIVERSITY of ADELAIDE

Contents

Abstract

Since the early 2000s, improvement of the student learning experience in university level laboratory activities in Australia has been sought by the Advancing Science by Enhancing Learning in the Laboratory (ASELL) project. The nation-wide project has made use of the ASELL Student Learning Experience (ASLE) survey to gather data and draw conclusions regarding student perspectives of their learning experiences, using trends observed in the data to inform pedagogy. Analyses of rating scale response format items on the ASLE survey have typically involved an integer value scoring system applied to the response categories. The appropriateness of such integer scoring techniques and the subsequent application of parametric statistical methods to ordered categorical data in this way is contested in statistical literature, which raises questions regarding the validity of ASELL project conclusions drawn in the past.

In this thesis, Rasch measurement is applied to a data set of ASLE survey responses, using the true interval scale measures gained to test the validity of the scoring techniques and parametric methods more typically applied to ASLE data. The role of student biases in survey response and 'objectivity' of any measures associated with learning experience quality are explored, yielding quantitative models of the student perception of laboratory learning experiences. The thesis culminates in the use of factor analysis to develop a Linear Logistic Test Model for a data set of over 9000 completed ASLE surveys, explaining the responses received as linear combinations of a small number of major factors in the student laboratory learning experience. The model is used to draw pedagogical conclusions from the ASLE survey data set uninfluenced by limitations of the integer scoring techniques usually applied.

The work has major implications for valid interpretation of ASLE survey data received both in the past and in future, suggesting that whilst integer scoring methods may be amenable to parametric statistics, the conflation of student dependent and student independent factors limits the generality of any conclusions drawn. Student independent measures obtained from Rasch analysis, however, reveal that the perceived relative quality of a laboratory exercise is largely consistent through the student population sampled. The Linear Logistic Test Model generated reveals a wide range of connections between different facets of the laboratory learning experience and this general perceived learning experience quality, informing effective science pedagogy. Pedagogical conclusions include strong connections between group work and understanding of theoretical content, the advantages of data analysis and individual work in development of more technical or practical skills, evidence for the importance of structuring activities appropriate to the ability level of the students, as well as ways to generate student interest and foster perceptions of a positive overall laboratory learning experience. A need for compromise between teaching objectives and learner preferences is highlighted, noting that the "best" way to design a laboratory activity largely depends on the intended purpose of the exercise.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

_____

Acknowledgements

First and foremost, I wish to thank my supervisors Simon Pyke, Natalie Williamson and John Willison for their continued and valued support throughout my writing of this thesis. I greatly appreciate their willingness to allow me to exercise freedom and creativity in my research, even following paths that are relatively unfamiliar for all concerned, whilst still being able to provide constructive feedback and guidance along the way.

I would also like to thank the many practical demonstrators who have assisted in the distribution of "my horrible surveys" over the course of numerous years, without whom I could never have gathered the wide data set which has been integral to the strength of this work's conclusions. I wish to particularly thank Annelie Karssen, who took over the collection of survey data whilst I was away briefly during 2015. Data gathered during that time has been some of the most useful for my final and most interesting conclusions, and I am grateful for her help. Laboratory technicians Peter Roberts and later Catherine Margach have also been of great help in accommodating my presence in the first year chemistry laboratory, and I am very grateful to them.

Another person deserving of particular thanks is Lyron Winderbaum, who readily offered his assistance in the dreaded art of MATLAB coding when I was in need. Without his help, I'm sure I would never have had it working. Continuing the mathematical theme, I wish to give thanks to Sivakumar Alagumalai, who first introduced me to Rasch analysis and kindly spared his time to provide me some advice in the earliest stages of my research.

I also wish to thank my friends and family for their continual support and kindness throughout my studies. My friends for giving me much needed distractions from my working life as well as sharing my dread of the editing process, and my family for continuing to support and encourage my continued study, even as a few expected deadlines sailed past.

Lastly, I would like to thank all of the first year chemistry students who have passed through the laboratory over the few years I have been present. Without them, my entire project would have been impossible and my working days would never have been so enjoyable.

1 *Introduction*

1.1 The Advancing Science by Enhancing Learning in the Laboratory (ASELL) project

1.1.1 Laboratory work in science education and history of the ASELL project

For more than a century,[1] the laboratory has served as a key component of university level science education. Despite queries as to whether benefits of laboratory sessions outweigh the costs,[2, 3] suggested key roles of laboratory activity in science education have persisted.[4-8] Development of hands-on practical skills, development of scientific and critical thinking skills, supporting learning of the subject matter as well as fostering more generic skills such as time management and effective work in teams are all frequent suggested benefits of laboratory activities.[9] These claims have been largely supported to varying extents, with more authentic research activities in undergraduate science additionally shown to act as powerful affective, behavioural and personal discovery experiences.[10]

Laboratories have historically been viewed as providing the opportunity for a strong inquiry-based environment, where inquiry is an integral part of the scientific process.[11, 12] Using laboratories as a tool to engage students with scientific concepts at a concrete, macroscopic level,[2, 13] students are allowed to forge connections between real world experimental observations and underpinning scientific concepts[14] whilst strengthening their scientific reasoning and broader grasp of how the scientific process works.[15] Developing skills in scientific writing has also been viewed as heavily dependent on laboratory work for this reason, since only in the laboratory are students able to "do" science rather than merely "learn about" science.[16]

Stimulation of student engagement and interest in science is also a key function of laboratory activities, notably since student interest levels have a powerful influence on learning.[17, 18] The appeal of laboratory experiments has previously been identified as one of the most prominent reasons for initial enrolment in chemistry,[19] justifying observations that positive laboratory experiences help encourage student retention in chemistry (and other science) majors.[20, 21] Despite this, difficulties in effectively implementing inquiry-based laboratories have meant that more expository "cook-book" laboratories[22] are prevalent, which miss out on many desirable (and potentially enjoyable) features of the experience.[23]

During the late 1990s, anecdotal evidence presented by academics attending research conferences around Australia suggested that increasing numbers of students were finding their physical chemistry laboratory sessions neither interesting nor motivating. These observations, coupled with associated decreases in enrolment and retention in physical chemistry courses, prompted the Committee for University Teaching and Staff Development (CUTSD) at the time to fund the Australian Physical Chemistry Enhanced Laboratory Learning (APCELL) project, which aimed to address the issue.[24]

Adopting a "research-led teaching" philosophy, the APCELL plan was initially formulated by researching the relevant education literature on laboratory learning and teaching, inquiring into the nature of the barriers to improvement, and gathering data on students' perceptions of their teaching and learning experiences. Data gathered on students' perceptions of their teaching and learning experiences in laboratories suggested that laboratory sessions could be perceived to lack relevance and be little help in achieving useful learning outcomes, reminiscent of other findings on effective laboratory teaching at the time.[25, 26] Given the pre-existence of education literature on effective laboratory teaching, the question arose as to

why the relevant recommendations had not already been adopted. It became apparent that despite attempts by individual institutions to improve their practice, each faced a variety of resource constraints which impeded their progress.[27] Individual institutions' limited access to physical resources (such as equipment), specialist expertise, pedagogical expertise and active student involvement meant that a multi-institution approach was necessary. The APCELL project therefore gained participation from multiple Australian universities, pooling the resources of individual institutions to collectively improve practice. The project's overarching objective to "*measurably improve the learning, motivation and enjoyment of chemistry laboratory experiences by students*"[28] was henceforth pursued in three ways:

1. Establishment of a network of physical chemistry educators and students to share expertise in on-going curriculum development
2. Development of a suite of physical chemistry experiments, based on sound pedagogy, that could facilitate improved student learning
3. Creation of an internet database including all of these experiments, complete with associated demonstrator notes and other documentation required to deliver the laboratory activity to students

Because of its student-centred philosophy and intended outcomes, the project sought to develop a template for considering existing laboratory teaching practices from a learner-focused perspective. At a workshop held in Canberra during July of 2000, academics from over 30 participating institutions were asked to reflect on and challenge their existing ideas and conceptions of teaching, addressing the issue at the level of their underlying ideas about teaching and learning rather than at the level of their teaching behaviours. The result was the refinement of the 'Educational Template' document,[24] which became central to the APCELL project and its successors. This template was designed to accompany experiments submitted to the APCELL review process, for potential inclusion in the online APCELL database of pedagogically sound experiments. The template was not designed to prescribe practice, but instead designed to promote consideration of existing practices from a learner perspective.

Following its review and subsequent amendments, the Educational Template document included several sections to be completed by the submitter as part of the experiment evaluation process. These sections included information on the context in which the experiment is run, the educational goals of the activity, how those objectives could be met and how both students and teachers could recognise they had been met. The template document was also designed to include an analysis of feedback provided by students who had conducted the experiment, in keeping with the learner-centred focus of the project. The student feedback contained within this final section of the completed template was to be gathered during hands on laboratory sessions, held during APCELL workshops.

Held in Sydney in February of 2001, the first APCELL workshop was principally designed around the "peer review" of experiments submitted to APCELL, largely scaffolded by the Educational Template. Both students and academics were provided the facilities to physically conduct the laboratory activities submitted to APCELL, providing feedback on their experiences. Submitters of the experiment acted as the "demonstrator" for each laboratory activity. The laboratory sessions held as part of the workshop not only gathered data on learner perspectives of each experiment, but additionally it placed academics in the 'shoes' of the students, opening up a dialogue between teachers and learners. Workshop delegates were also provided time to discuss the submitted experiments at length, evaluating strengths and weaknesses of the task

design, potential improvements which could be made or amendments to the completed Educational Template document submitted along with the experiment.

Attendees at the first APCELL workshop responded very positively to the experience, reporting their heightened awareness of what the student experience constitutes and the issues in running an effective laboratory learning exercise from the student perspective. The value and necessity of evaluating experiments in a hands-on, interactive environment in this way was acknowledged strongly by the workshop delegates, and it was agreed that all experiments submitted to APCELL should be put through an extensive and rigorous review, incorporating this process.

Additional APCELL workshops were organised subsequent to the workshop in Sydney, largely to facilitate the review of large numbers of submitted experiments in bulk whilst also reinforcing the student-centred learning concept with participating academics. Because APCELL experiments now had a structured 'peer review' process associated with their evaluation, submitters could elect to pursue publication of the experiment details and evaluation results in a peer-reviewed education journal subsequent to their acceptance into the APCELL database. A partnership was established with the Australian Journal of Education in Chemistry,[29] which published numerous APCELL experiments and their evaluations.[30-41]

Following the success of APCELL, the project was expanded in 2006 to involve chemistry more broadly and relabelled as ACELL: The Australian Chemistry Enhanced Laboratory Learning project.[42-46] Over the course of its development, the review process for experiments submitted to A(P)CELL had evolved to incorporate evaluation at the submitter's home institution, gathering survey feedback from students completing the experiment as part of a course in which they are enrolled. A standard recommended procedure for analysing the survey feedback was developed,[47] and this data (both qualitative and quantitative) could be presented in the published form of the evaluated experiment. This process was incorporated into the ACELL experiment review scheme, as outlined in Figure 1.

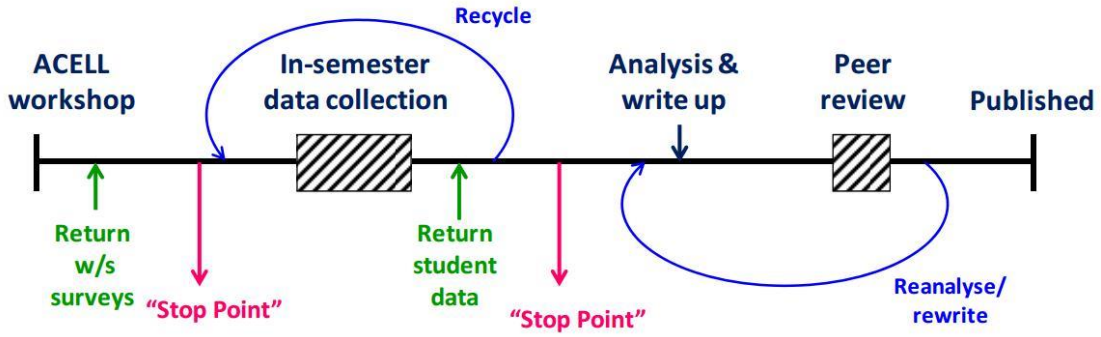


Figure 1: Experiment review process

Experiments submitted to the review process undergo a workshop review by both students and academics, an in-semester review by the student cohort of the home institution and a peer review process prior to publication. Stop points exist between these phases where the experiment may be modified and re-evaluated based on feedback. Image is reproduced with permission from Pyke *et al*.[48]

This scheme has been utilised for the evaluation of a number of experiments, some of which have been published in peer reviewed journals alongside their associated student feedback at both the workshop and at the home institution, as well as their evaluation using the Educational Template document.[49-51]

ACELL's success continued, with numerous workshops run across Australia. Following calls for even further expansion into other disciplines of science[46, 52] the project eventually evolved into the current Advancing Science by Enhancing Learning in the Laboratory (ASELL) project,[48, 53, 54] which has now seen involvement from outside of Australia[54] and is beginning to also encompass education in schools.[55-57] The increased volume of data generated from workshop evaluations and home institution evaluations also allowed the project to begin contributing to education research more broadly, using survey response data to investigate large scale trends in laboratory learning experiences.[58, 59] To date, ASELL and its predecessors have gathered data from over 120 experiments, with contribution from over 25,000 students, 350 academics and 30 deans affiliated with 28 universities across Australia.[58] Additional unknown volumes of data gathered using ASELL surveys across a variety of institutions also exist separately to the ASELL project database, such as data presented within this thesis.

The project has maintained its experiment review structure (Figure 1), and still presents accepted experiments (both from current and past forms of the project) in the ASELL online database, available on what is now the ASELL website.[56] Experiments may also still be evaluated using surveys designed by ASELL at home institutions, where students conducting the experiment may provide feedback. Data gathered using these surveys, much like data previously gathered by ACELL and APCELL, has also been used to contribute to laboratory learning education research beyond the experiment review process, in both comparative[60] and correlational studies.[58, 61]

1.1.2 Data analysis and interpretation

Whilst the experiment review process used by ASELL and its predecessors involves multiple stages, this thesis will primarily have its focus on survey data received during workshops and most notably home institution analyses. The ASELL Student Learning Evaluation (ASLE) survey instrument, used for home institution analyses, was designed with the intent of providing academics with a simple, easy to analyse tool for capturing key elements of the student experience during laboratory activities. The survey is comprised of fourteen Likert-type (rating scale) response format items, each of which allows students to respond in one of five ordered response categories, as well as five open response format items (Table 1). Survey items were initially formulated based on recurring themes evident in original open response comments gathered during early APCELL workshops,[61] as well as education literature regarding benefits of inquiry based laboratory exercises[62] and teamwork.[23] The precise phrasing of the questions included on ASELL project surveys has continually evolved with the project; however the content of the questions has remained generally equivalent. The rating scale or 'Likert-style' response format items are typically used for quantitative analysis, whilst open response format items are used for more qualitative purposes.

Responses to ASLE surveys (used for home institution analyses) and to analogous surveys used in ASELL workshops are used to gauge the relative quality of the experiment from the student perspective. Likert-type data in particular may be more easily subjected to statistical comparisons, revealing any differences between evaluations of different experiments or between separate evaluations of the same experiment in different contexts. Often these investigations are treated as exploratory in nature, due to small sample sizes and hence poor generality of conclusions and little statistical power. This limitation is especially an issue for the workshop surveys, the responses for which are limited by small numbers of workshop participants. Studies capable of drawing more generalised and less informal conclusions

require large numbers of responses, typically available only from in-semester survey collection, or from the full collated data set of many evaluated experiments.

Table 1: Items included in the ASELL Student Learning Experience (ASLE) survey

#	Item	Response Format
1	This experiment helped me to develop my data interpretation skills	Likert - style
2	This experiment helped me to develop my laboratory skills	Likert - style
3	I found this to be an interesting experiment	Likert - style
4	It was clear to me how this laboratory exercise would be assessed	Likert - style
5	It was clear to me what I was expected to learn from completing this experiment	Likert - style
6	Completing this experiment has increased my understanding of chemistry	Likert - style
7	Sufficient background information, of an appropriate standard, is provided in the introduction	Likert - style
8	The demonstrators offered effective supervision and guidance	Likert - style
9	The experimental procedure was clearly explained in the lab manual or notes	Likert - style
10	I can see the relevance of this experiment to my chemistry studies	Likert - style
11	Working in a team to complete this experiment was beneficial	Likert - style
12	The experiment provided me with the opportunity to take responsibility for my own learning	Likert - style
13	I found the time available to complete this experiment was	Likert - style
14	Overall, as a learning experience, I would rate this experiment as	Likert - style
15	Did you enjoy doing the experiment? Why or why not?	Open
16	What did you think was the main lesson to be learnt from the experiment?	Open
17	What aspects of the experiment did you find most enjoyable and interesting?	Open
18	What aspects of the experiment need improvement and what changes would you suggest?	Open
19	Please provide any additional comments on this experiment here.	Open

Response categories:
Items 1-12: "Strongly Agree", "Agree", "Neutral", Disagree", "Strongly Disagree"
Item 13: "Way Too Much", "Too Much", "About Right", "Not Enough", "Nowhere Near Enough"
Item 14: "Excellent", "Good", "Average", "Poor", "Very Poor"

The workshop survey for experiment evaluation and the ASLE survey are similar in their content, despite some differences in the questions posed. For this reason, responses obtained from both surveys yield comparable information and are typically treated using identical analysis strategies. The ASLE survey exclusively represents the views of students at the home institution, whilst the workshop survey is completed by ASELL workshop delegates, predominantly inclusive of academics. Information such as this may be used in conjunction with open response comments received to gauge the quality of an experiment from the student perspective, identifying areas of possible strength or weakness. Work conducted by Crisp *et al*.[51] in evaluating student and staff perceptions of the same experiment illustrates the complementarity of the workshop survey and the ASLE survey, whilst also showcasing the way in which Likert-type data may be used as an indicator of participant perceptions (Figure 2, page 7).

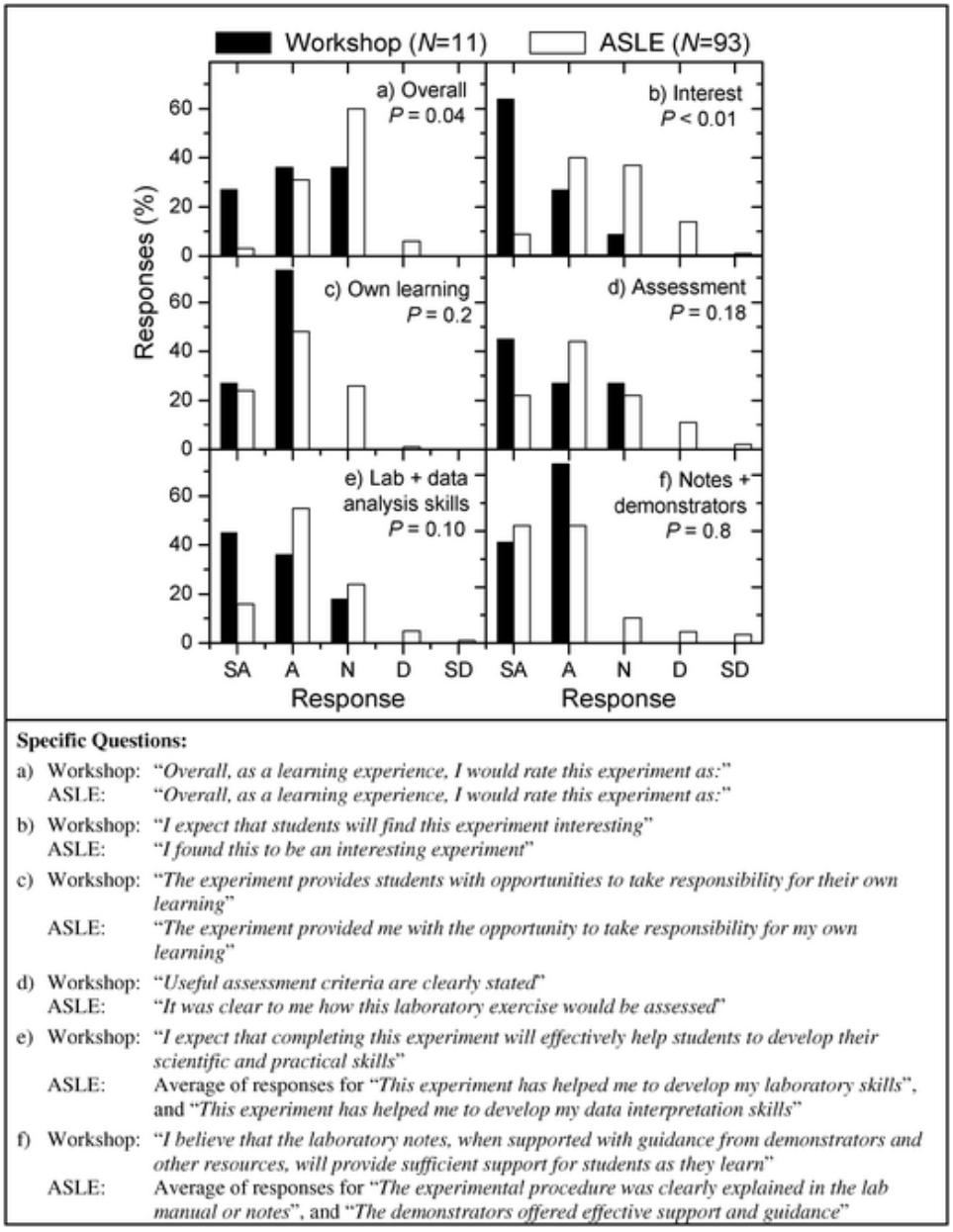


Figure 2: Exploration and contrast of perceptions using Likert-type response data obtained using the workshop survey and ASLE survey

The frequency of responses observed in each of the rating scale categories gives a general insight into perceived quality of the experiment. In this figure, reproduced with permission from Crisp *et al.*[51], the workshop survey is used to illustrate the perceptions of academics, whilst the ASLE survey is used to gauge the perspectives of students.

Over time, a standard technique for the analysis of survey responses has been developed with the project. For open response items, a procedure of categorising comments based on their content and enumerating the number of comments which fall into each category is employed[47] (Figure 3). For Likert-type items, a scoring system is implemented whereby each of the five successive response categories are assigned successive integer values. As an example, response categories "strongly disagree", "disagree", "neutral", "agree" and "strongly agree" (available for items 1 – 12) would be assigned scores of -2, -1, +0, +1 and +2 respectively. Response options for item 13 and for item 14 are treated similarly. The average scored response may then be reported for each item (Figure 4).

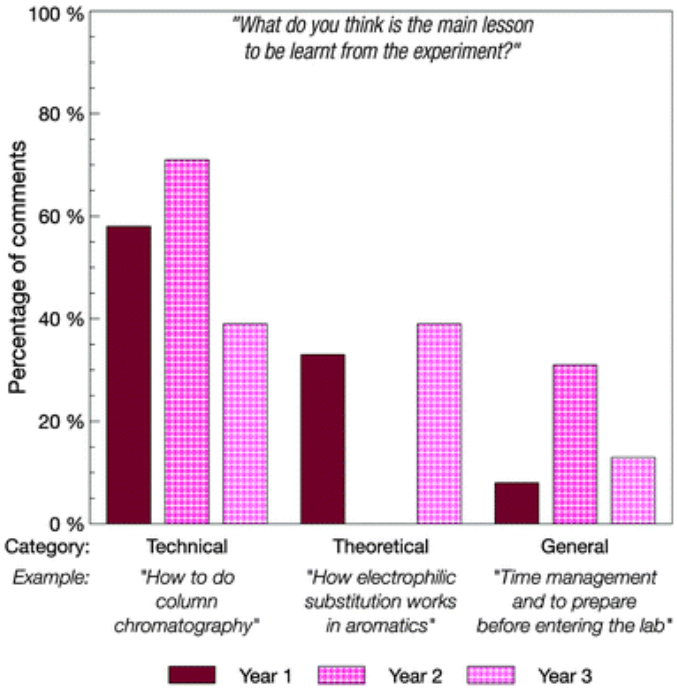


Figure 3: An example of the recommend analysis of ASLE survey open response items

Open response comments have been classified into researcher-defined categories based on their content. Frequencies of comments in each category have then been enumerated for the purposes of comparisons (in this case, between three iterations of the same experiment run in three different years). Image has been reproduced with permission from Southam *et al*.[60]

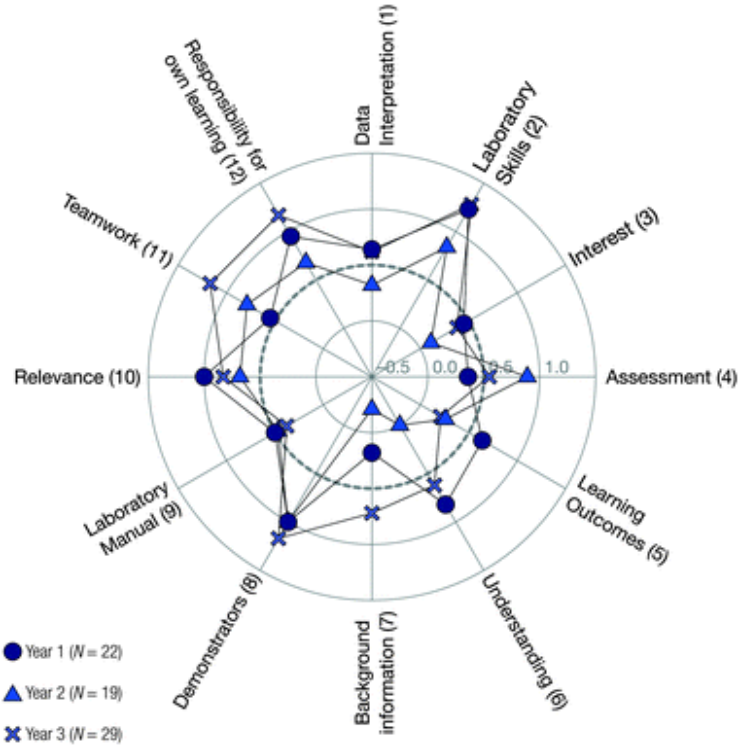


Figure 4: An example of typical 'scoring' of ASLE survey Likert-type response format items

Successive integer value scores have been assigned to individual Likert-type item responses, enabling mean scores for each survey item to be tallied and used for comparisons (ASLE items 1 – 12 are shown). In this case, student perceptions have been contrasted between different years the same experiment was run. Image has been reproduced with permission from Southam *et al*.[60]

Similar techniques have also been applied with other surveys utilised over the course of the existence of the ASELL project and its predecessors. Other surveys used at ASELL workshops to gauge the perceptions of participants have been subjected to similar treatments, contrasting perceptions between staff and students as well as comparing responses between different workshops. Yeung *et al*[53] extensively applied both techniques described above in evaluating the success of the first ASELL workshop, identifying key themes in qualitative comments received and applying the same integer scoring methodology to survey responses, yielding mean scores as a measure of perception.

However, Yeung *et al.* take the integer scoring technique a step further than previous ASELL based studies by using the calculated mean scores in parametric statistical tests. In the study, responses to individual Likert-type items are assigned successive integer scores in the usual way, then used to calculate both mean scores and standard deviations. These values, in conjunction with sample sizes, were used multiple times to conduct both Student's t-test and ANOVA; practices generally restricted to interval scale data rather than ordered categorical data such as rating scale responses. The distribution of student responses shown in Figure 5 below was characterised as having a mean score of +1.52 ($\sigma = 0.57$), concluded not to significantly differ from the distribution of staff responses, characterised by mean score of +1.33 ($\sigma = 0.76$), using Student's t-test ($t = 1.21$, $df = 77$, $p = 0.231$).

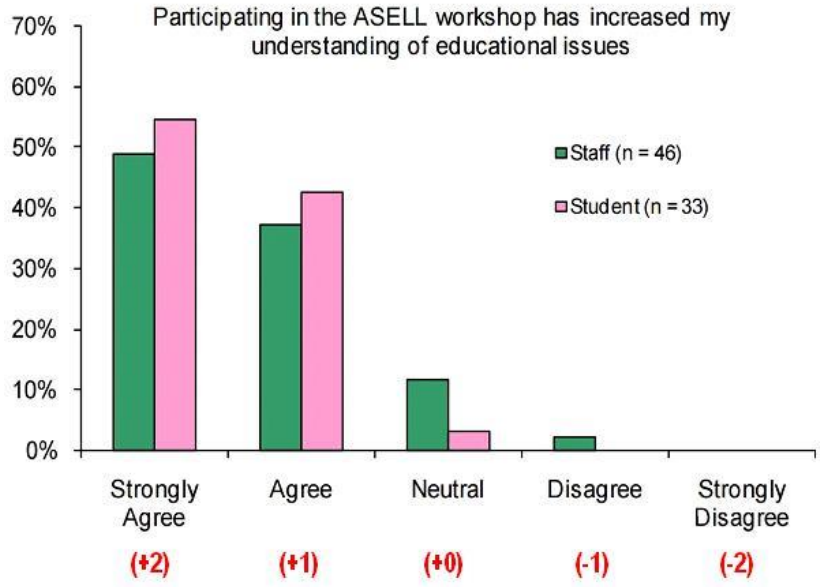


Figure 5: Assignment of scores for the purpose of statistical testing.

Image is reproduced with permission from Yeung *Et al*,[53] with the addition of scores associated with each response category (shown in red in parentheses).

Standard deviations had been reported alongside mean ASELL scores previously as a rough measure of the spread of responses (see the characterisation of the "thermodynamics think-in" experiment by Kable and Read,[50] for an example). However, they had never before been used to explicitly quantify probabilities associated with statistical tests. Procedures such as this represent a shift from interpreting response scores as a rough indicator of perception to interpreting them as a quantified, interval scale measures fit for parametric statistical comparisons. Other ASELL papers have chosen not to use scores in statistical tests, instead resorting to non-parametric methods such as the Wilcoxon rank sum test, as is more usual for the analysis of ordered categorical data.

Whilst ASELL survey results have commonly been used for the purposes of evaluating individual experiments, student perception data in the form of mean scores have also been used to probe the laboratory learning experience more deeply. Large volumes of data are available from the multitude of workshop evaluations and home institution evaluations conducted over the course of the ASELL project and its predecessors' existence. This has enabled more reliable and generalizable statistical conclusions to be drawn, meaning ASELL survey data has emerged as a tool to investigate more fundamental questions about generating a positive student experience in the laboratory.

A notable example of the use of larger volumes of ASELL data is the correlation of scored responses received for one survey item against the scores received for another, in pursuit of identifying factors contributing to a positive laboratory experience. The role of 'interest' in generating a positive 'overall learning experience' was exemplified by George *et al*.[59] in this way, using both mean scores and scored individual responses to the 'interest' and 'overall learning experience' Likert-type items from a large number of ASELL evaluated experiments (Figure 6). Similar correlations were evaluated between the "overall" item and the other Likert-type response format items of the survey, yielding correlations between overall experience and items 1, 3, 6 and 12 of the ASLE survey.

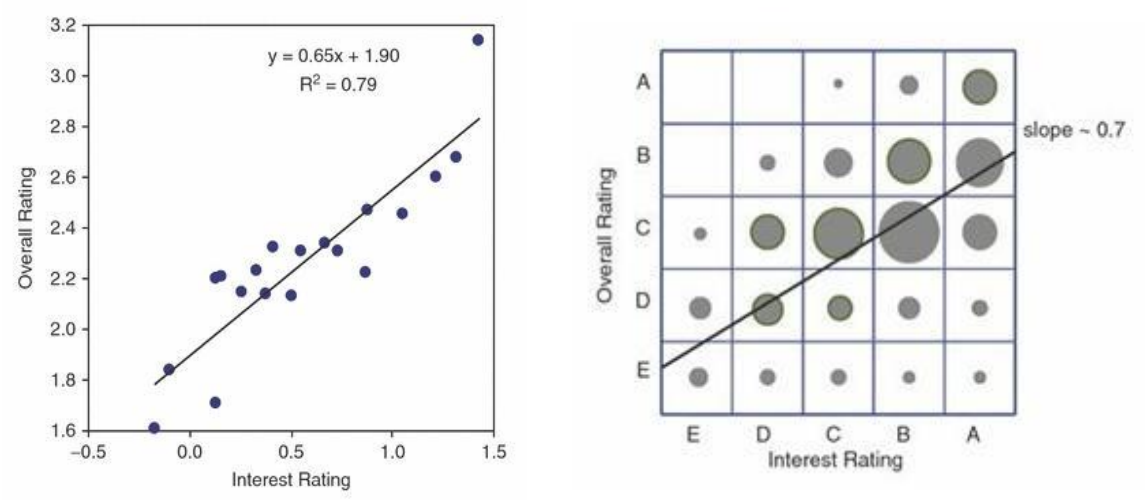


Figure 6: Correlated scored ASLE survey responses for item 14 (overall) and item 3 (interest)

LEFT: The mean values of scored responses have been correlated, with each data point representing values from a different experiment in the ASELL database. RIGHT: Individual responses received for two different Likert type items of the ASLE survey have been plotted against one another. Data point sizes are proportional to the frequency of response. The slope value, indicative of the rate at which the "overall" rating changes as the "interest" rating changes, has been calculated based on assigning successive integer values to the five rating scale categories, here labelled as A (the most positive response) through to E (the least positive response). Images have been reproduced with permission from George *et al*.[59]

Correlations such as these were stated to *"provide a valuable insight into the factors that significantly influence students' learning experiences"*.[59] In this specific instance, because items 1, 3, 6 and 12 concern factors reflecting affective and cognitive engagement, this result was taken to show that "*students' evaluations of the learning aspects of a laboratory activity appear to derive from the high-level engagement and deep learning for which we strive*".[59] Conclusions such as these rest on the validity of the integer scoring method applied, as well as the valid interpretation of the meaning of survey responses by researchers utilising ASLE surveys.

1.2 Validity and ASELL

1.2.1 Quantitative methods: categorical data, parametric statistics

Strictly speaking, Likert "scales", for which Likert-type response format items are usually used, emerge from summating or averaging scores obtained from an entire set of items, rather than the individual items themselves within that set as ASELL and its predecessors have done.[63] Many studies have shown that Likert scales emerging from multiple summated items can validly be treated as interval scale and are fit for parametric statistics, but not in the case of analysing individual items – a practice recommended against.[64, 65] Recommendations in the literature suggest that Likert-type response format data should be treated as ordinal rather than interval scale, implying that non-parametric methods should be used[66-68] such as the chi-squared test[69], Mann-Whitney U test (also known as the Wilcoxon rank sum test),[70-73] the Kruskall-Wallis test,[74] Kendall's tau[75] and Spearman's Rho,[76] whilst parametric methods such as reporting means and standard deviations, use of Student's t-test for equal[77] or unequal[78-80] variances, ANOVA[81-83] and Pearson's correlation coefficient[84] should be avoided and considered invalid.[85] Parametric comparisons of mean values such as t-tests and ANOVA are said to be inadvisable for individual items due to difficulties in obtaining normally distributed data, whilst use of Pearson's correlation is considered particularly inappropriate because it is influenced by the range of the score values used.[86]

However, some statisticians have no issue with scoring ordered categorical data for the purposes of correlations[87] and other parametric methods. F-tests, specifically those utilised by ANOVA, have long been demonstrated as being extremely robust to violations of the interval data assumption[88] and Pearson's correlation has been shown to be "insensitive to extreme violations of the basic assumptions of normality and the type of scale".[89] As such, some disagree that these methods are inappropriate for Likert-type data. Some even go so far as to claim that scoring Likert-type data for the sake of conducting t-tests is not only acceptable, but is superior to using rank-based tests such as the Wilcoxon, which should be avoided.[90] Norman[91] concludes:

> *"Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of ''coming to the wrong conclusion''. These findings are consistent with empirical literature dating back nearly 80 years. The controversy can cease (but likely won't)."*

Perceived key limitations in the integer scoring technique concern the lack of sample independence of scored rating scale responses, as well as the difficult interpretation of scores and differences between them. A difference between two scores may mean a certain magnitude of difference in the evaluated attribute at one location on the scale, whereas the exact same score difference may imply an attribute difference of an entirely different magnitude at another point on the scale. For example, the progression from "agree" to "strongly agree" may not reflect an improvement of equal magnitude to the progression from "disagree" to "neutral", as the equal score difference would appear to suggest. Related to this idea is the fact that averaging or summating two scores may not make qualitative sense. Paraphrasing Kuzon Jr *et al.*[92], Jamieson[85] states:

> *"the average of 'fair' and 'good' is not 'fair-and-a-half'; this is true even when one assigns integers to represent 'fair' and 'good'!"*

It is objections like these which are used to justify the usual recommendations that Likert-type response format data should not be treated with simple scoring techniques or parametric statistical methods, as ASELL and its predecessors have done.

Barrie *et al.*[61] employed the technique of correlating scored responses to individual Likert-type items in a recent publication. The authors subjected scored ASLE response data to principal component analysis: a procedure which uses observed (Pearson) correlations between scale variables to extract a number of major dimensions characterising the full data set. These extracted dimensions or 'principal components' are linear combinations of the original dimensions of the data set (in this case, the individual ASLE survey items). One of the principal components extracted from the ASLE data was identified by the authors to be a "resources" factor, comprised of the survey questions relating to the demonstrators, laboratory notes and background material provided. The authors presented the following analysis (Figure 7) of the items comprising this "resources" factor, this time using mean scores obtained from a variety of experiments.

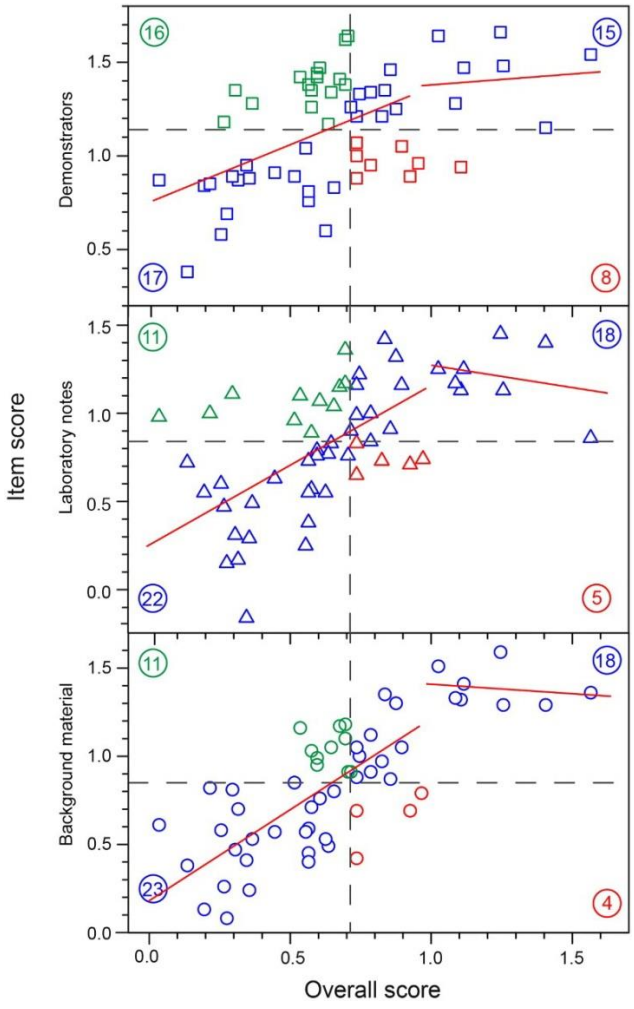


Figure 7: Example of scored ASLE data being treated as an interval scale measurement

Mean scored responses calculated for different items of the ASLE survey have been plotted against one another in order to show evident correlations. Lines shown in red indicate the inferred rates of change in responses given for the item on the vertical axis with respect to change in mean score for the overall learning experience item, making the assumption that the calculated mean scores can be treated as interval scale measures of the subject of the survey items. Circled number values indicate the number of data points in the relevant quadrant of the plot. Image reproduced with permission from Barrie *et al.*[61]

Not only are Pearson correlations used in this technique, but the individual data points involved are also presumed to be interval scale variables, for each singular survey item: both are issues of contention in suggested practice with ordered categorical data, as discussed previously. The caption to this figure presented in the original paper included the statement: *"A break in the regression line is evident at an overall score of 1.0, indicating that improvement in the overall laboratory experience no longer depends on resources once a certain standard is reached".* Such a statement inherently presumes that the mean scored response is linearly related to the variable underpinning that response. Conceivably, however, it may be the case that a much larger improvement is needed to shift from "good"(scored +1) to "excellent"(scored +2) than is required to shift from "average"(scored +0) to "good"(scored +1). In the figure above, the apparent lack of improvement could therefore simply be an artefact of the response scale used, rather than a genuine plateau in perception. This illustrates the way typical analyses of ASLE survey responses inherently assume an 'equidistant' nature of the response categories, influencing the conclusions drawn from the data.

The validity of treating the ASLE rating scale data in the usual manner is therefore unclear, and investigation needs to be conducted in order to establish the appropriateness or otherwise of these methods, which have been applied to a vast array of data spanning back over a decade. The reliability of past conclusions drawn from data to which these methods have been applied rests on the validity of treating ASLE survey rating scale item responses with this scoring methodology. More fundamental questions additionally exist, however, concerning the ability to validly measure experiment quality in the first place.

1.2.2 Qualitative interpretations: what does ASLE data really reflect?

The earliest, simplest concept of "valid" measurement is simply that the instrument used (in this case the ASLE survey) measures what it purports to measure.[93] Since the statement of this simple definition, multiple types and perspectives of validity emerged. A number of these came to be commonly viewed as facets of "construct validity"; broadly defined as whether empirical relations between observed outcomes are consistent with theoretical relations within a nomological network.[94] Others still argued the topic covered a broader range of ideas than this,[95] such as the social and ethical consequences of test use.[96] This expansion of the validity concept and the defining of various different types of validity continued to persist,[97-100] leading Shaw and Newton to recently conclude that agreement over the meaning of the term is unlikely, therefore recommending the term "validity" be abandoned.[101]

Borsboom, Mellenbergh and van Heerden have attempted to consolidate and simplify this convoluted and expansive understanding of valid measurement by reducing it down to two essential criteria; a measure of an attribute is valid if and only if (a) the attribute exists and (b) variations in the attribute cause variation in the outcome.[102] On this understanding, the concept of valid measurement involves both the mathematical techniques used to obtain numerical values from observed data, as well as truth of the presumed correspondence between those number values and the trait purported to be measured with those numbers. There must exist some attribute which may consistently be assigned a meaningful numerical value, which can be said to be a "measure" of that attribute. Further, changes in that measurable attribute, if it exists, must change the observed outcomes of surveys (or other measurement instruments) from which the researcher obtains the purported measures of that attribute.

In the case of the ASLE surveys' Likert-type items, what the survey "purports to measure" is intimately tied to the surveys' common uses. These surveys have, since their creation, been distributed and analysed with the intention to evaluate the quality of an experiment as experienced by students, with each different Likert-type item's responses often taken to reflect student perspectives of different experiment qualities. Of note here is that it is the quality of the *experiment* that is inferred from survey responses, not the disposition of the student cohort performing the evaluation. The fact that variation in student dispositions towards positive response could in theory alter the responses obtained was recognised well by Southam *et al*,[60] who stated:

> *"There are limitations with this study, most obviously the convenience sample at a single institution using a self-report instrument. This brings forth issues of equivalence as data from different samples are compared, combined and inferences drawn."*

What is being recognised here is the lack of sample independence in survey responses obtained. It is acknowledged here, explicitly, that responses obtained do not exclusively reflect properties of the experiment itself: measurement of experiment quality is confounded by student dispositions. This is potentially problematic, as conclusions drawn from ASLE survey data often use survey responses to gauge experiment quality in a more sample independent sense. For example, Read and Kable make the following statements about the "thermodynamics think-in" experiment, following evaluation using ASLE surveys[50] (emphasis and added commentary is shown in bold):

> *"Analysis of the data shows that students enjoy working on the practical, and report it **[the practical]** to be a beneficial learning experience that effectively develops their understanding of thermodynamic principles. **The practical** also fosters significant interest, and through a process of collaboration and cooperation aids the students in further developing their generic thinking skills."*

> *"Clear evidence has been presented that **this experiment fosters** cooperative learning and teamwork, triggers and maintains student engagement and interest, and is perceived to be highly relevant."*

It is clear here that survey data has been used to draw conclusions about the experiment itself as a tool of education. Further, there is an implied sense that this experiment, if run with other students, could be used to elicit similar educational benefits.

In the case of the ASLE surveys, the attributes which must be shown to exist in order to satisfy Borsboom, Mellenbergh and van Heerden's first requirement of valid measurement are therefore the (measurable) experiment qualities targeted by each specific Likert-type item of the survey, *as true for most students*. Unless these attributes have a broadly student-independent component, they cannot be said to be qualities of the experiment itself, but rather qualities of the student body selected to perform the evaluation. Establishing measure validity in this case therefore requires the demonstration that experiment quality can reasonably be said to exist in a student-independent and somewhat "objective" sense, to which a number value can consistently be ascribed. Scored ASLE responses need not reflect this "objective" experiment quality *exclusively*, but certainly must reflect this *predominantly* if mean scores are to be used as presented above. For scores to be treated as reflecting a

generalizable measure of experiment quality, there must exist a generalizable sense of experiment quality for scores to reflect. Further, factors specific to the student sample must not confound survey responses so extensively as to obscure the "objective" experiment quality beyond retrieval.

A related issue is the presumed correspondence between a change in experiment quality and a change in observed survey response. For example, it is often presumed that student perception outcomes could be improved by making amendments to the design of the exercise evaluated. An experiment may be subjected to the ASELL review process, then may be revised and ideally improved based on the suggestions or comments made. This assumption that changing the experiment design may improve survey outcomes is reflective of Borsboom, Mellenbergh and van Heerden's second requirement of valid measurement: that variation in the measured attribute must cause variation in the outcome. The "outcome" here is the (scored) set of ASLE survey responses received, whilst the relevant "attribute" is an attribute of the student learning experience, theoretically emerging from the exercise's design. The theoretical connections between experimental design and the (measurable) attributes of the experiment targeted by the items of the survey are the nomological network discussed in the concept of "construct validity"; similarities and differences in observed survey outcomes must directly map to the predictions of these theoretical connections. Changing experiment design should in theory change the measured attributes, which should therefore change the observed outcomes. Failing this, recommended practice suggested by ASELL project research may not yield the benefits it claims. Unfortunately, a detailed theoretical understanding of the connection between experiment design and measured attributes of the student learning experience is not yet understood. For this reason, the ability to satisfy Borsboom, Mellenbergh and van Heerden's second requirement of valid measurement is limited for the ASLE survey Likert-type item scores. It is not yet possible to confidently and precisely predict an expected change in survey outcomes, given a specific change in experiment design. Some crude, qualitative expectations may currently be possible; however, the experiments evaluated using ASLE surveys frequently differ in multiple respects, making expectations based on these crude understandings alone less clear and lacking in certainty.

Strictly speaking, to validly claim *interval scale measurement*, (the measurement presumed in applying parametric methods to scored ordered categorical data) changes in observed scores should occur in fixed proportion to the magnitude of changes in the underlying trait they are claimed to reflect. For example, a change in score of magnitude +1 should reflect a fixed magnitude of change in the underlying trait; a progression from a "good" experiment to an "excellent" experiment should only yield the identical change in score as the progression from a "poor" experiment to an "average" experiment if those progressions are in fact of the identical magnitude. Verifying this would require a quantitative understanding of both the underlying "objective" experiment quality attributes themselves as well as their precise connection to the scored responses observed. Quantitative models of survey responses, able to make more specific and testable predictions, need to be formulated in order to probe these connections further. One technique of obtaining such quantitative models is the use of Rasch analysis.

1.3 Rasch analysis

1.3.1 Measures as opposed to scores

In the administration of tests and in the field of psychometrics, a stark distinction is made between "scores" and "measures". A wealth of literature exists discussing "classical test theory"[103, 104] and "latent trait theory"[105] (often referred to as item response theory).[106, 107] Classical test theory is based upon the integer scoring techniques commonly applied to tests (in the form of adding "marks" to obtain a final score) and survey results (for example in the case of the ASLE surveys), whilst item response theory takes the observed outcome to be a result of some latent trait underpinning respondents' propensity to provide various different responses. Whilst it has often been shown that little difference exists between the values obtained from either theory,[108-112] "scores" of classical test theory are viewed as having limitations that the "measures" of latent trait theory do not possess. The fact that scores theoretically cannot be treated as interval scale whilst measures can is one such limitation.

Rasch measurement provides a means of avoiding the controversies of bridging the gap between observations in ordered categories and interval scale measurements[113-116] and has been claimed to be the only mathematical formulation capable of converting observed counts into true "measures" as opposed to mere "scores".[113] The Rasch model has been recognised as useful in educational research because of this property[116-119] and has been used for survey validation in this field previously.[120-122] Based upon the works of Georg Rasch,[123, 124] the model was initially developed for dichotomous responses (yes/no , correct/incorrect etc) allowing for the estimation of measurements associated with survey or test items independent of the persons sampled and similarly, person associated measures independent of the survey or test items posed.[123-125] The model has since been expanded beyond dichotomous responses, though all types of Rasch model still express the probability of response in each available category as a function of some latent trait measure underpinning the response, usually broken up into a person specific term and an item specific term.[115, 126] The Rasch model (both the original dichotomous response model and its extensions) may be derived directly from the need to maintain "specific objectivity" of each facet of the model (eg. person, survey or test item, etc).[127, 128] On a test, for example, the measure of ability for one person relative to the measure of ability for another must be independent of the question asked. Similarly, the measure of difficulty for one question relative to that of another question must be independent of the person responding. The Rasch model has been demonstrated to be the only rating scale model capable of this measure objectivity, which is necessary for scientific comparisons.[129]

Generally, parameters of the model include a measure for each separate person, each of which reflects that individual's "objective" propensity to respond higher or lower on the rating scale, and an item parameter for each survey item which reflects the "objective" difficulty of responding in higher categories faced by persons responding to that item. Parameters defining the point of equal probability of responding in either of two adjacent categories are also included where the rating scale has more than two options. The clear advantage of this treatment of the response scale over the usual scoring methods is that response categories are no longer assumed to be equidistant. Rather, the category structure is estimated based on the observed data. Because of this, values obtained from Rasch modelling are not subject to the same controversies as scored responses when parametric statistics are applied to them.

Rasch models may also be constructed in a variety of different ways, each reflecting a different conceptualised interaction between observed responses and underlying latent variables (discussed later in section 2.2). Rasch models therefore not only provide an alternate, more sophisticated means of quantitative analysis for the ASLE surveys, but additionally enable the exploration of more qualitative aspects of survey data interpretation. Crucially, an array of fit statistics may be used to test the fit of observed data to any given Rasch model, opening the possibility to explicitly test the construct validity of any model posed. This means Rasch measurement permits the ability to test whether "objective" measures associated with survey responders or items can reasonably be assigned in almost any manner suggested.

1.3.2 The Rasch model as a tool of validation

Drawing on the concept of an equation relating attribute measures to some substantive theory, as well as a measurement mechanism relating the attribute measures to observed outcomes,[130] Stenner *et al*. advocate a similar view of validity to Borsboom, Mellenbergh and van Heerden in presenting the benefits of considering causal rather than merely descriptive Rasch models.[131] That is, models where the observed responses are interpreted to be *caused* by the fact that person and item measures take the value that they do, as opposed to the measures merely being descriptive of general trends in the observed data. In justifying this interpretation, techniques have been implemented to demonstrate that item parameter values are a direct function of the features of the task at hand. That is, the measure is validated: changes in attributes of the task are shown to cause a change in the observed measures and outcomes obtained. Through experimentation, an equation may be determined which derives the value of the measure directly from the attribute. This is known as a "specification equation".[132] Because such an equation may be used to make quantitative predictions, accuracy of the equation can be explicitly tested using observed data, as is the case for all Rasch models.

An example of the successful use of a specification equation is the derivation of text readability measures (termed "Lexiles") computed directly from the text, which are able to predict student reading test scores.[133-135] In this example, the (Rasch) item measures reflecting text readability are shown to be a mathematical function of elementary features of the text itself: the log mean sentence length and mean log of word frequencies. Because of this direct relationship between the text and the Rasch measures, Rasch measures are known to reflect the trait they are purported to reflect and are therefore, by definition, valid measures. Changing the text results in changing the Rasch measure by a known quantity, thus reading test scores may also be predicted to change by a quantifiable degree. This case exemplifies the power of developing a specification equation for a data set.

Figure 8 illustrates the role of the specification equation and causal Rasch model as integral components of the theoretical construct connecting laboratory activity design and responses students provide on the ASLE surveys. The act of the researcher is to infer latent trait measures from observed data, then seek patterns in these measures to discern which features of laboratory activity design influence their values (and hence the observed responses). This is achieved by estimating latent trait measures through Rasch modelling (ensuring construct validity of those measures), then discerning how these measures vary given the experiment design via either controlled experimentation or observational studies. The specification equation and causal Rasch model then constitute the theorised connection between activity design and measurable experiment attribute, and between measurable experiment attribute and observed survey responses respectively.

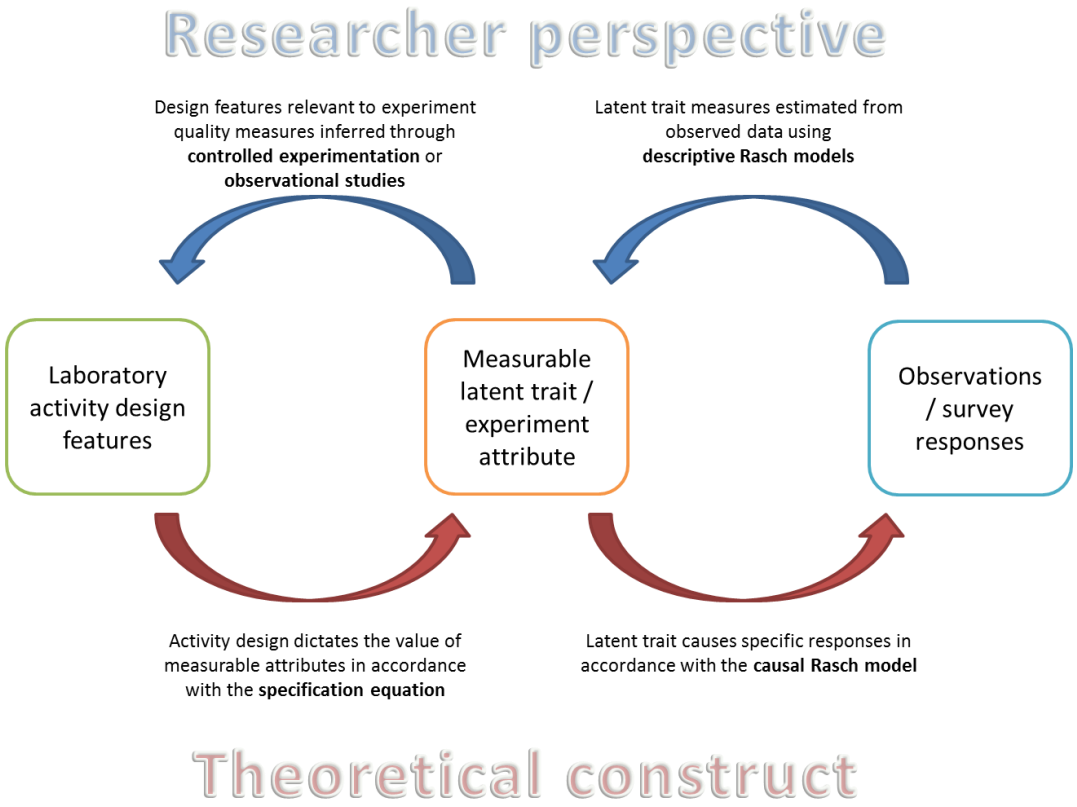


Figure 8: Perspectives on the connection between laboratory activity design and observed student response data

The requirements of measure validation for the ASLE surveys, that measures exist and are a function of the laboratory design, are inherently met if a specification equation and causal Rasch model are determined. Establishing this connection between observations and measures via the measurement mechanism (causal Rasch model), and developing a substantive theory of why the measures take on the values that they do (expressed in the specification equation) not only serves to validate the measurement techniques and interpretations, but also allows an in-depth, quantitative model of the measured attributes of interest, able to be used predictively and therefore able to be experimentally supported or refuted. A complete mathematical formulation connecting laboratory activity design and observed survey response such as this would allow the quantitative prediction of ASLE survey outcomes for any proposed laboratory activity design. This would not only be of practical use in designing or improving laboratory activities, but it would also allow scientific investigation of the way laboratory experience operates via the testing of predictions made from the current model.

1.4 Outline of this thesis

1.4.1 Immediate aims and hypotheses

The primary aim of this thesis is to investigate the validity of using ASLE survey Likert-type data to draw inferences regarding students' perceived laboratory learning experience. In so doing, an additional aim is to establish and characterise how relevant features of experiment design influence perceptions of the laboratory learning experience from the student perspective. Pursuing these aims, the work presented in this thesis can be described as addressing three core hypotheses, each of which can be tested using Rasch measurement.

> ***Hypothesis 1:***
> *Conclusions drawn from the ASLE survey data using typical scoring techniques resemble conclusions drawn using sample independent, interval scale measures extracted from the same data.*

This hypothesis underpins a large quantity of work that has been performed using the ASLE surveys in the past, most notably the analyses which utilise parametric statistical techniques such as calculation of mean scores, standard deviations and correlations. As Rasch modelling provides the only means of converting ordered categorical counts into sample independent, interval scale measures, this hypothesis must be tested by contrasting Rasch measures for a specific experiment and survey question with their corresponding ASELL scores. This major theme will be addressed in multiple ways within this thesis.

In the first instance, a typical score-based investigation into laboratory learning experience using the ASLE surveys will be presented. This study will then be revisited with Rasch modelling techniques, critically evaluating the validity of the score-based study conclusions. A more in depth analysis of the mathematical relationships between score and measures will also be presented. Rasch models embodying the usual presumptions underpinning ASLE survey data may be formulated, thereby estimating interval scale measures associated with the data, presuming those measures are valid. These measures may then be contrasted with their corresponding mean ASELL scores, evaluating the relationship between the two. Additionally, the distribution of scored responses expected for any given Rasch measure may be generated directly from Rasch model parameters, enabling an assessment of the appropriateness or otherwise of parametric statistical methods when applied to scored data. All of these techniques, collectively, test Hypothesis 1 from various perspectives.

> ***Hypothesis 2:***
> *Student independent contributions to the ASLE survey responses exist and are measurable.*

This hypothesis, as well as the next, is drawn directly from the requirements of measure validity. Even if scored ASLE responses resemble interval scale measures estimated via Rasch analysis, it is still presumed that the Rasch model is a valid description of the way the observed data operates. Given that the Rasch model is (as previously described in section 1.3.1) the only mathematical formulation capable of converting observed counts in a set of ordered categories into interval scale measures, the fit of the observed data to the Rasch model provides a means of assessing whether any valid interval scale measures of quality can be assigned to the surveyed experiments in the first place. Other means of testing Hypothesis 2 also exist, and these may be employed concurrently with methods which also test the third hypothesis of these works. The presumption that ASLE data reflect properties of the

experiments evaluated and therefore properties independent of the students responding, which can reasonably be assigned a single number value, rests on these hypotheses.

> ***Hypothesis 3:***
> *Student independent measures obtained from ASLE survey data reflect qualities of the experiment evaluated.*

A great deal of flexibility in the construction of Rasch-type models is permissible, thereby allowing for most conceivable hypotheses concerning the connection between experiment design and student perception to be incorporated into Rasch models and evaluated. This not only means that a Rasch model which embodies the way student, survey question and experiment are presumed to interact may be tested, but so can other Rasch models reflecting different conceptualisations. Formulating different Rasch models for the observed data and calculating corresponding fit statistics for each therefore allows for the assessment of which model best explains the data observed. The best model may or may not contain student independent facet(s), thereby testing Hypothesis 2 above. The experiment specificity of any student independent measures identified in the best explanatory model determined may also support or refute the third hypothesis of this work.

1.4.2 Long term goals

Some Rasch models, such as the Linear Logistic Test Model (LLTM, detailed in later sections: see section 2.2.2) achieve direct mathematical links between experiment structure and likely student perception, thereby enabling student responses to be predicted quantitatively. These predictions can then be empirically contrasted with observation, leading to reformulation of the Rasch model such that it provides more accurate predictions. Utilising statistical techniques to compare various models for their efficacy as an explanation of the observed responses, it is conceivable that iterated re-formulation of Rasch models in this way could lead to the generation of a specification equation for the ASLE surveys. That is, an equation expressing the measures of quality for an experiment (as true for most students) as a direct mathematical function of the design of the experiment itself.

Development of a specification equation would not only entirely validate the notion that measures obtained genuinely do reflect qualities of the experiment, but far more crucially would reveal why some experiments are perceived more or less positively than others. This would be invaluable to educators, as the specification equation could be exploited to structure laboratory exercises which produce the circumstances most likely to be appealing to most students. This has uses ranging from improving student engagement and potentially therefore improving knowledge retention in laboratory activities, through to improving student enrolment and retention in science courses. These objectives were primary goals of the ASELL project at its inception, and are potentially achievable through the use of Rasch modelling in the manner described.

Investigations presented in this thesis therefore address the topic of connections between laboratory activity design and estimated Rasch measures, inherently also investigating the truth or falsity of Hypothesis 3 above. This is to be pursued firstly by investigating whether student independent Rasch measures may validly be interpreted as specific to the experiment conducted (as described above) and secondly, if possible, by identifying the components of experiment design which contribute to the value of those measures. This may be facilitated by formulating models such as the LLTM, provided those models serve as an adequate

explanatory model of the observations. By investigating connections between ASLE survey responses and components of experiment design quantitatively via the Rasch models, it is hoped that this work may lay the foundation for future investigations identifying and refining a specification equation for the ASLE survey responses.

2 General methods

2.1 Data collection: surveying first year chemistry laboratory sessions

2.1.1 Ethical approval

Approval to gather ASLE survey data discussed throughout this thesis was granted by the University of Adelaide Human Research Ethics Committee on the 25th of July, 2012 (Approval number H-2012-097). Copies of the information presented to participants are available in the supporting information (see section 7.1).

2.1.2 Student cohorts

Data utilised in the analyses presented in this thesis originate from two distinct cohorts of first year undergraduate chemistry students at the University of Adelaide. Students enrolled in the courses Chemistry IA and Chemistry IB (run in semesters 1 and 2 respectively) are required to have attained prerequisite levels of achievement in high school chemistry, whereas students enrolled in the courses Foundations of Chemistry IA and Foundations of Chemistry IB (again in semester 1 and 2 respectively) are not. Typically, approximately 550 students are enrolled in Chemistry IA or B courses, whilst approximately 450 students enrol in the Foundations of Chemistry courses from a diverse range of backgrounds, though these numbers vary from year to year.

Students enrolled in these courses were provided with the opportunity to complete the ASLE survey (see Table 1 in section 1.1.1) at the end of their laboratory sessions. Surveys were presented to students as optional, and in the early stages of data collection, anonymous. During latter stages of data collection, surveys provided students with the opportunity to supply their student identification number, assured that the number would never be used to directly identify them by name. In the case the student's laboratory demonstrator was also an analyst of the survey data, student provision of an identification number was not made possible, and anonymity was ensured. All students who chose to complete the survey had the option of not providing their identification if they wished.

2.1.3 Experiments surveyed

Experiments were conducted by students during fortnightly laboratory sessions, in which they were allocated 3 hours to complete the experiment procedure and accompanying laboratory booklet questions for assessment. Online pre-laboratory questions were also required to be completed for each experiment prior to that experiment's laboratory session.

In the earlier years in which data used in this thesis were collected (prior to 2012), both the Chemistry IA/B and Foundations of Chemistry IA/B shared a number of experiments in common, with no alignment with lecture material and a randomised sequence of experiments during the semester. Students enrolled in these courses were randomly assigned practical groups, with each group conducting experiments in a different sequence during fortnightly laboratory sessions. Chemistry IA/B courses included six laboratory activities per semester, whilst Foundations of Chemistry courses included only five. However, the first laboratory session in each semester for the Foundations of Chemistry courses was not a 'wet' lab session, leaving four practical laboratory experiments conducted each semester by the Foundations of Chemistry cohort. These four experiments were all experiments also conducted by the Chemistry IA/B students. The original list of experiments conducted during the earlier years of

data collection is given in Table 2, including descriptions of the laboratory activities, semester in which they were conducted and whether students worked in pairs or individually.

Table 2: Initial list of laboratory experiments conducted by students

Experiment title	Description	Conducted by
Biological buffers *(Experiment 1)*	Students generate titration curves with the aid of technology to graphically investigate the pKa and effective range of buffer solutions, including histidine.	Both cohorts, semester 1 pairs
Thermochemistry *(Experiment 2)*	Students conduct simple calorimetry experiments and perform the appropriate calculations in order to calculate the enthalpy of formation of ammonium chloride using Hess' law.	Chemistry IA students only pairs
Vapour pressure *(Experiment 3)*	Students measure the vapour pressure of a number of mixtures of cyclohexane and ethanol at a range of different mole fractions with the aid of technology. The results, in conjunction with the application of Raoult's law and Dalton's law, are used to show the mixture is non-ideal and hence that the two substances have differing intermolecular forces.	Chemistry IA students only pairs
Melting points and recrystallization *(Experiment 4)*	Acetanilide is recrystallised from a crude sample, and the melting points of both the crude and purified samples are obtained and used to briefly assess purity.	Both cohorts, semester 2 individuals
Quantitative techniques *(Experiment 5)*	Students test the precision and accuracy of a volumetric pipette by using the measured mass of water pipetted. Students then determine the concentration of a sulfuric acid solution by titration against sodium hydroxide, including associated error calculations.	Both cohorts, semester 1 individuals
Reaction kinetics *(Experiment 6)*	Students react iodide and persulfate ions a number of times, varying reactant concentrations, temperature and the presence of a catalyst. This is used to draw inferences about the rate of the reaction, including determination of the rate law. The experimentally derived rate law is used to support or refute proposed reaction mechanisms.	Both cohorts, semester 2 pairs
Liquid-liquid extraction and TLC *(Experiment 8)*	Students perform an acid-base liquid-liquid extraction in order to separate an acidic compound (salicylic acid) and a neutral compound (3-nitroacetophenone). The results are then analysed by thin-layer chromatography (TLC)	Both cohorts, semester 2 individuals
Synthesis of aspirin *(Experiment 9)*	Aspirin is synthesised from salicylic acid and acetic anhydride. The product is then recrystallised and analysed against pure aspirin and pure salicylic acid samples by Thin-layer chromatography to test for purity.	Chemistry IB students only individuals

Experiment title	Description	Conducted by
Coloured complexes of iron *(Experiment 10)*	Complexation between iron(III) and the acetyl acetonate (acac) bidentate ligand is used as an example to demonstrate equilibrium and Le Chatalier's principle. The addition of sodium acetate is shown to shift equilibrium in favour of the formation of the tris acac complex, demonstrated using pH measurements as well as visual inspection of solutions. Solubility of the different iron complexes in polar or non-polar solvents is used as a tool to observe the equilibrium shift visually.	Both cohorts, semester 1 pairs
Analysis of spinach extracts *(Experiment 11)*	Liquid-liquid extraction is used to isolate coloured organic compounds from spinach leaves, with the resulting green solution analysed by thin-layer chromatography.	Chemistry IB students only individuals
Ion exchange chromatography *(Experiment 12)*	Iron complexes are prepared in the presence of three different conditions: dilute hydrochloric acid, concentrated hydrochloric acid and concentrated hydrochloric acid with added heat. Students perform ion exchange chromatography on the coloured products, rationalising the differences in observed results depending on the reaction conditions used.	Both cohorts, semester 1 pairs
Determination of copper(II) ion concentration *(Experiment 13)*	Students use a standard solution of copper sulphate and perform serial dilutions to obtain a variety of different concentrations, each of which has its absorbance measured at a selected wavelength of light. With the aid of technology, students generate a calibration plot with these measurements, then use their plot to determine the concentration of a solution from its absorbance by applying Beer's law. Students also briefly observe the relationship between wavelength and colour by adjusting a bench top spectrophotometer and observing the colour of a laser beam.	Both cohorts, semester 1 pairs

At the beginning of 2012, the Foundations of Chemistry courses were modified. Practicals (experiments) were no longer randomised, but conducted in the same sequence for all students in an order designed to align with lecture content as closely as feasible. Some specific experiments were modified to suit the Foundations of Chemistry cohort in small ways, including small alterations to experimental procedures, questions asked in the answer booklet and information provided in the instruction document for the experiment. The specific experiments conducted were also changed in some instances, with the Foundations cohort being presented with experiments they had not been in previous years. Three new experiments were also devised explicitly for the Foundations of Chemistry cohort. The laboratory components of the Chemistry IA/B courses remained as they had been in previous years.

A list of experiments conducted by the Foundations of Chemistry cohort from 2012 onwards is provided in Table 3, all of which had small alterations to the introductory material provided in the laboratory instruction manual for the experiment. Some small further revisions were also made to some of these experiments after 2012, and these changes are also noted. The order

in which these experiments were presented also changed from one year to the next in a small number of cases. Experiments are listed in their initial 2012 order in Table 3, with later amendments noted.

Table 3: Foundations of Chemistry cohort experiments from 2012 onwards

Experiment title	Description	Notes
Introductory experiment	Students complete a number of questions involving basic chemistry concepts such as atomic structure and simple calculations. Initially not a 'wet' lab session.	Conducted individually. Later revised (2013) to include a video of some possible laboratory observations students were asked to comment on. A second, later revision (2014) included the pipetting section previously a part of "Quantitative techniques", and no longer this observations video.
Quantitative techniques	Initially equal to the experiment of the same title described previously, however students were not required to perform error calculations.	Conducted individually. Later revised (2014) not to include the section related to pipetting. Presented second in semester 1 during 2012 and 13, presented third in 2014.
Determination of vitamin C concentration	Students determine the concentration of vitamin C in a commercial brand of apple juice by redox titration against iodine solution. Students must standardise the iodine solution first by titration against a known solution of vitamin C they make themselves, then use the iodine solution of now known concentration to determine the concentration of vitamin C in the commercial juice by titration.	Conducted individually. Presented third in semester 1 during 2012 and 13, presented fourth in 2014.
Equilibrium and Le Chatelier's principle	Equivalent to the experiment previously labelled "Coloured complexes of iron", with some amendments to the question booklet.	Conducted in pairs. Presented fourth in semester 1 during 2012, presented fifth in 2013 onwards.
Visible absorption spectrophotometry	Equivalent to the experiment previously titled "Determination of copper(II) ion concentration", with some modifications. Students record absorbance values from communal machines rather than each working pair having access to their own, and construct their calibration curve on paper rather than using technology. Students also do not observe the relation between wavelength and colour.	Conducted in pairs. Presented fifth in semester 1 during 2012, presented fourth in 2013, presented second in 2014.

Experiment title	Description	Notes
Aromachemistry	Students are provided with a number of sample vials containing unknown compounds, and are asked to identify which vials correspond to which molecules, having been told the list of compounds and their aromas. Students then answer questions regarding the systematic nomenclature of organic molecules and regarding basic reactions such as oxidation of organic compounds.	New as of 2012. First experiment in semester 2. Conducted individually.
Analysis of spinach extracts	Equivalent to the experiment previously described under the same title, with the exception that students work in pairs rather than individually.	Conducted in pairs.
Thermochemistry	Equivalent to the experiment previously described under the same title, with the exception that students are not asked one final question about entropy in the answer booklet.	Conducted in pairs.
Metal activity series	Students place a number of solid metals into a range of solutions containing metal ions, observing which cases result in metal displacement reactions. From these results, a metal activity series is derived. Students perform a similar process using halogen waters and halide solutions to derive an activity series for halogens.	New as of 2012. Conducted in pairs.
Reaction kinetics	Equivalent to the experiment previously described under the same title.	Conducted in pairs. The order and phrasing of some questions in this experiment's answer booklet were later revised for 2013 onwards.

2.2 Rasch model formulations

2.2.1 Unidimensional Rasch models

A substantial amount of research presented in this thesis involves fitting data to various forms of the Rasch model. A generalised Rasch-type partial credit model may be expressed as the following:

$$ln\left[\frac{P(X=x_k)}{P(X=x_{k-1})}\right] = \varphi - \tau_{q,k} \qquad \textbf{1}$$

Where P(X=x_k) denotes the probability that the observed response X, is equal to the k^{th} category of the rating scale x_k, the parameter φ is the latent trait measure and $\tau_{q,k}$ is the Rasch-Andrich threshold between the k^{th} category and the $(k-1)^{th}$ category of the rating scale, for the q^{th} scale group (a scale group being a set of items all with the same rating scale structure). The latent trait measure φ thus determines the probability of responding in each category of the rating scale. The τ parameters would be absent from a model with only two possible responses for each item.

The above expression has collapsed the latent trait measure φ into a single term. However, Rasch models generally express this term as a difference between respondent "ability" β and item "difficulty" δ; the person and item specific measures previously mentioned (see section 1.3.1). This produces a two-facet partial credit model (if using a response scale of more than two options):

$$ln\left[\frac{P_{n,i}(X=x_k)}{P_{n,i}(X=x_{k-1})}\right] = \beta_n - \delta_i - \tau_{q,k} \qquad \textbf{2}$$

Where the subscripts *n* and *i* identify the variables as being specific to the n^{th} respondent or i^{th} item respectively, and the i^{th} item is part of the q^{th} scale group. The partial credit model by Masters[136] would be one in which all items have their own scale group, whilst the Andrich rating scale model[137, 138] would be one in which all items are within the same scale group.

Splitting the latent trait parameter into only person and item terms is not the only possibility, however. Many facet Rasch models[113, 139] utilise multiple different terms, with each term specific to a different "facet". The multiple different facets interact to generate the latent trait measure, which interacts with the rating scale (or partial credit scale) to predict the probability of observing each category of response. A typical example is a scenario in which students are graded by multiple different judges, for a number of tasks.[140] Student, judge and task would each be assigned their own facet, splitting the latent trait measure φ from Equation 1 into three separate terms:

$$\varphi = S_n - J_m - T_j \qquad \textbf{3}$$

Where S_n is the "ability" of the n^{th} student, J_m is the "harshness" of the m^{th} judge, and T_j is the "difficulty" of the j^{th} task. The sum of the effect of these measures then interacts with the rating scale threshold parameters to determine the probability that the n^{th} student, rated by the m^{th} judge on the j^{th} task will result in an observation in each of the rating scale categories.

The way in which the latent trait parameter φ is split helps define the theoretical construct underpinning the observed responses by establishing which variables contribute to the final resultant observation in the rating scale. The latent trait underpinning responses to the ASELL surveys, for example, could be conceptualised in a number of ways. By changing the different facets included in the Rasch model used to explain the observed data, the factors theorised to be contributing to observed responses are correspondingly changed and thus so is the theorised mathematical mechanism connecting latent trait measures to observed responses.

2.2.2 Multidimensional Rasch models

Some types of Rasch model are able to express the measures described previously as a function of smaller, elementary components. That is, rather than expressing survey responses as a function of a singular latent trait varying along a singular dimension, the model may be reformulated such that multiple different measurable dimensions contribute to responses. Person measures, for example, may be explained as a linear combination of smaller variables. The Multidimensional Partial Credit Model (MPCM)[141] expresses each person measure β as a linear combination of M different person attributes. Here, a multitude of person attributes apply differently to different survey items/circumstances. This reflects the notion that students draw on different predispositions in response to different questions (and in different circumstances).

$$\beta_{i,n} = \sum_{m=1}^{M} w_{i,m}\theta_{n,m} \qquad \textbf{4}$$

In the expression above, $w_{i,m}$ serves as a weighting factor, scaling the degree of contribution of the m^{th} student attribute to the response of the question/circumstance assigned the i^{th} δ value. The values of $\theta_{n,m}$ reflect the relative magnitude of the n^{th} student's m^{th} attribute, when applicable. This model may be of use in reflecting the fact that a student may, for example, have a different tendency to provide positive response $\theta_{n,1}$ in the case an experiment contains mathematics, compared with their tendency to provide positive response $\theta_{n,2}$ in cases the experiment does not.

Another example is the Linear Logistic Test Model (LLTM),[142-144] where the item parameters are broken down in the following way:

$$\delta_i = \sum_{j=1}^{J} q_{i,j}\eta_j \qquad \textbf{5}$$

Here, the δ value is expressed as a linear combination of J many smaller components. Each component has a parameter η reflecting its relative contribution to δ if applicable, whilst the q values act as scalar 'weighting' factors reflecting the degree to which each component contributes to the measure for the i^{th} δ value.

A very simple example of the LLTM, where δ values correspond to a measure of quality of the i^{th} experiment (with respect to some specific ASLE survey question for simplicity of the example), would be a case where q_j takes the value 1 if the j^{th} component is relevant to the quality of the i^{th} experiment, and 0 otherwise. Component j could, for example, concern mathematical content. If the i^{th} experiment has mathematical content, q_{ij} would take the value 1, and so there would be an additional contribution of magnitude η_j to the experiment quality

(δ_i) because it contains mathematical content. A more nuanced example could be where the q values take different values reflecting the relative amount of mathematical content in the experiment, with experiment quality changing linearly as degree of mathematical content changes. Examples like this demonstrate this model's capability to clearly link the experiment quality measures (δ) directly to the design of the experiment.

The difficulty of applying this model lies in the identification of the different components which contribute to the measure of the δ facet and the assignment of the various q values weighting their contribution. Developing a matrix of j×i different q values (weighting j components linearly combining to explain i different δ measures) is not an issue unique to the LLTM. Establishment of "Q-matrices" as they are called has historically been relevant in the categorisation of test questions and the study of student misconceptions.[145] These matrices can in fact be estimated from observed data through generation and testing of random Q-matrices until an optimal Q-matrix is found,[146] however this procedure is computationally demanding. Steps for developing a Q-matrix commonly involve the following:[147, 148]

1. Identification of a set of components contributing to the response, usually informed by experts in the relevant field
2. Coding each item based on which components contribute to it and which do not. This develops the initial Q-matrix
3. Analyse the data with reference to the Q-matrix (for example using cognitive diagnostic models[149] or the LLTM)
4. Modify the initial Q-matrix based on observed output statistics associated with the modelled parameters, as well as theory
5. Repeat steps 3 and 4 until an acceptable Q-matrix is determined

However, for the ASLE surveys, this approach would require prior knowledge regarding the features of a positive laboratory experience, and the degree to which they contribute to each of the ASLE survey question topics. A large part of the ASELL project's purpose is to determine these contributing factors in the first place, and therefore any expert suggestion of likely contributing factors to include in the Q-matrix may be somewhat speculative in nature. A technique is needed to identify the factors of a positive laboratory experience and their relative contribution to each item of the ASLE survey before a meaningful Q-matrix can be constructed.

Both of these variations of the Rasch model: the decomposition of both person measures and item measures into linear combinations of many attributes or components respectively, are incorporated into the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM),[150] also known as the ConQuest model (implemented in the ConQuest Rasch measurement software).[151] As such, this model represents a general form of a large family of Rasch models; all models mentioned thus far are in fact specific cases of this general model, achievable by imposing various constraints on the MRCMLM.[152] The notation below expresses the parameters using the dot product of two vectors as opposed to summation notation, and expresses the model at the individual category probability level as opposed to the common log-odds form used thus far.

$$P(X = x_k) = \frac{\exp\left(\vec{w}_{i,k} \cdot \vec{\theta} - \vec{q}_{i,k} \cdot \vec{\eta}\right)}{\sum_{k=1}^{K_i} \exp\left(\vec{w}_{i,k} \cdot \vec{\theta} - \vec{q}_{i,k} \cdot \vec{\eta}\right)} \qquad \textbf{6}$$

Here, the θ vector contains the M many different student attributes which sum linearly to the student measure (previously β), whilst the w vector contains the scalar weighting factors which apply to each respective θ value for a specific survey item/circumstance (indexed by i). Similarly, the η vector contains the relative contribution of the J many different components of experiment quality which sum linearly to provide the experiment quality measure (previously δ), whilst the vector q specifies the weightings applied to each of these components respectively, depending on the survey item/circumstance (again, indexed by i).

A notable additional feature of this model is that it allows for variation in the weighting of the θ and η parameters depending on which category of response is being considered, indicated by the subscript k next to the w and q vectors (where K_i is the number of response categories possible for the i^{th} item). The rating scale category specificity of the q values incorporates the rating scale category structure in this model, achieved using a series of τ_k parameters in previously described Rasch models. The rating scale category specificity of both the w and q vectors is also of great use in modelling tests. For example, person trait θ_1 may be far more significant a contributor to obtaining the second mark of a question as compared to the first, whilst person trait θ_2 may contribute equally to both. This could be incorporated into the Rasch model by changing the value of w for different marks in the same question; progressively higher marks achieved being directly analogous to progressively higher scored response categories. Similar cases could be conceived for the different components contributing to the difficulty of those items (η).

These nuances, however, can reasonably be presumed not to contribute to the ASLE survey responses. The main purpose of introducing the existence of the MRCMLM here is that it serves as the general model of which a wide array of diverse Rasch models may be considered specific cases. It justifies the simultaneous decomposition of both the person measures and experiment quality measures of the ASLE surveys into several component parts or facets, some of which may take on different (or equivalent) values based on complex considerations. Rules defining when different facets apply and whether they take different values can be explained as different formulations of the matrices of w and q vectors in the MRCMLM.

2.3 Data treatment: generation of Rasch models

2.3.1 Rasch measurement software

Studies presented in this thesis make use of two programs designed specifically for Rasch measurement. The *Winsteps* program[153] is designed exclusively for two-facet Rasch models, akin to those described by Equation 2. It provides a comprehensive set of associated statistics with the model generated, including bias analyses and variance decomposition. The *Facets* program[154] is capable of many-facet Rasch measurement (see section 2.2.1) and therefore able to model a much broader array of conceptualisations of the ASLE data than the *Winsteps* program. *Facets* is also capable of generating accurate measure estimations for other models such as the Linear Logistic Test Model (section 2.2.2, Equation 5) by carefully defining and structuring the facets included.[155] The *ConQuest* program[151, 156] often used for these models was therefore not required, though may prove useful for future extensions to the work presented in this thesis.

The Winsteps and Facets programs both converge to optimised measure estimates in two phases. The first phase, PROX, obtains an initial rough estimate by assuming normally distributed measures.[157, 158] These estimates are then used as the initial values for joint maximum likelihood estimation (JMLE), which produces measures for each facet 'independent' of the other facets in the analysis.[125]

2.3.2 Confirmatory and exploratory applications: treatment of misfit

Whilst Rasch models are often considered a specific type of item response theory model, the conceptualisation of the relationship between data and theory differs substantially between item response theory and Rasch theory.[159] In contrast to item response theory's emphasis on structuring models which fit the data, Rasch modelling typically analyses data in the context of a pre-specified model (a Rasch model), assessing the fit of the data to the model as opposed to the reverse.[160] The reason for this is intimately tied to the question of validity: purported measures must be verified to fulfil the relevant criteria of appropriately being labelled a measure of a trait. Given that the Rasch model is the only mathematical formulation capable of converting observed counts into true "measures" as previously discussed (see section 1.3.1), misfit to the Rasch model is therefore interpreted as evidence of poor construct validity as opposed to inaccurate formulation of the mathematical model itself.[161, 162]

Consequently, response patterns exhibiting poor agreement between observation and Rasch model predictions may be removed from consideration as a matter of routine practice in Rasch measurement, as the inclusion of "misfitting" responses may compromise the measurement properties of the scale generated and perturb the estimated category structure of the instrument.[162] In the context of ASLE survey analysis, misfitting students are essentially interpreted as "donkey votes": those which do not follow a pattern which makes sense in light of the way most others respond to the array of experiments surveyed, given the Rasch model. This may occur because the student treats the response scale significantly differently to other students, views the set of experiments significantly unlike the way the other students do, or possibly even because their responses reflect nothing to do with the experiment at all. Once these misfitting students are removed, the results reflect the best estimates of the category structure and experiment measures that appear to be the case for most students. It is, however, important to note the number of misfitting students who do not adhere to this resulting model. Item measures in the Rasch model may also misfit, and in this case the

interpretation would be that the item cannot be assigned a true measure that appears to be reflective of any trend consistent for most students responding, in the context of the Rasch model utilised. Broadly speaking, poor fit in Rasch measurement implies invalidity of the measure construct: the values purported to be measures appear not to provide a true 'measure' of some consistently evident attribute.

Because of these considerations, Rasch analysis is typically confirmatory in nature: a specific Rasch model is presumed to be the correct expression of the measurement construct, in the case the attributes in question are measurable, and the fit statistics of the model generated are used to confirm this presumption. This does not, however, preclude studies which aim to determine which Rasch model is most appropriate for modelling the data. Use of the LLTM, MPCM and MDRCMLM mentioned previously (section 2.2.2) are often justified by statistically comparing these models to much simpler analogous Rasch models which do not model the measures obtained as linear combinations of multiple elementary variables. The initial descriptive Rasch model is contrasted with the more parsimonious model (such as the LLTM) to ensure that the data still fit the model to a comparative degree despite the decomposition of the person or item measures into smaller component parts.[152, 163] Section 2.5.4 describes the statistical techniques often used for these comparisons.

Comparative studies like those mentioned above are examples of cases where different formulations of the Rasch model are tested for their efficacy of explaining the observed data, despite the usual confirmatory nature of Rasch techniques. In cases such as these, the removal of misfitting observations would be in error. Removal of misfitting students during a study explicitly designed to contrast the fit of two alternate models would introduce bias into the comparative test, in favour of the model for which misfits had been removed. For this reason, misfitting data points were only removed from consideration in studies presented in this thesis in the cases where the objective was to determine the best estimate of the measurement construct, under the presumption that a particular model or interpretation is known to be appropriate and valid. Studies described where alternate models were compared did not remove misfitting data points.

2.3.3 Measurement construct issues: extreme and disconnected responses

It is feasible that all observed data points relevant to the estimation of a given measure may lie in the extreme positive or extreme negative response category. In cases such as this, the latent trait measure which gives rise to the observed responses cannot be precisely measured, as an infinite number of values beyond a certain point would all predict the same extreme set of observations. Assigning definite measures in these cases is therefore problematic[164] and as such, persons, items or other facet elements for which all observed data points are at the same extreme do not contribute to the measure estimation procedure and do not contribute to the model's various fit statistics. For this reason, the removal of extreme persons or items from consideration is common in the analyses discussed in this thesis.

All measurements output from Rasch model estimation are ideally within one frame of reference, and can be understood as being in a definite location on the scale relative to the other measures. This ideal, however, is not always realised. The possibility exists for different subsets of the observed data to be entirely disconnected[165] from one another. A simple example may be a case where one group of students (group A) provides survey responses for items 1 to 5, whereas an entirely different set of students (group B) provides responses only for items 6 to 10. In this case, measures for items 1 to 5 would not be directly comparable to

measures for items 6 to 10, as nowhere in the analysis does there exist a student who provided responses in both subsets of the data, which would otherwise enable the measures for these items to be assigned a numerical value relative to one another. Relative location of the measures on the scale is only assured within each of the two isolated subsets of data (group A and items 1 to 5, or group B and items 6 to 10). Scenarios akin to this may exist purely as a result of unfortunate patterns in sampling, or may exist as an artefact of the way various facets of the Rasch model are defined. Because measures are not comparable across subsets, data points appearing within small, isolated subsets separate from the connected bulk are on occasion (where stated) removed from consideration in the analyses described here.

In the event isolated subsets of data need to be made comparable, techniques are available. "Equating" techniques,[166, 167] as they are known, have the goal of placing the previously isolated measurement subsets into the same reference frame. Often this necessitates "anchoring" some measures to have specific values reasonably selected by the researcher, though this often comes at the cost of making an assumption. For example, two isolated subsets may be equated by presuming some items to have equivalent measures in the two different subsets, or it may be presumed that the distribution of student measures in each subset has the same centre (requiring both student measure subsets to be "group anchored" at the same value). Alternately, data sets may be ensured to be within the same reference frame for measurement if they share common persons responding, items the respondents are posed with, or common elements of another facet.

2.4 Data analysis: general statistical procedures

2.4.1 Statistical testing and family-wise error

Typically, statistical tests are conducted by reporting probability (p) values of the observed data being sampled under the presumed truth of some "null hypothesis". Individual statistical tests are deemed to refute the null hypothesis at $p<\alpha$, where the value of α reflects the probability of a type I error: the incorrect rejection of a true null hypothesis.

An important issue which arises when multiple different hypothesis tests are conducted is the problem of "multiple comparisons". In the case where multiple statistical tests are conducted on the same data set, the chances of incorrectly rejecting at least one true null hypothesis are raised, purely by virtue of the fact many tests are conducted. In general, if k many statistical tests are conducted, each deeming significant results at significance level α, then the probability of at least one type I error occurring, also known as the "family wise error rate" is given by:

$$\bar{\alpha} = 1 - (1 - \alpha)^k \qquad \textbf{7}$$

This implies that, for example, if fourteen statistical tests are used to detect difference in responses to any one of fourteen Likert-type items of the ASLE survey between two different evaluated experiments, the probability of inferring at least one significant difference at $p<0.05$, in the case the two experiments are in truth equal, is as high as 51%. This is one reason why the "shotgun approach" of testing for any difference between each singular Likert-type item individually when contrasting two different survey evaluations is heavily criticised.[35, 36]

A way of controlling for this highly undesirable effect is the application of the Bonferroni correction.[168] The Bonferroni correction operates by reducing the selected value of α simply and conveniently in such a way as to ensure the family wise error rate is at least as low as desired by the analyst. For a specified "family" of *k* hypotheses, the Bonferroni correction recommends deeming significant difference at α/k, where α here is the significance criterion which would ordinarily be applied were only a single test being conducted. As $\bar{\alpha}$ is always less than or equal to α/k, this ensures the family wise error rate is sufficiently small, and is in fact conservative methodology. This technique is applied in numerous cases, by necessity, in this thesis.

2.4.2 The normal distribution assumption

A number of statistical procedures, namely "parametric" methods, require data to follow a normal distribution. That is, the data are distributed as follows:

$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(\frac{-(x-\mu_x)^2}{2{\sigma_x}^2}\right) \qquad \textbf{8}$$

Where p*(x)* yields the probability density function of the variable of interest *x*, distributed with mean value μ (also known as the "expected value") and standard deviation σ. That is, given possible values of x are normally distributed, Equation 8 yields the probability that an observation randomly sampled from the population of *x* values will yield that specified value of *x*. Similarly, the integral of the function from $-\infty$ to some specified x value gives the probability of observing that specific value of *x* or less (a one-sided test). This can be used to evaluate the

probability that the observed value of *x* will lie between or outside of specified values (a two-sided test), as well as "confidence intervals" defining the range at which x is likely to be observed with a specified level of probability. A more convenient, equivalent notation is to simply state $x \sim \mathcal{N}(\mu_x , \sigma_x^2)$,where σ_x^2 (the square of the standard deviation), is often called the variance.

The value μ_x, termed the population mean, reflects the central location of the normal distribution. The mean is simply the average of all possible *x* values weighted by their probability density, and may be estimated from a finite sample of n observations by obtaining the sample mean, labelled $\bar{x}$, given in Equation 9.

$$\mu_x = \sum x\, p(x) \qquad \cong \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textbf{9}$$

The value σ_x, termed the standard deviation, gives a measure of the spread of the distribution about the mean value, equal to the root-mean square difference between the mean and each possible *x* value, weighted by the probability density. This value may be estimated from a finite number of observations n to yield the sample standard deviation, labelled s_x:

$$\sigma_x = \sqrt{\sum p(x)(x-\mu)^2} \qquad \cong \qquad s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \textbf{10}$$

One reason the normal distribution is assumed by a number of comparative statistical tests is due to the distribution of sample mean values expected to be achieved from repeated experiments. This relationship is expressed by the central limit theorem,[169] which roughly states that as the sample size *n* of each of the individual experiments increases (in which *n* many observations of the variable of interest *x* are made), the distribution of the sample means ($\bar{x}$, estimated from each of the repeated experiments) approaches a normal distribution $\mathcal{N}(\mu_x , \sigma_x^2/n)$. That is, a normal distribution centred about the population mean of the observed variable *x*, with standard deviation in the estimated sample means of $\sigma_x/\sqrt{n}$, known as the "standard error" (SE) in the mean value of *x*.

$$\bar{x} \sim \mathcal{N}(\mu_{\bar{x}} , SE(x)^2) \quad ; \quad \mu_{\bar{x}} = \mu_x \quad , \quad SE(\bar{x}) = \frac{\sigma_x}{\sqrt{n}} \qquad \textbf{11}$$

This is true regardless of whether the distribution of the sampled variable *x* is normal. However if *x* is not normally distributed, larger values of *n* are required before the distribution of expected mean estimates is sufficiently normal.

Measures of skewness and kurtosis may be used to quantify departure from a normal distribution, with skewness roughly expressing a difference between the mean value and the centre of the distribution (the median)[170] and kurtosis roughly expressing non-normal proportions between the centre and tails of the distribution.[171] Equations commonly used for calculating skewness (γ_1) and excess kurtosis (γ_2), both of which are zero for the normal distribution, are given by the following:

$$\gamma_1 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^3}{{\sigma_x}^3} \right) \quad \textbf{12}$$

$$\gamma_2 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^4}{{\sigma_x}^4} \right) - 3 \quad \textbf{13}$$

Skewness may be used to calculate the sample size required for expected sample means to be sufficiently normally distributed. Cochran's rough guideline[172] of 25×*skewness* and Boos and Hughes-Oliver's[173] suggestion of $(5.66 \times skewness)^2$ for a two-sided test may be used to justify common rules of thumb that sample sizes greater than 25 – 30 observations are usually sufficient for moderately skewed or unskewed data (presuming the magnitude of skewness is 1 or less). Meeting this sufficient sample size implies the distribution of expected sample means obtained from repeated experimentation is approximately normal, and therefore "confidence intervals" of the true population mean's likely location may be estimated based on the sample mean and standard deviation estimates. Kurtosis and skewness are both individually used to assess departure from the normal distribution here, in conjunction with their standard error values (used as described in the next section, 2.4.3).

Three main statistical tests are implemented here to test for normally distributed data; the Kolmogorov-Smirnov test,[174-176] the Shapiro-Wilk test[177] and Rasch measurement based chi squared tests output by Rasch measurement software for each facet. All three of these methods test the null hypothesis that the values or measures observed or estimated are sampled from a normal distribution. Probability (p) values reported correspond to the probability that this is true, given the observed values. Visual techniques for the assessment of whether data appear normally distributed also include the use of Q-Q plots,[178] which should appear as a straight line in the case of normally distributed data. With the exception of Rasch-based chi squared tests, the assessments of normality mentioned here are briefly and effectively explained by Ghasemi and Zahediasl.[179] Of note, the Shapiro-Wilk test is recommended as being more powerful than the Kolmogorov-Smirnov test, both of which have low power at small sample sizes.[180] Conversely, as with most statistical tests, larger sample sizes will imply smaller departures from normality are reported as significant. These considerations justify the use of multiple different assessments of normality in this thesis.

2.4.3 Z and T statistics

Discussed in the previous section (2.4.2) was the ability to use the normal distribution to construct confidence intervals in which the observed variable may be observed with a specified probability. This, coupled with standard error values, can be used to obtain the probability that a calculated variable estimate is equal to, less than or greater than a specific value as desired. A convenient technique is to use z values, which converts values of the variable of interest to their location on a scale transformed to be distributed $\mathcal{N}(0,1)$ rather than their original values distributed $\mathcal{N}(\mu_x,\sigma_x^2)$. This is convenient because confidence intervals may then be simply defined by stating the number of standard deviations about the mean the observed value may fall between. For an observed variable *x* expected to be distributed $\mathcal{N}(\mu_x,\sigma_x^2)$, the corresponding z-value may be computed simply as:

$$z = \frac{x - \mu_x}{\sigma_x} \quad \textbf{14}$$

This may be used to test the probability that the observed value of *x* is equal to the expected value μ_x, under the "null hypothesis" that estimates of x are normally distributed about the value μ_x. In a two-sided test, the regions of the normal distribution outside of $\mu_x \pm z\sigma_x$ are summated, yielding the probability that the observed value of x would be randomly sampled outside of the region of $\mu_x \pm z\sigma_x$ if the null hypothesis were true. A variety of statistics are quoted in this thesis along with their standard error values and z tests like the above may be used to test their difference from, or equality to, specified values in this way.

Typically, values of μ and σ are not known, but rather estimated from sample data. In this case the procedure for computing z stated in Equation 14 produces a variable which approaches normality as larger samples are used for the estimates, but strictly speaking is not exactly normally distributed. The variable is instead said to follow a t-distribution of a specific number of "degrees of freedom". The degrees of freedom directly relate to the sample size used to estimate the relevant values, with the t-distribution approximating a normal distribution more closely as sample size (and therefore the degrees of freedom) increases. The t distribution may be used in a similar manner to the z statistic for computing the probability that an observed value x, estimated from a sample of size n, is equivalent to some expected value k. Here, the test statistic follows a t distribution with degrees of freedom n-1.[77] As previously discussed with reference to the central limit theorem, the standard deviation in the sample estimate distribution is termed the standard error value, and hence takes its place in Equation 14.

$$t = \frac{x - k}{SE(x)} \quad \textbf{15}$$

The z and t distributions may thus both be used for statistical comparisons, with the z statistic being appropriate when the sampling distribution is known to be sufficiently normal, and the t statistic appropriate more generally. A key example is the testing of skewness and kurtosis values, which are quoted in this thesis alongside their standard errors to evaluate whether data are normally distributed. Approximations to the standard error in a proportion also exist, and this may be used to test whether an observed proportion ($\hat{p}$) is equal to an expected value.

$$SE(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n} \quad \textbf{16}$$

The normal approximation to the standard error in a proportion $\hat{p}$ estimated from a random sample of size n is given by Equation 16, which is often stated to be approximately accurate provided $\hat{p}$×n and ($\hat{p}$-1)×n are both greater than 5. Despite the prevalence of this approximation, there are superior methods, however.[181] A more accurate formulation is the use of the Wilson score interval,[182] whereby a desired value of z may be input into Equation 17 to yield the upper and lower bounds of the desired confidence interval of the proportion observed.

$$\frac{1}{1 + \frac{1}{n}z^2}\left(\hat{p} + \frac{1}{2n}z^2 \pm \sqrt{\frac{1}{n}\hat{p}(1 - \hat{p}) + \frac{1}{4n^2}z^2}\right) \quad \textbf{17}$$

Generally, observed values greater than 1.96 standard deviations away from their expected value (therefore $|z|>1.96$) reject the null hypothesis at $p<0.05$ in accordance with the normal distribution. Other statistics are also commonly converted to their "z-standardised" equivalents for easy interpretation. Infit and outfit statistics, quoted commonly in Rasch analysis, are good examples (see section 2.5.2.1).

It is also possible to use the normal distribution approximation to compare one value to another, if the sampling distribution of both has been estimated. The difference between two observed variables, x_1 and x_2 estimated from independent samples, may be evaluated using Student's t-test. The test makes use of the standard error in the two values to compute the probability that x_1-x_2 is equal to zero, thereby testing the null hypothesis that the two values are equal. As before, because the standard error estimates in each value are dependent on the sample sizes used to estimate the values in question, the t distribution (from which p values are computed) varies depending on its "degrees of freedom" (df). Though forms of the t-test exist which presume the variances (and hence standard deviations) about the two values compared to be equal,[77] statistics literature recommends the unequal variances form of the t-test, also known as Welch's test[78, 79] unconditionally when sample sizes are unequal, forgoing common tests of equality of variance.[183, 184] The test has degrees of freedom given by the Welch-Satterthwaite equation,[80] where SE(x) is the standard error in x, given by the usual relation between standard deviation and sample size in the case x is a mean value estimated from a sample of n observations (Equation 11).

$$t_{(Welch)} = \frac{x_1 - x_2}{\sqrt{SE(x_1)^2 + SE(x_2)^2}} \quad ; \quad df = \frac{(SE(x_1)^2 + SE(x_2)^2)^2}{SE(x_1)^4/(n_1 - 1) + SE(x_2)^4/(n_2 - 1)} \qquad \textbf{18}$$

An extension of this technique to test the equality of more than two values simultaneously is one way ANOVA.[81, 82] This method tests the equality of the "within group variance" and the "between group variance", each of which have their own associated degrees of freedom, using an F statistic; equal to the t statistic squared in the case of only two values being compared. This typical technique again presumes equal variances of the individual samples compared, however. An alternative which does not make this presumption is Welch's ANOVA[83]. These techniques are all applied in various contexts within this thesis. In each case, however, the p-values reported reflect the probability that the values being compared are equivalent.

2.4.4 Chi squared statistics and nonparametric comparisons

Variables which follow chi squared (χ^2) distributions are common within Rasch analysis and within many other general statistical methods. A variable composed of the sum of the squares of *k* independent standard normal variables (i.e., squared z values) follows a chi squared distribution with *k* degrees of freedom. These statistics are often used to assess degrees of fit and to quantify the extent of deviation from a predicted or expected value, often across entire models or data sets. This may be the assessment of the fit of observed data to the Rasch model (see for example infit and outfit values or global fit statistics discussed in section 2.5.2) or the fit of the distribution of observed measure estimates to a hypothesised trend. Key examples of the latter include the "random(normal) chi square" and "fixed(all same) chi square" values reported by Rasch measurement software, which test whether the measures estimated for the elements of a specific facet appear to be randomly sampled from a normal distribution or whether they appear to be of equivalent values respectively. Probability values can be obtained from chi squared values using the relevant number of degrees of freedom

(similar to other statistics discussed) and used to perform hypothesis tests. In all cases, the probability value obtained corresponds to the probability that the obtained chi squared value be observed under the null hypothesis that the observed data is equivalent to the expected model or model being tested against (such as the normal distribution, Rasch model predictions etc). Chi squared distributed values can also be derived from the likelihood that the obtained data is observed given a specified model, as discussed in sections 2.5.2.2 and 2.5.4.1. Detailed discussion of a variety of applications of Chi squared statistics and the appropriate associated methodology is provided by Delucchi.[66]

Chi squared values can also be converted to approximately standard normal (z) statistics via the Wilson-Hilferty transformation.[185] If Y is a statistic following a chi squared distribution with degrees of freedom df, then the corresponding approximately standard normal (z) value is given by the transformation (W) shown in Equation 19. This transformation is often performed on chi squared values used to assess fit to the Rasch model (see section 2.5.2.1).[126]

$$W(Y) = \frac{\left(\frac{Y}{df}\right)^{1/3} - \left(1 - \left(\frac{1}{9}\right)\left(\frac{2}{df}\right)\right)}{\sqrt{\left(\frac{1}{9}\right)\left(\frac{2}{df}\right)}} \qquad \textbf{19}$$

A particular application of variables following the chi squared distribution is the chi squared test of independence, which tests the null hypothesis that one categorical variable is independent to another. The test is conducted by structuring a "contingency table" with each column corresponding to a specific category of one variable (the column variable) and each row corresponding to a specific category of another (the row variable). In each cell of the table is the number of sampled data points observed in the relevant category of both the appropriate row and column variable. The test uses the observed numbers of responses to generate expected values for each cell in the event the row variable and column variable are statistically independent (i.e. the expected frequencies under the null hypothesis). The difference between these expectations E and the observation O are then used to generate a χ^2 value. From this value the probability that the row variable is statistically independent to the column variable is obtained. For each observed count $O_{i,j}$ in row i and column j of the table, in a table with r rows and c columns, with N total observations, the calculation of expected values $E_{i,j}$ and the test statistic χ^2 are given as follows, where the chi squared value has degrees of freedom (r-1)×(c-1).[69]

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N} \quad ; \quad \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \qquad \textbf{20}$$

One notable case in which the chi squared test of independence is performed in this thesis is in the application of Mood's median test,[186] which tests whether the frequency of observed data points either above or below the "grand median" (the median taken using all sample groups compared) is independent of the sample group. This provides a non-parametric means to test for difference between the central point of two distributions of observed data. The chi squared test of independence may be improved upon, however, in the case that either the row or column variable only has two possible categories. Whilst the above technique only yields approximate probability values, Fisher's exact test[187] yields exact probability values, though is more computationally intensive. The exact test is at times utilised in this thesis, notably in the

comparison of categorised responses to open response format ASLE survey items between sample groups (see section 3.1).

Other non-parametric methods of contrasting two data sets include the Mann-Whitney U test,[72] equivalently known as the Wilcoxon Rank Sum test,[70, 71] and their extension to more than two data sets being compared, the Kruskall-Wallis test.[74] These non-parametric methods utilise methods of assigning ranks to observed data points and do not make assumptions of normally distributed data. This makes them useful alternatives to the parametric methods described previously when testing for equality between two or more independent samples of data.

2.4.5 Correlation and linear models

Statistics of the normal distribution are used commonly in structuring models associating two or more variables. Given two variables x and y, modelled as being related by some mathematical function $f(x_i) = \hat{y}_i$ (where $\hat{y}_i$ is predicted value of the ith observed y value, y_i corresponding to the ith observed x value, x_i), a commonly quoted statistic is the coefficient of determination, labelled R^2.

$$R^2 = 1 - \frac{\sigma^2_{residuals}}{\sigma^2_{total}} \quad ; \quad \sigma^2_{residuals} = \sum_i \frac{(y_i - \hat{y}_i)^2}{n} \, , \qquad \sigma^2_{total} = \sum_i \frac{(y_i - \bar{y})^2}{n} \qquad \textbf{21}$$

As can be seen in Equation 21 above, the R^2 value subtracts the proportion of the total variance in y unexplained by the model from 1, yielding the proportion of observed variance in y explained by the model. In the case x and y are related by a linear model, the coefficient of determination is equal to the square of Pearson's correlation coefficient.[188] Pearson's correlation coefficient[84] ($\rho_{x,y}$) is calculated as:

$$\rho_{x,y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad ; \quad \sigma_{xy} = E\big[(x - \mu_x)(y - \mu_y)\big] \quad \cong \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \textbf{22}$$

where σ_{xy} is the covariance of x and y. This value, much like the mean and standard deviation, may be estimated from a sample of finite size n as shown in Equation 22. The E() operator represents the "expected value": in this case the mean value of the term within the parentheses. The sample correlation coefficient, which approximates $\rho_{x,y}$ whilst maintaining its relationship to R^2, may therefore be calculated as follows:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \qquad \textbf{23}$$

In the case it is required to statistically compare sample correlations to specific values, or to other correlations, the Fisher z transformation may be performed.[189, 190] The value of r_{xy} is transformed to a new value which approximately follows a normal distribution as shown in Equation 24. The Fisher transformed correlation may therefore be compared with specific values or other Fisher transformed correlations, using its standard error of $1/\sqrt{(n-3)}$.

$$F(r_{xy}) = \frac{1}{2} ln \left| \frac{1 + r_{xy}}{1 - r_{xy}} \right| \, , \qquad F(r_{xy}) \sim \mathcal{N}\left(F(\rho_{xy}), \frac{1}{n-3}\right) \qquad \textbf{24}$$

In simple linear regression, the sample correlation coefficient (Equation 23) is of use in computing the line of best fit expressing a modelled linear relationship between variables x and y. Given a set of observed values of x and y, the linear relationship which expresses the i^{th} estimated y value ($\hat{y}_i$) from the i^{th} x value (x_i) may be formulated as shown in Equation 25.[188, 191, 192]

$$y_i \quad \cong \quad \hat{y}_i = a + bx_i \quad ; \qquad b = r_{xy}\frac{s_y}{s_x} \quad , \qquad a = \bar{y} - b\bar{x} \qquad \textbf{25}$$

The standard error in the slope of the line b is also known, enabling comparative statistical tests between the slope and a specified value, or between two different estimated slopes. The degrees of freedom for t-tests (see section 2.4.3) comparing a slope value (estimated from sample size n) to a specific number and for comparing two slopes to each other (estimated from sample sizes n_1 and n_2 respectively) are n-2 and n_1+n_2-2 respectively.

$$SE(b) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum(x_i - \bar{x})^2}} \qquad \textbf{26}$$

2.4.6 Factor analysis

Linear models may be formulated not only for the association of two variables, but also for multiple observed variables. Methods of formulating linear associations between a set of observed variables include principal components analysis (PCA) [193-195] and factor analysis.[196-198] Though similar, a number of key differences exist between PCA and factor analysis,[199] the most notable of which is that unlike the results of factor analysis, the results of PCA may not be interpreted as representing an underlying construct of the data.[200, 201] Rather, PCA serves as a data reduction technique. In PCA, the set of responses to an array of observed variables are reduced to a smaller number of principal components, where each component is a linear combination of the initial observed variables. Factor analysis, however, expresses the observed variables as a linear combination of underlying factors. Additionally, PCA accounts for the totality of observed variance in the observed data, whereas factor analysis only accounts for the shared variance between observed variables.[202] In keeping with the objectives of this thesis, which involve exploring the factors underpinning observed ASLE survey responses, factor analysis is the more appropriate of these alternatives in this instance.

For an array of observed cases (indexed by i), each consisting of observations of N different variables (indexed by n), the factor model seeks to explain each observed variable as a linear combination of F underlying factors, where F<N. This is achieved by converting the observed variables into their z standardised forms (see section 2.4.3), then fitting the data to Equation 27 below. The value of $z_{n,i}$ corresponds to the z standardised form of the i^{th} case of the n^{th} variable ($x_{n,i}$), whilst each value of $\varepsilon_{f,i}$ represents the measure of the f^{th} underlying factor in the i^{th} observed case. The $l_{n,f}$ values, termed "factor loadings", weight the contribution of the f^{th} factor to the n^{th} observed variable. The $E_{n,i}$ parameter is simply an error term, equivalent to zero for a perfectly fitting model.

$$z_{n,i} = \sum_{f=1}^{F} l_{n,f}\varepsilon_{f,i} + E_{n,i} \qquad \textbf{27}$$

A number of techniques are available for factor "extraction". That is, techniques used to isolate the underlying factors of the model and estimate their loadings. These techniques include (but are not limited to) canonical factor analysis[203] (which utilises the same model as PCA), common factor analysis,[204] alpha factoring[205] and image factoring.[206] Many extraction methods may result in nonsensical so-called "Heywood cases", where the modelled factors are computed to explain more than 100% of the variance in the observed data in some instances.[207] This does not occur for image factoring, however.[208] The number of factors extracted during the analysis may be specified by the researcher or may be selected based on various statistical considerations[209] such as eigenvalues[210] (not recommended) or scree plots.[211] Following extraction, "rotation" methods may be employed in order to reformulate the extracted factors in a manner more easily interpreted by the analyst.[212] Orthogonal rotation methods, which maintain that estimated factors are uncorrelated, include varimax[213] and quartimax.[214] Oblique rotation methods, which permit factors to be correlated, include direct oblimin[215] and promax.[216] Rotation methods are not limited to those listed here.

Two common statistical techniques used to assess the adequacy of correlated sample data for factor analysis are Bartlett's test of sphericity and the Kaiser-Myer-Olkin (KMO) measure of sample adequacy.[217] KMO measure of sample adequacy[218, 219] ranges from zero to one and is deemed unacceptable at values below 0.5, whilst Bartlett's test of sphericity[220, 221] tests the null hypothesis that all variables are uncorrelated. The rejection of this null hypothesis implies the data are appropriate for factor analysis.[221]

2.5 Data analysis: Rasch model related statistics

2.5.1 Observed, expected and fair scores

If implementing the traditional integer scoring methods applied to the ASLE survey data, the observed average score A corresponding to a set of observed counts in each of the K many available response categories (described by the vector $\vec{X}$ = [c_1 , c_2 , … , c_K]) may be computed using the following:

$$A(\vec{X}) = \frac{1}{N}\sum_{k=1}^{K} a_k c_k \qquad \textbf{28}$$

Where a_k is the score value assigned to the k^{th} response category, whilst N is the total number of observations: the sum of the c_k values. Usually, the five score values utilised for the five ASLE item response categories are the integer values from -2 to 2, however Rasch measurement software often reports score results using a scoring system beginning with the value zero for the first category, then proceeding with successive integer values for progressively higher categories. This is not problematic, as the usual ASELL-type score may be obtained directly from the score reported by Rasch software by subtracting 2, however it is worth noting for the purposes of reading and interpreting ”observed average scores” as reported by Rasch measurement programs.

Once a Rasch model has been estimated for the data, expected average response scores associated with individual persons, items or elements of other facets may be computed. This is achieved by taking P(X = x_k) values obtained from the Rasch model directly (Equation 1) to obtain the expected count of responses in each category c_k, then applying Equation 28 to obtain a mean score based on expected response counts, rather than observed response counts. A value labelled as the “fair average” score may be also reported in the case of using the *Facets* program, and this is approximately equivalent to taking Rasch measures of all other facets as being their average value, then applying this technique.[222]

The point-measure correlation gives the observed correlation (see section 2.4.5) between observed scores and the associated Rasch measures.[223] Values of these correlations expected under the presumption that the data perfectly fit the Rasch model can also be computed and used to assess whether the observed values are excessively high or low. The observed and expected correlation values between score and measure are often labelled as “ptmea” and “ptexp” respectively in *Winsteps* or *Facets* outputs.

2.5.2 Rasch model fit statistics and descriptive values

2.5.2.1 Infit, outfit and discrimination

A variety of statistics are available for describing the fit of data to the Rasch model. These include both local and global fit statistics including Infit (inlier-fit) and outfit (outlier-fit) values. Infit and outfit statistics[76, 77] provide measures of how closely the data fit model predictions, with respect to inlying and outlying observations respectively.[161] They are computed first by contrasting the observed response score (X_n) for each individual data point (indexed by n) and contrasting with the expected mean response (E_n) for that single data point, computed using procedures described previously (see section 2.5.1). The standard deviation in the expected mean value (S_n) is also calculated, using the population standard deviation formula (Equation

10, section 2.4.2) and the expected frequencies of scored responses in each category for that data point. The observed score, expected mean score and standard deviation in the expected mean are then used to obtain a z score (see Equation 14, section 2.4.3), termed the "standardised residual" value for that data point (Z_n). Infit and outfit mean square values ($MnSq_{INFIT}$ and $MnSq_{OUTFIT}$ respectively) are then computed using these standardised residuals, which are evaluated for each data point relevant to the facet element number the statistics are being quoted for (for example all data points contributed to by a specific person, specific survey question, etc, of which there are N).

$$MnSq_{OUTFIT} = \frac{1}{N}\sum_{n=1}^{N} {Z_n}^2 \quad , \quad MnSq_{INFIT} = \frac{\sum_{n=1}^{N}(X_n - E_n)^2}{\sum_{n=1}^{N} S_n} \qquad \textbf{29}$$

Values quoted for infit and outfit may be mean-square values calculated as shown above, or may be corresponding standard normal z values obtained using the Wilson-Hilferty transformation[139, 224](see sections 2.4.3 and 2.4.4, Equation 19). Because the mean square values are chi squared statistics (see section 2.4.4) divided by their degrees of freedom, they therefore have expected values of 1. The corresponding z values may be interpreted as per other z statistics.

Values substantially above expectation (termed "underfit") may imply measures do not behave in the manner presumed by the construct of the Rasch model generated, indicating inconsistency between prediction and observation; a key component of construct validity. Conversely, values below expectation imply data accords with model predictions so closely as to be unexpected (termed "overfit").[162] The z statistic may be preferable to the use of the mean square value in some cases due to the mean square's insensitivity to variance in the measures.[225] However, z statistics reflect the statistical significance of the departure from expectation rather than its magnitude, and as such studies with vast numbers of responses may show significant values of z despite very small deviations from Rasch model prediction.[226] This justifies use of the mean-square values, however this is also problematic as there are no clear 'rules' defining which values are extreme and which are acceptable (though ranges of roughly 0.6 to 1.4 would be acceptable for ASLE survey purposes).[227] Generally, z statistics are useful for determining whether data fit the model perfectly, whilst mean squares are useful for determining whether data fit the model 'usefully'.[228] A statistic related to the infit and outfit values is the RMSR or root mean square residual, where the residual is equivalent to the difference between expected scored response based on the Rasch model and observed scored response.

The estimated discrimination values reported are in this instance best interpreted as a measure of model fit: the Rasch model uses discrimination values of 1, and these statistics describe what value the discrimination would take were this variable allowed to change. It expresses the degree of change in observed response relative to the change in the latent trait variable. Values below one suggest that observed responses change less drastically than expected as the latent trait (Rasch) measure changes, whilst values above one indicate observed responses change more drastically than expected as the latent trait measure changes. The latter may be indicative of a second, undesirable variable which correlates to the latent trait of interest influencing responses.[229]

2.5.2.2 Global fit and variance explained

Other statistics reflecting the fit of observed data to the model are the log-probability value and the log-likelihood chi-square value. The log-probability value is equivalent to the natural logarithm of the probability of sampling the relevant observed data points given the estimated Rasch model parameters, whilst the log-likelihood chi square value is simply -2 $\times$ the log-probability value.[230] Both of these statistics may be quoted locally for individual data points, for measures relevant to a large set of data points, or for the Rasch model as a whole. The log-likelihood chi-square value is a common measure of global fit, and a key component of statistical tests contrasting the fit of different Rasch models of the same data (see section 2.5.4). It may also be used to test whether the observed data fits the Rasch model perfectly, though perfect fit is generally unexpected.[231] The value is approximately chi-squared distributed with degrees of freedom approximately equal to the number of free parameters estimated subtracted from the number of (non-extreme) data points in the analysis.

Rasch measurement software is also capable of reporting the proportion of variance in observed data explained by the Rasch measures estimated via variance decomposition techniques.[232] This can be useful in assessing the degree to which various specific facets (such as persons or items) or the Rasch model as a whole contributes to variation in the observed data. This proportion is influenced by a variety of factors such as the inherent imprecision in the categorical response scale, the relative degree of variation in some facets compared to others, random error and other factors, so care must be taken in drawing conclusions from these values.

2.5.2.3 Separation and reliability

Each facet in a Rasch model may be assigned separation and reliability values. The separation statistic is related to the number of statistically different performance strata identifiable in the sample,[233] whilst the reliability value (which ranges from zero to one, one being optimal) provides a measure of the reproducibility of the observed order of measures estimated.[234, 235] These values improve given an increased number of data points.[236]

In the context of the usual two-facet Rasch models (see Equation 2 in section 2.2.1), low person separation implies the hierarchy of person measures cannot be distinguished well given the data available, and low item separation generally implies that the person sample is not large enough to confirm the item measure hierarchy. If the reliability value is low this can be improved chiefly by increasing the sample size, by making the survey instrument better targeted (the mean person measure and the mean item measure are close) or by gathering data from a broader range of the other measure. For example, item measure reliability can be improved by gathering data from a group of persons with a broader range of person measures, or conversely, person measure reliability can be improved by gathering data from items with a wider range of item measures.

2.5.2.4 Rating scale associated statistics

Both *Winsteps* and *Facets* report a range of statistical values associated with the rating scale structures estimated. Each Andrich threshold listed corresponds to the τ parameter (see Equations 1 and 2) defining the point at which a student is equally likely to respond in either of the two relevant adjacent categories, whilst the Thurstone threshold is the measure at which a student has a 50% probability of responding in the lower of these two categories or below, and a 50% probability of responding in the upper category or above.[237]

The category measure is the point on the latent trait variable at which the expected score (see section 2.5.1) is equal to the assigned score for that category. These values are useful for assessing the equidistant nature of the categories or otherwise, as are the category "ranges", which are defined at the "half point thresholds": the points at which the expected score is equal to the average of the two assigned scores of adjacent categories. Also reported may be the observed average latent trait measure for observations in that category, or the expected value of this statistic presuming the Rasch model (labelled as observed and expected average measure respectively).[238] The *Facets* program also reports the category peak probability, which is simply the measure at which the probability of observing that category is at its maximum.

The coherence statistics are also available, and describe the ability to draw inferences between the observed response category and the Rasch measure for the latent trait variable. The C=>M (category implies measure) value describes how frequently the observed response category correctly predicts the latent trait measure, whilst the M=>C (measure implies category) value describes how frequently the latent trait measure correctly predicts the observed response category.[239]

2.5.3 Analysis of bias interactions

Differential Item Functioning (DIF) and Differential Person Functioning (DPF) are examples of bias evident in data modelled using Rasch analysis. DIF refers to cases where items appear to adopt significantly different measures for different persons or person groups, whilst DPF refers to cases where person measures appear significantly different when estimated using different items or sets of items. Similar bias interactions can also occur between other facets in the case of many facet Rasch models.

Two statistical techniques of assessing bias interactions such as these are available in the *Winsteps* and *Facets* programs. Mantel statistics,[240] a multiple response category extension of Mantel-Haenszel statistics,[241] are computed via a procedure of dividing the data into strata based on the measures associated. In the case of incomplete data these statistics are less accurate and in some cases not able to be estimated, with alternative methods therefore preferable.[242] Both *Facets* and *Winsteps* also provide alternative Bias analysis statistics using a Rasch-based methodology. These statistics, reported as the results of t-tests (see section 2.4.3), have been shown to be superior to Mantel-Haenszel statistics.[243-246] For this reason, and also due to the high prevalence of missing data in the ASLE survey responses, these t-test statistics were used to assess bias interactions in the studies described in this thesis. The probability (p) values reported by these tests correspond to the probability of obtaining the observed data under the null hypothesis of no bias. Reported alongside these likelihood based statistics are group-level fit statistics,[247, 248] which test the hypothesis that observed responses to entire groups of facet elements (for example groups of persons) accord with Rasch model expectations.

Whilst bias analyses such as DIF may determine that bias is present, a phenomenon known as 'artificial DIF' exists whereby the presence of bias in one item results in statistical tests reporting bias for other items artificially. This means that whilst bias analyses are useful for detecting the presence or absence of bias in a facet generally, it may be unclear which specific elements of the facet (eg. which persons or which items) take different values as a result of bias interactions and which do not. Differentiating 'real' from 'artificial' bias interactions may be achieved by "resolving" the differential measures.[249]

For example, a series of items may appear to exhibit DIF depending on which of two person groups is responding. The item exhibiting the most significant DIF may be 'resolved' by assigning it two measures as opposed to one: one measure applicable to each person group. This is equivalent to treating it as two separate items: one item for which only one person group responds, and a second item for which only the second person group responds. Following this resolution, any artificial DIF induced by this item's real DIF would no longer be evident. It is therefore possible to differentiate between real and artificial bias interactions in this way, though the analysis required to do so completely may be extensive.

2.5.4 Model selection

2.5.4.1 Variance explained and the likelihood ratio test

Given two alternate models of the ASLE survey data, the need may arise to contrast the models' capability to explain the observed data. In the case of nested models, those where one model formulation is obtained by restricting the parameters of the other (the parent model), the likelihood ratio test may be used to contrast the proportion of observed variance explained by the two models. The test statistic:

$$D = \chi_1^2 - \chi_2^2 \qquad ; \qquad \chi^2 = -2\ln(\mathcal{L}) \tag{30}$$

where χ_1^2 and χ_2^2 refer to the log-likelihood chi square values of the parent and restricted model respectively and $\mathcal{L}$ is the likelihood of observing the data given the estimated model parameters, approaches a chi squared distribution with degrees of freedom equal to the difference between the degrees of freedom of the two original χ^2 values ($df_{\text{parent model}} - df_{\text{restricted model}}$) as more data points are included.[250] This may be useful to test the null hypothesis that both models explain the same proportion of the observed data. In general, the degrees of freedom associated with log-likelihood chi square values are given by *df = n-k*, where n is the number of data points, all of which must be common to both models and k is the number of free parameters estimated.

"Free" parameters are those for which values are estimated rather than mathematically necessitated. In general, the free parameters associated with a model include one parameter for each element of each facet, minus the number of "centred facets" (those for which the average measure is defined as zero, therefore meaning one element's measure is the negative sum of the measures for the other elements), plus the number of free parameters estimated in calculation of the Andrich thresholds. As the Andrich threshold values for each scale group are defined as summating to zero and threshold values exist between categories, the number of free parameters estimated for each scale group is two less than the number of rating scale categories.

Though often applied to establish that no explained data is lost when applying the LLTM in place of less simplified Rasch models,[144, 251] a limitation of the likelihood ratio test is that it does not take into account the number of parameters required to achieve the observed proportion of data explained by the models proposed. Estimating a larger number of parameters in a model will invariably explain a greater proportion of the variance in observed data points, even if those extra parameters do not reflect any genuine trends in the data previously unaccounted for. Additional parameters allow models to fit better to random "noise", thereby reporting a higher proportion of variance explained. An improvement to the

use of the likelihood ratio test is therefore the implementation of a procedure which takes the parsimony of the model proposed into account.

2.5.4.2 Parsimony and the corrected Akaike Information Criterion

The Akaike Information Criterion,[252] and its corrected form[253] (AICc) provide a means of selecting the best explanatory model of the observed data in a statistical manner. The statistic takes into account the global fit of the model to the data (in the form of the log-likelihood chi square value) as well as the parsimony of the model (related to the number of free parameters needed to be estimated) to yield an AICc value for each proposed model. The model with the lowest AICc value is taken to be the best explanatory model for the data, for reasons outlined well by Burnham and Anderson.[254] Specifically, the difference in AICc value from that of the minimum AICc model (ΔAICc) corresponds to the Kullback-Leibler information[255] loss experienced if working under the alternative model rather than that with the lower AICc, whilst *exp(ΔAICc/2)* yields the likelihood of the proposed model given the data, relative to the best model proposed. In comparison to the lowest AICc model, alternate models with ΔAICc ≤ 2 have "substantial" comparative support, alternate models with 4 ≤ ΔAICc ≤ 7 have "considerably less" support, whilst those with ΔAICc ≥ 10 have "essentially no support", irrespective of the actual magnitude of AICc values being compared.[254] The formula for the corrected Akaike Information Criterion is given in Equation 31, where $\mathcal{L}$ is the likelihood of the estimated parameter values given the observed outcomes (equivalent to the probability of the observed data given those parameter values), k is the number of free parameters estimated in the model and n is the number of data points used.

$$AICc = -2\ln(\mathcal{L}) + 2k + \frac{2k(k+1)}{n-k-1} \qquad \textbf{31}$$

The value of -2ln($\mathcal{L}$) is often labelled as the log-likelihood chi square value and is commonly used as a measure of global fit of the observed data to the Rasch model as described previously (see section 2.5.2.2). The remaining terms in the AICc equation serve to penalise a large number of parameters used to estimate a comparatively small number of data points, meaning sufficient parsimony is a key factor in determination of the best model. The statistic has been applied to the selection of appropriate Rasch models previously.[150]

3 *Quantitative methods and the ASLE survey data*

In this section the application of common integer value scoring techniques to ASLE survey data is explored from a range of perspectives, contrasting the conclusions able to be drawn with those which would be suggested through Rasch modelling. Results presented in this section collectively serve to test the first primary hypothesis of this thesis:

> ***Hypothesis 1:***
> *Conclusions drawn from the ASLE survey data using typical scoring techniques resemble conclusions drawn using sample independent, interval scale measures extracted from the same data.*

Section 3.1 presents a study reminiscent of typical ASLE survey use. Rating scale items of the ASLE survey are analysed with the usual integer value scoring methodology, supporting any conclusions drawn using comments received on open response items of the survey. This score-based study will serve as a point of comparison in the subsequent section.

Section 3.2 includes an in-depth analysis of the identical data used for the previous section, this time using Rasch analysis. The student independence or otherwise of score-based data is particularly highlighted, testing the datasets gathered for evidence of sampling bias. The impacts of these effects on conclusions of the initial score-based study are identified and discussed, in so doing contrasting scoring methodology with Rasch methodology. This study, in conjunction with the previous, serve as a specific example of how any limitations in score-based methods may impact research conclusions.

Much of the data and discussion presented in sections 3.1 and 3.2 have been published (though with some differences) in the Journal of Chemical Education,[256] presenting an investigation contrasting student perceptions of two different technological interfaces used in laboratory activities.

Section 3.3 presents a far more generalised investigation into the measurement properties of integer value scored data. Unidimensional Rasch models are generated for each item of the ASLE survey, using these models to contrast sample independent, interval scale Rasch measures of with the analogous score-based values expected. The relationship between scores and measures is revealed both at the level of individual responses and group level statistics. The validity of applying parametric statistical methods to scored data is also investigated.

3.1 Typical score-based analysis of ASLE survey data: an example

3.1.1 Outline

Rasch analysis, whilst revealing information of much more depth than traditional scoring methodologies, is not readily accessible to all researchers and has not been applied to the ASLE surveys prior to the works presented in this thesis. The study documented in this section serves as an example of the way traditional scoring methods may be applied to a data set, despite their suggested limitations. This study makes use of the usual integer scoring techniques applied to Likert-type items of the ASLE survey, heavily drawing upon qualitative comments received on the same set of survey responses to provide additional support to any conclusions. In this way, pedagogical implications may still be confidently drawn, whilst illustrating a typical study conducted using ASLE survey data. In the section following this study (3.2), Likert-type data used to draw conclusions here will be re-analysed using Rasch analysis. This study and the next will thereby serve to contrast the conclusions of integer scoring methods and Rasch methods, testing the first primary hypothesis of this thesis: that "*conclusions drawn from the ASLE survey data using typical scoring techniques resemble conclusions drawn using sample independent, interval scale measures extracted from the same data".*

Though the primary objectives of this thesis concern validity of ASLE survey methodology and past conclusions, the specific investigation presented in this section as a vehicle for later investigating scoring methodology validity has its own notable implications. Investigation into the validity of the scoring methods implemented is reserved for the section following (3.2), using the conclusions of this study as a baseline for comparison. Consequently, the discussion of results here will exclusively focus on conclusions revealed about student perceptions and effective design of experiments, not the validity of scoring methods used to draw those conclusions. The study presented involves a contrast between two technological interfaces which students may be required to use as part of laboratory activities: a handheld graphing data logger and analogous software installed on a laptop computer. It is conclusions regarding these technological interfaces which will be discussed here, reserving an analysis of the validity of the scoring methodology used for section 3.2.

3.1.2 Background: Microcomputer based laboratories

Laboratory work provides a wide range of benefits for learning in chemistry. [7] Practical laboratory work has historically been claimed to be beneficial for multiple reasons, [5, 6] including exposing students to concrete experiences with objects and concepts mentioned in the classroom. [4, 13] Connecting concrete, macroscopic observations to the abstract representations and symbolisations used in science is well understood to be a key hurdle in the understanding of chemistry concepts[257-260] and is a task hailed as being one of the most difficult challenges facing science teachers, as well as one of the most important.[261] In part to assist in the teaching of abstract concepts, and to engage students with technology they may use as working scientists, technology plays an increasingly large role in science education, including data collection and display in the laboratory setting.[262] A large body of research exists concerning activities in which computing devices are used in conjunction with measurement devices (probeware) to gather and display data in laboratory teaching exercises. Students may be required to use handheld graphing data logger devices[263-265] specifically designed to display and analyse data collected from associated probeware, or alternately the probeware may be

connected to a laptop or desktop computer equipped with the necessary software. [266-269] These activities have been termed microcomputer based laboratory (MBL) activities.

Because of their ability to pair events with their graphical representation in real time, MBL activities have been suggested to assist in making the connection between the concrete and the abstract,[270-272] notably in the form of improving students' graph interpretation skills. More active student engagement in constructing understanding has been suggested as an additional benefit, arising from increased focus on data interpretation instead of data collection, and increased student collaboration.[273, 274] These MBL activities have been said to have the ability to "transform" laboratory activities due to these advantages.[275]

Whilst a number of studies have supported these claims,[265, 276-282] particularly in the context of inquiry based learning,[263, 283, 284] overall the results of implementing MBL activities appear mixed[285] and successful implementation of MBL activities appears to be a complex issue.[286, 287] Studies exist which counter the suggested benefits concerning student understanding of graphs[288] and the connection between the macroscopic and more abstract.[289] Other prominent issues appear to be the possibility that students may watch uncritically as the computer 'does all the work for them', as well as students encountering difficulty using the computing devices themselves.[286, 290]

Studies documenting student perspectives of this technology reveal that student views of MBL vary to a great extent.[291] Negative issues raised again include students feeling disengaged as the data logger does all the work, as well as a lack of technical familiarity.[292] Students are reported to claim the technology is complex and difficult to use[293] and that they do not have the time to 'play' with the technology and undergo trial and error processes of learning like they would do with home computers.[292] These issues, including that students have difficulty manipulating and using MBL technology, notably in "older" forms, have also been recognized in teacher views.[290] The disadvantages of having to learn how to use the technology as well as the exercise's learning objectives has been observed to outweigh the advantages of MBL in the past,[294] and it has been suggested that MBL activities may be better suited to those who have a better idea of both content and the handling of sensor technology associated with the microcomputer devices.[295] This is reminiscent of classroom based studies suggesting students benefit more from computer-based exercises if they are comfortable with their use[296] and it is not unreasonable to expect that the same is true of computing devices in the laboratory setting. Recent review of studies concerning MBL activities in secondary school chemistry suggests more research needs to be conducted to discover what can be done to assist students in overcoming these issues they express.[291]

This study reports differences in student perception data received from two different cohorts of students in their first year undergraduate chemistry laboratory sessions; one cohort using a handheld data logger device to collect and display data in the experiments studied, and the other cohort performing the identical tasks, instead using a laptop computer. This change in technological interface was made in response to negative views expressed by students regarding the data logger devices, gathered using ASLE surveys distributed with original intentions other than this specific study. ASLE surveys were then used to monitor student perceptions the following year after the change had been made, and the results are presented in the following discussion. The observations made are suggested to be of use in moving towards overcoming the reported student difficulties associated with so called "microcomputer based laboratories".

3.1.3 Specific methods

3.1.3.1 Experiments Conducted

During the year 2011 at the University of Adelaide, experiments studied within this research involved student use of the *PASCO Xplorer GLX* handheld graphing data logger device. [264, 265] Student feedback data, collected for other research purposes using the ASLE instrument, revealed negative perceptions of these devices, which had been utilised in the teaching laboratory for a number of years prior. In response to this feedback, the devices were replaced with laptop computers equipped with software replicating the capabilities of the data loggers; *PASCO DataStudio*.[266] The only differences between the two years, aside from this technological interface change, include the laboratory demonstrators and the portion of the instruction manual devoted to use of the technology implemented (in the form of an isolated appendix to the rest of the manual). All other features of the relevant experiments remained identical, including the tasks performed using either the data logger or the laptop computer. A total of three experiments were studied, with the perception of each contrasted between the two years. The utilised data measurement tools and functions of the data loggers or laptops were different in each of these three experiments. Experiments studied include "Vapour pressure", "Biological buffers" and "Determination of copper(II) ion concentration", described previously in Table 2 (section 2.1.3).

3.1.3.2 Data Treatment

Qualitative comments received in response to the open response items on the ASLE instrument (items 15-19) were assigned codes based on their content, and also whether the comment was of a positive, negative or neutral nature. The thirteen content-specific codes used for all survey items except item 16 were pre-established; devised for the purposes of separate research conducted in previous years (unpublished data). Codes used for item 16, in the case this item was used in this research, were devised as appropriate for the specific experiment's learning objectives. Frequencies of comments which were and were not assigned each of these pre-established codes were enumerated for each survey item. Fisher's exact test (see section 2.4.4) was used to compare these frequencies between the two forms of each experiment individually.

Responses to Likert-style items were assigned scores corresponding to their position on the five point scale. Responses to items 1 – 12 of the ASLE instrument were assigned successive integer scores from +2 to -2 ("strongly agree" to "strongly disagree") with zero ("neutral") as the midpoint, +2 being the optimal response. Item 13 responses, concerning time availability, were also scored from +2 to -2 ("way too much" to "nowhere near enough") with zero ("about right") as the midpoint and optimal response. The final Likert-style item, concerning overall learning experience, was similarly scored from +2 to -2 ("excellent" to "very poor") with zero ("average") as the midpoint and +2 as the optimal response. Mean values of response scores received from each student cohort, from each year, for each experiment, for each Likert-type response format item on the surveys were calculated for the purposes of comparison (labelled as m_{2011} and, m_{2012} for mean values from 2011 and 2012 data respectively). This is in line with standard methodology of the ASELL project (see section 1.1.2).

Mean scored responses to the Likert-type response format items were compared between years for each of the three experiments using the T-test for unequal variances (Equation 18, section 2.4.3), using this test in preference to the equal variances test in all cases as recommended in the statistics literature for data sets of unequal sample size. The value of

alpha (α) was selected to be 0.05 for the purposes of statistical testing, and two tailed probability values were obtained for all comparisons made. In order to account for the issue of multiple comparisons and control the family-wise error rate, the unweighted Bonferroni method was applied (see section 2.4.1). All hypothesis tests conducted to compare the same experiment between the two years of study were taken to be of the same family of hypotheses, thereby yielding one family of hypotheses tests for each of the three experiments in the study. Consequently, statistically significant difference was inferred at $p < \alpha/n$, where n is the number of hypothesis tests conducted to compare the relevant experiment's two different forms. Statistical tests and values were calculated using *Microsoft® Excel® 2010*, with the exception of Fisher's exact test, which was conducted using *VassarStats*.[297]

3.1.3.3 Student Cohorts Sampled

Data featured in this study were obtained from students enrolled in the first year undergraduate courses Chemistry IA and Chemistry IB (see section 2.1.1 for a more detailed description). The number of students sampled was variable between experiments and between years, with students presented with the non-compulsory ASLE instrument at the end of their laboratory sessions. Given the sample-dependence of scored ASLE survey data, this sampling technique presents the possibility of introducing bias in the sample: deviating from a true representation of the student population. An investigation of the same data set using more sample independent analysis techniques (Rasch analysis) is to be presented in the following section (3.2) to overcome this limitation in the scoring methods used and evaluate the validity of conclusions drawn (see section 3.2.4.1: Summary of sample adequacy implications).

3.1.4 Results and discussion

3.1.4.1 Improved Perception of Overall Learning Experience

Details of all statistical tests conducted and frequencies of responses received for both qualitative and quantitative items of the survey are presented in the supporting information (section 7.2).

The effects of replacing the data logger interface with the laptop interface appear to have yielded noticeable improvements in student perception of overall learning experience. When responding to the Likert-type item "*Overall, as a learning experience, I would rate this experiment as*", students conducting the vapour pressure experiment responded significantly more positively ($m_{2011} = -0.02$, $m_{2012} = 0.69$, $t(149.0) = -5.06$, $p = 1.22 \times 10^{-6}$). Once the data logger interface was replaced, this experiment in particular showed apparent improvement in a large number of respects judging by Likert-type item responses. These improvements, displayed in Figure 9, will be elaborated upon throughout the discussion. Change in overall perception of the vapour pressure experiment was also clearly evident in the open response comments. When asked if they enjoyed the vapour pressure experiment (item 15), a significantly greater number of students gave positive responses (18 of 55 in 2011, 54 of 79 in 2012, $p = 5.34 \times 10^{-5}$) and fewer gave negative responses (39 of 55 in 2011, 31 of 79 in 2012, $p = 4.09 \times 10^{-4}$). The biological buffers experiment may also have been perceived more positively overall judging by responses to item 14 ($m_{2011} = 0.60$, $m_{2011} = 0.88$, $t(185.3) = -2.52$, $p = 1.26 \times 10^{-2}$), however this result could not be deemed statistically significant when accounting for family-wise error.

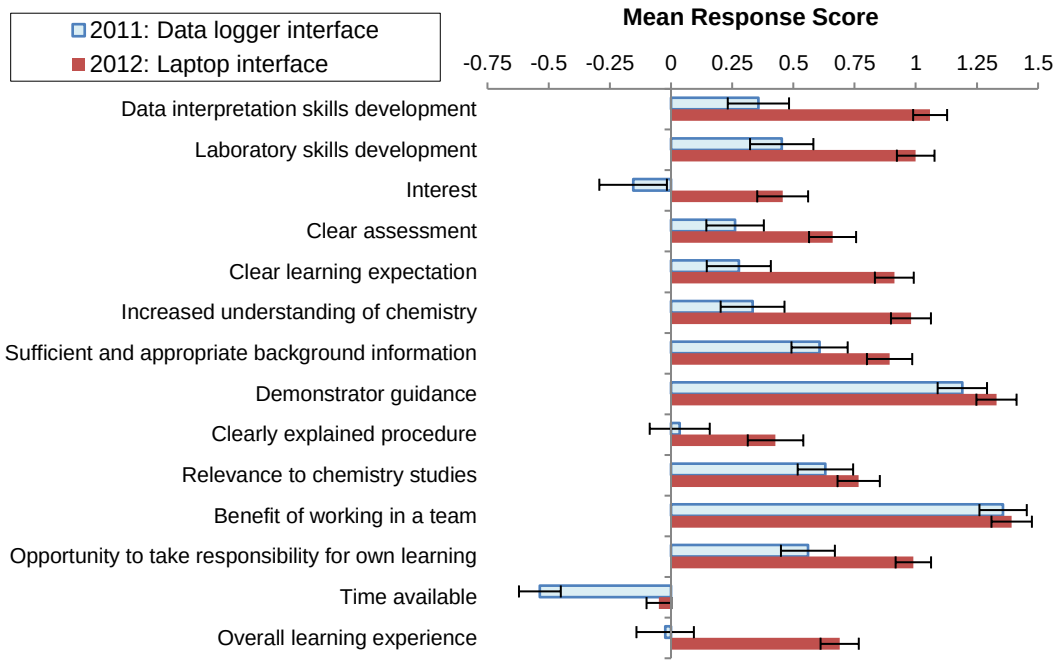


Figure 9: Mean Likert-type item response scores for "Vapour Pressure"

Mean response scores obtained using the laptop interface appear significantly higher than mean scores obtained using the data logger interface. Error bars represent the standard error in the mean value. Further statistical details for these comparisons are available in the supporting information (see section 7.2.1)

In the case of "Vapour pressure" specifically, data presented later in this thesis (section 4.3.3.2, p.126) can be used to show that students not only found the data logger form of the experiment to be poor, but gave noticeably unfair ratings compared to all other experiments (Figure 43, p.127). A drop in approximately 0.5 score units across all ASLE survey questions for this experiment appears to be due solely to negative student bias, adding to the already low "fair" mean score values. Whilst this effect is not revealed by traditional integer scoring techniques alone, it is still supportive of the general pedagogical implications of the study discussed in this section. This large scale negative bias is only evident for this single experiment, and disappears once the data logger is replaced with the laptop.

3.1.4.2 Reduced Negativity about the Equipment

Examining the qualitative comments received for the three experiments, there are a number of indicators that the laptop interface was better received than the data logger interface. During 2011, of the 24 students who stated they did not enjoy the experiment in response to the question *"Did you enjoy the experiment? Why or why not?"*, 21 of those referenced the equipment used. A number of these comments mentioned that the data logger devices were prone to error and difficult to use. Some examples from the biological buffers experiment include:

> *"It was ok, the GLX thing was difficult to use"; "it was annoying using the GLX"; "No. The explorer GLX is hard to use"; "I do not like the pasco"; "no glx do not work well, most people had problems"*

Comments from the vapour pressure experiment for this same question further reveal issues with the use of the data loggers:

"yes however I found the xplorer GLX difficult to use and very frustrating. The video, written instructions and demonstrator could not supply sufficient information on using the xplorer GLX"; "No. I do not like using the explorer GLX"

Once the handheld data loggers had been replaced, students conducting the biological buffers experiment made negative comments about the equipment used significantly less often in response to this same survey item, both amongst all comments received for this item (21 of 51 comments in 2011, 4 of 54 comments in 2012, $p = 5.96 \times 10^{-5}$) and also when considering only those comments which were negative (21 of 24 comments in 2011, 4 of 19 comments in 2012, $p = 2.44 \times 10^{-5}$). However, this was not clearly evident in the other two experiments, potentially due to the fact that different functions of the technology were used in each experiment.

When the question *"What aspects of the experiment need improvement, and what changes would you suggest?"* was asked of students, significantly fewer equipment related negative comments were received for the biological buffers experiment (25 of 35 comments in 2011, 6 of 32 comments in 2012, $p = 2.16 \times 10^{-5}$). There was also some indication of a similar effect in the copper(II) ion concentration experiment, (18 of 43 comments in 2011, 12 of 66 in 2012, $p = 4.61 \times 10^{-2}$) however this could be attributed to family-wise error. Amongst the negative comments which were received for this question about the copper ion experiment, the equipment was mentioned significantly less frequently amongst the improvements listed (25 of 31 negative comments in 2011, 6 of 25 negative comments in 2012, $p = 3.38 \times 10^{-5}$).

In the case of the vapour pressure experiment, comments viewing the equipment negatively could not be said to differ significantly. However, there appeared to be a positive shift in the number of students who found the equipment appealing, often due to the new and unfamiliar use of technology (in this context, 'unfamiliar' refers to students not having seen technology used in this manner in the laboratory previously). In response to item 15, the level of familiarity or unfamiliarity was mentioned more frequently as a reason for liking the experiment (1 of 55 total comments in 2011, 15 of 79 comments in 2012, $p = 2.20 \times 10^{-3}$). Although this difference could be attributed to family-wise error, these comments almost exclusively made very positive reference to the enjoyable experience of the novel use of technology, and this was something not seen when using the data logger interface.

Additionally, these comments cite the ability to watch the graph change in real time in a number of cases. Some examples of most enjoyable aspects of the experiment students described include:

"Working with the computer program to see live data of changes in temp (of water bath) and vapour pressure"; "watching the graph on the screen as it changed"; "recording the pressure of a system onto a real time computer graph".

These comments are reminiscent of some of the benefits suggested in the literature of the use of microcomputers in laboratories, and no comments referenced this until 2012 when the data logger devices were replaced with the laptop computers. Citing the equipment as the most enjoyable aspect of this experiment (in response to item 17), though increased, did not increase to a level able to be deemed statistically significant (14 of 43 total comments in 2011, 35 of 79 comments in 2012, $p = 7.82 \times 10^{-2}$).

3.1.4.3 Greater Sense of Understanding

A broad range of indicators across all three experiments show evidence that students who used the laptops felt a greater sense of understanding the experiments than those using the data loggers. The vapour pressure experiment in particular shows clear improvement. Significant positive changes in the mean response scores for items 5 and 6 were observed in this experiment, providing reason to suspect a potential increase in student perception of clarity of the expected learning outcomes ($m_{2011} = 0.28$, $m_{2012} = 0.91$, $t(138.4) = -4.15$, $p = 5.76 \times 10^{-5}$) and an increased sense that the experiment increased the students' understanding of chemistry ($m_{2011} = 0.33$, $m_{2012} = 0.98$, $t(142.9) = -4.21$, $p = 4.58 \times 10^{-5}$).

Open response comments for this same experiment corroborate this finding. In response to the question *"Did you enjoy the experiment? Why or why not?",* the relative occurrence of positive and negative comments relating to understanding appeared to shift in favour of more positive comments in 2012 (1 positive and 11 negative in 2011, 6 positive and 2 negative in 2012, $p = 4.44 \times 10^{-3}$). Understanding was also negatively mentioned less frequently in response to this question, both in the context of all item 15 comments received (11 of 55 comments in 2011, 2 of 79 in 2012, $p = 1.69 \times 10^{-3}$), and considered amongst the negative responses only (11 of 39 in 2011, 2 of 31 in 2012, $p = 2.91 \times 10^{-2}$). The copper(II) ion concentration experiment also shows some indication of an increased understanding, with positive comments related to understanding stated as a reason for liking the experiment (item 15) more often in 2012 (8 of 81 in 2011, 21 of 95 in 2012, $p = 4.05 \times 10^{-2}$). Whilst these differences may reasonably be attributed to family-wise error, these observations collectively serve to reinforce the significant differences already observed in the Likert-type data. Considering the multitude of indicators of increased understanding, in the vapour pressure experiment at the very least, there is clear evidence that student perception of understanding increased when the laptops were used in place of the data logger devices.

Whether this improved sense of understanding is a genuine reflection of deeper learning remains unknown. A significant increase in perception of data interpretation skills development ($m_{2011} = 0.36$, $m_{2012} = 1.06$, $t(131.7) = -4.90$, $p = 2.76 \times 10^{-6}$) was detected in responses to the first survey item for the vapour pressure experiment, however the truth of this increased perception is unknown. In response to the question *"What did you think was the main lesson to be learnt from the experiment?",* students conducting the vapour pressure experiment in 2012 did cite at least one of the main concepts (Dalton's law or Raoult's law and when it applies, intermolecular forces, non-ideal mixtures) in their responses more frequently (20 of 45 comments in 2011, 47 of 71 in 2012, $p = 3.33 \times 10^{-2}$). Comments including Raoult's law or Dalton's law were also more frequent (14 of 45 comments in 2011, 37 of 71 in 2012, $p = 3.48 \times 10^{-2}$). However, both of these results may be attributed to family-wise error and hence are inconclusive.

3.1.4.4 Increased Perceived Simplicity

In addition to the improvement in student perception of understanding, the students using the laptop computers also appear to have reported a view that the experiment was "simple" (a term commonly used by students in their open responses) more frequently than the cohort using the data logger devices. In response to being asked if they liked the experiment and why (item 15) students conducting the vapour pressure experiment using the data logger interface made no positive comments about the level of simplicity of the experiment, whereas 7 negative comments were received stating the lack of simplicity was their reason for disliking

the experiment. In 2012, using the laptop interface, this perception was entirely reversed. No students mentioned a lack of simplicity, and 12 listed the experiment's simplicity as their reason for liking the experiment, a clear significant improvement ($p = 1.99 \times 10^{-5}$).

As a consequence of this, the vapour pressure experiment exhibits a number of differences in the tests comparing frequency of reasons for liking and disliking the experiment between the two years regarding the experiment's simplicity. In response to item 15, simplicity was more frequent amongst reasons given for liking the experiment (0 of 18 comments in 2011, 12 of 54 in 2012, $p = 2.99 \times 10^{-2}$) and was less frequent amongst reasons given for disliking the experiment (7 of 39 comments in 2011, 0 of 31 in 2012, $p = 1.50 \times 10^{-2}$). When considering all comments received for this survey item, simplicity was mentioned positively more often (0 of 55 comments in 2011, 12 of 79 in 2012, $p = 1.45 \times 10^{-3}$) and mentioned negatively less often (7 of 55 comments in 2011, 0 of 79 in 2012, $p = 1.55 \times 10^{-3}$). These differences are not significant beyond attribution to family wise error, but may well be genuine since they are reminiscent of the significant difference already confirmed.

There is also some evidence that students conducting the copper(II) ion concentration experiment in 2012 when using the laptop interface reported a perception of simplicity more frequently than the 2011 cohort, though it is possible that there exists a degree of bias in samples taken for this experiment. Simplicity was mentioned as a reason for liking the experiment more often in response to item 15, both relative to all responses received (12 of 81 comments in 2011, 32 of 95 in 2012, $p = 4.98 \times 10^{-3}$), as well relative to other reasons given for liking the experiment (12 of 66 comments in 2011, 32 of 86 in 2012, $p = 1.18 \times 10^{-2}$). These differences are not large enough to exclude the possibility of being artefacts of family-wise error, but again are reminiscent of effects already seen elsewhere.

3.1.4.5 Further Comments

Following the technological user interface change made, student perception of the vapour pressure experiment in particular was subject to widespread improvement, with students reporting a greater perception of understanding of the experiment and clarity of the learning outcomes, a complete reversal in the initial perceived lack of simplicity of the experiment and a more positive perception of overall learning experience. The novel equipment was seen in a more positive light, and students began making comments more reminiscent of the usual benefits of real time graphing technology in MBL exercises. The biological buffers experiment also received fewer negative comments about the equipment, showing the laptop interface to be the more positively received option. The copper(II) ion concentration experiment did not show any clear difference between the two forms of the experiment, but the data do show some signs of improvement in students' perceived understanding and the perceived simplicity of the experiment.

Reasons for the lack of major difference observed in the copper(II) ion concentration experiment are unclear, although it is possible that sampling bias could prevent any differences being detectable. It is also possible that differences for this experiment were not observable as the microcomputer technology is used in a comparatively smaller portion of the task in this case. Relative portions of the task in which the microcomputer technology was used could also explain why widespread significant improvement was more evident in the vapour pressure experiment than for the biological buffers experiment.

It is conceivable that these observed improvements arose because a portion of the instructions were re-written to accommodate the different technology in 2012, but there is little data to suggest this is the case. An exception to this is the case of the vapour pressure experiment, in which the manual was reported as a reason for disliking the experiment (item 15) less often in 2012 (12 negative comments of 55 total comments in 2011, 5 of 79 in 2012, $p = 1.53 \times 10^{-2}$). However, this difference is not significant to the degree that it could not be attributed to family-wise error. Far stronger evidence exists that changing the manual was in fact detrimental rather than beneficial in the case of the biological buffers experiment. The material provided for the experiment was negatively mentioned significantly more frequently when asking students for potential improvements to the experiment (item 18), both in the context of all comments made (3 of 35 in 2011, 17 of 32 in 2012, $p = 1.10 \times 10^{-4}$), and considering only those improvements listed (3 of 31 in 2011, 17 of 25 in 2012, $p = 1.11 \times 10^{-5}$). It is possible that without this detrimental change in the instructions, some of the positive effects of replacing the handheld graphing data loggers seen in the vapour pressure experiment would also have become apparent in the biological buffers experiment. In any case, the changes in the manual do not appear clearly responsible for the significant improvements observed, meaning these effects are reasonably concluded to be genuine consequences of the technological change.

The influence of having different practical demonstrators between years is not known with certainty. However, responses to item 8 of the survey, asking students about effective supervision and guidance of the demonstrator, do not differ between years in any case even if multiple comparisons are unaccounted for. From this it appears reasonable to conclude that demonstrators were similarly effective in the students' views, and are unlikely to have differently influenced the perceptions of students in the two different years compared.

3.1.5 Conclusion

When conducting experiments in 2011 using the *PASCO Xplorer GLX* handheld graphing data logger, a number of negative student comments were received about the devices, often concerning difficulty of their use. Replacing these with laptop computers equipped with *PASCO DataStudio* software in 2012 resulted in numerous improvements in student perception data as compared with the previous year's cohort, including a significant reduction in these comments amongst other benefits. Based on the reduction in comments about the difficulty of using the technology, and the fact that the user interface was the only facet of the procedure altered, the data suggest that a more easily used technological interface plays a key role in positive student perception of these experiments. A recognition of the fact that the user interface of technology can influence student perceptions as strongly as has been observed here may significantly assist in alleviating student issues with microcomputer based laboratory activities reported in the literature. It is suggested that a familiar user interface is a vital element of the teaching laboratory when computers are used, and that in order to allow full access to the potential benefits of microcomputer based laboratory activities from the student perspective, handheld graphic data logging devices are far less preferable than alternatives equipped with the same capabilities that are more easily used by students, such as laptop computers.

The study presented in this section exemplifies a typical ASLE survey-based study utilising common integer value scoring of Likert-type items. Using the conclusions drawn here as a baseline for comparison, alternate techniques may now be applied to the same data set to

confirm that the same conclusions may be drawn, despite limitations of integer scoring methodology. Such a comparison is presented in the section following.

3.2 Justifying the conclusions of a scored analysis: Rasch techniques applied to the technological interfaces study

3.2.1 Outline

The previous section states its conclusions in such a way as to imply the observed differences in survey response were caused by a change in a feature of experiment design: the use of a laptop computer in place of a handheld data logger (see section 3.1). The claim made at the study's conclusion is not that the improvements reported were due to a more easily satisfied student cohort. Rather, there is presumed to be some "objective" measure of experiment quality which has been raised, leading students to broadly report a better outcome. Ultimately, ASLE scores obtained for an experiment are presumed to act as measurements of this "objective" experiment quality, which may be considered independent of the particular students surveyed. Rasch analysis presents an opportunity to explicitly test the correspondence between scored ASLE data and more sample independent measures, as it has the capacity to separate student dependent and student independent contributions to survey response. The present section therefore re-analyses the identical data set that was used in the previous section, this time using Rasch modelling techniques. Errors introduced due to the scoring techniques' lack of sample independence are investigated, inherently testing the first primary hypothesis of this thesis: that "*conclusions drawn from the ASLE survey data using typical scoring techniques resemble conclusions drawn using sample independent, interval scale measures extracted from the same data".*

It was claimed in section 3.1 that the data logger interface studied was less preferable to a laptop equipped with equivalent software, implying that use of the laptop interface was a superior way to design the experiments. Claims like these, typical of ASLE survey research, convey the idea that one experiment design is better than another design, in a sense true for students generally (see section 1.2.2). In this way the conclusion is implied to hold a degree of "objectivity": whilst it is acknowledged that some variation in student preferences may exist, there is a broader sense in which one option is claimed to be preferable to the other once individual student variations are 'averaged out'.

As discussed previously in the introductory material to this thesis (see section 1.3.1), requiring this "objectivity" of any measurements estimated is one means of deriving the Rasch model. Rasch recognised this requirement for the items students respond to, but also for measures assigned to students themselves, notably test scores. He termed this requirement of valid comparisons "specific objectivity", stating:[124]

> *"comparisons between individuals become independent of which particular instruments -- tests or items or other stimuli -- have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class -- measuring the same thing -- independent of which particular individuals, within a class considered, were instrumental for comparison."*

In an ASLE survey context, the "stimuli" Rasch refers to above are the experiments students conduct and the survey question to which they respond; Rasch states it ought to be possible to compare different experiment designs in a way that is independent of the student responding. This is exactly the presumption discussed above, where it is presumed there is an "objective" sense in which one experiment design is better or worse than another. Because ASLE survey mean scores are currently used as the measures of quality for an experiment, it is these scores

which must embody this specific objectivity. As such, mean scores should remain roughly equivalent regardless of the students providing the survey responses. Failing this, scores must either be inaccurate or unreliable estimates of "objective" experiment quality, or experiment quality cannot be assigned a numerical value in any objective sense in the first place. The possibility of the latter of these two options will be addressed in subsequent studies of this thesis (see section 4.1). The accuracy and reliability of scores as measures, however, can be evaluated in part by testing the extent to which student dependent variables influence the scored results observed. This is the primary objective of the study presented here.

Because Rasch measurement is the only means of estimating student dependent and student independent measures separately, Rasch analysis is required in order to explicitly test the contribution (or lack thereof) of student dependent effects on ASLE mean scores observed. To this end, the following investigation uses Rasch measurement to evaluate the distributions of student "biases" in the gathered samples used to conduct the previous, score-based study. Samples are analysed for their comparability between years, as well as their probable resemblance to a representative sample of the student population. If the results of score-based analyses are to resemble the results obtained using Rasch measures as per the first primary hypothesis of this thesis (see section 1.4.1), then these student biases must not confound the results obtained by score-based methods to a substantial degree.

3.2.2 Specific methods

The same data used to conduct the previous investigation (section 3.1) were used in this study. Sample sizes were adequate for Rasch analysis, as measures defining the item-response construct are reasonably informative for polychotomous survey items (such as is the case in this study) for sample sizes above approximately 50.[298] Cases in which students responded exclusively in the most extreme positive or negative categories of the survey were removed from consideration in these analyses, as justified previously (see section 2.3.3). The number of these cases for each data set is noted, with commentary on the likely impact of their presence. Data obtained from different sample groups were disconnected (see section 2.3.3) owing to a lack of student identification on survey responses gathered. This requires that whilst student bias distributions can be examined for each sample group in this study (defined by year and experiment conducted), it is not possible to contrast the distributions obtained from two sample groups directly. Similarly, it is not possible to directly contrast the absolute values of any student independent measures estimated from Rasch modelling directly between sample groups. However, as will be seen, this limitation does not imply that the influence of student dependent effects on scored results cannot be assessed.

To gain a measure of each student's bias towards answering positively, a separate "rating scale" Rasch model (see Equation 2 presented previously in section 2.2.1) was generated for each of the two cohorts, for each experiment, using the *Winsteps* Rasch measurement software. Student "ability" measures (β) specific to the person responding on that occasion and item "difficulty" measures (δ) specific to the question asked for that experiment were estimated in each model,[i] then used to test the apparent adequacy of student samples and their comparability across different occasions. The rating scale model was used rather than a partial credit model (see section 2.2.1) due to the small number of data points.

[i] This formulation of β and δ measures is later determined to be the best general explanatory model of the ASLE data. See section 4.1.3.

In order for data collected to be representative of the broader student population, it should be the case that the sample does not include a disproportionately high number of students biased towards providing either especially negative or especially positive responses. That is, the distribution of student measures (biases) within the sample should not be skewed and should represent the extremities and centre of the distribution of student biases with appropriate proportion. In a Rasch measurement context, this 'bias' towards positive response in general corresponds to the student "ability" measure (β). The distribution of student measures was therefore examined for normality in each case.

Presuming a representative sample to be distributed normally (a presumption made in the initial stages of Rasch model estimation; see section 2.3.1), skew in the distribution of student measures was taken to indicate that that the more positively biased students and the more negatively biased students were potentially represented disproportionately in the sample, whilst kurtosis was taken to suggest that the middle of the distribution and/or the extremities of the distribution could be represented in inappropriate proportion. It is acknowledged that this methodology presumes that non-normal distributions of student biases would not be expected to occur at the population level.

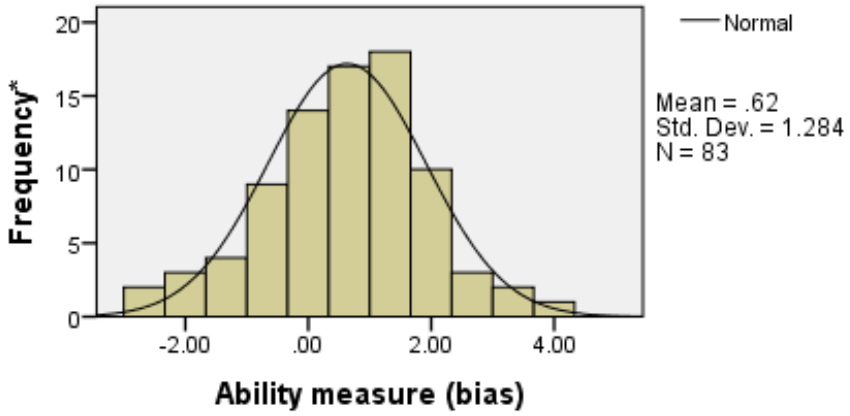
In order to test the comparability of the 2011 and 2012 student groups, an experiment which had not been altered between the two years was selected to act as a "negative control". Data from both the 2011 and 2012 iterations of the Analysis of Spinach Extracts experiment were accordingly merged, then used to generate another (single) rating scale Rasch model. Given the structure of this experiment remained unchanged between the two years, the "objective"' quality of the experiment (modelled by the δ measures) may be reasonably presumed identical. Presuming equality of student independent measures in this way allows data connectivity between the two years in the case of this specific experiment, unlike for other experiments. The student measures (β) obtained were compared for the 2011 and 2012 groups to test comparability of the student cohorts sampled in the two years , whilst DIF analysis (see section 2.5.3) was performed to confirm equality of the item measures (δ) between the two years as assumed in the analysis.

Differential item functioning (DIF) analysis is robust against large differences in student propensity to respond positively when coupled with Rasch analysis.[299] In the cases where the week of the semester in which surveys were collected was known, DIF analysis was used to identify any inconsistencies in student independent measures (δ) between different weeks of the semester. However, it was not inferred which specific items differed between weeks, as significant DIF has the potential to be "artificial" rather than genuine (see section 2.5.3 for a more detailed discussion of DIF). The distribution of student dependent measures (β) was subsequently contrasted between weeks in light of the DIF analysis results, in order to test whether student samples from different weeks appeared to have equal propensity toward positive response.

3.2.3 Results

3.2.3.1 2011 Vapour pressure experiment

One student responded in the most extreme negative category for every question posed in this sample. Consequently, the Rasch analysis assigned them an arbitrarily low "ability" measure, such that their predicted responses all appear at the extreme low categories. For this reason, responses from this person were excluded from consideration in the data presented.

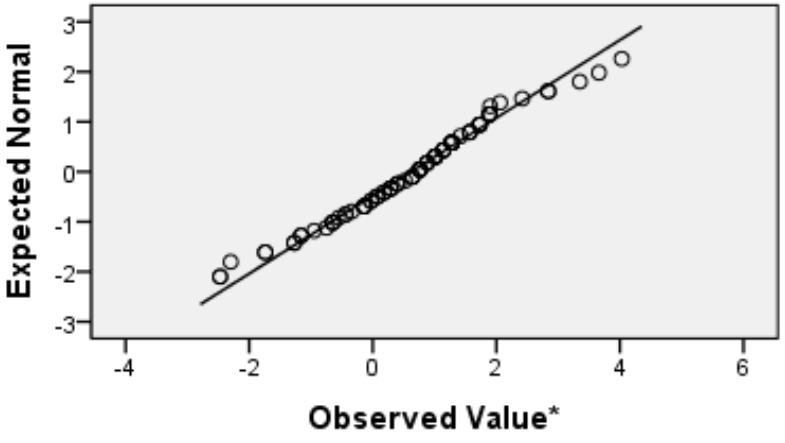


* one extreme response excluded from consideration

Figure 10: Student measures for the 2011 Vapour pressure experiment

It is possible that this student's response is a "donkey vote", given they are well outside the distribution of the other students (assigned "ability" measure of -6.11, well beyond three standard deviations about the mean). Given the magnitude of significant differences observed in the Likert-style data for this experiment, it is not expected that this single student would impact the results of the main comparative study in such a way as to alter the conclusions drawn.

The distribution of student measures observed appears approximately normal, as is evident not only in the results of the Kolmogorov-Smirnov ($D = 0.074$, $df = 83$, $p \geq 0.200$) and Shapiro-Wilk ($W = 0.983$, $df = 83$, $p = 0.363$) tests, but also by graphical inspection (see Figure 10 and Figure 11). The distribution exhibits neither significant skew (skewness = -0.082, S.E. = 0.264), nor significant kurtosis (kurtosis = 0.505, S.E. = 0.523). Overall, with the exception of one outlying student response, the sample of students appears not to favour either positively biased or negatively biased students excessively, and the centre of the distribution and tails of the distribution of student biases are represented in keeping with a normal distribution.



* one extreme response excluded from consideration

Figure 11: Q-Q plot of student measures for "Vapour pressure" in 2011

3.2.3.2 2012 Vapour pressure experiment

The overall student bias distribution for this data set was found not to be normally distributed by the Kolmogorov-Smirnov (D = 0.102, df = 102,p = 0.010) and Shapiro-Wilk (W = 0.949, df = 102, p = 0.001) tests.

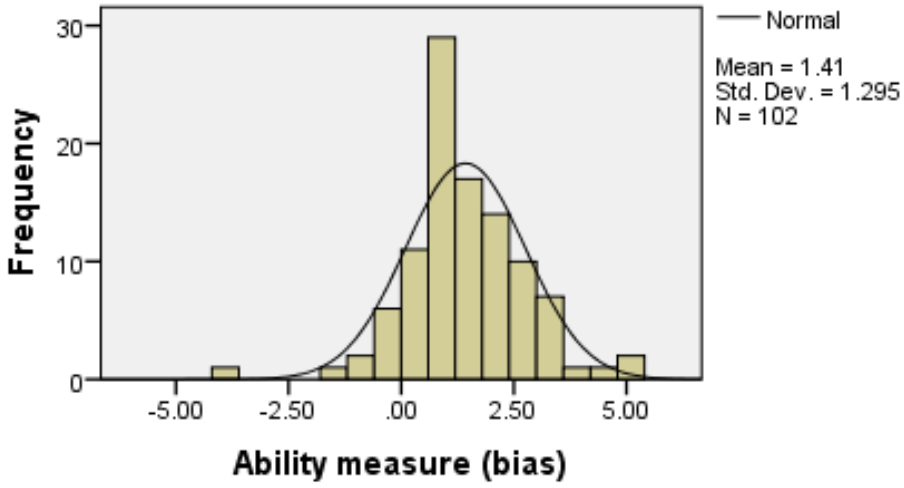


Figure 12: Person measures obtained for the Vapour pressure experiment in 2012

Contributing to non-normality is the presence of a highly negatively biased student, most clearly visible to the bottom left of the Q-Q plot (Figure 13). However, unlike in the 2011 version of the experiment, this student's response does not appear to be a possible "donkey vote", as they have not responded in the extreme negative category in every case, and should be left in the analysis.

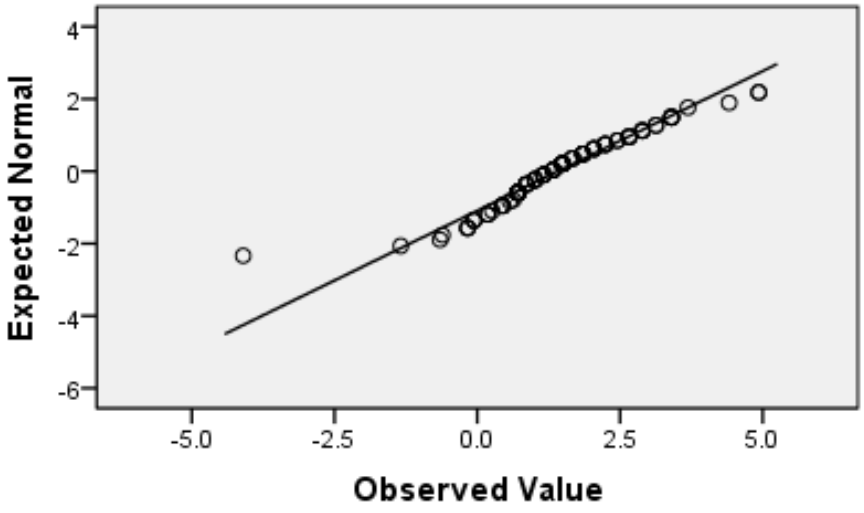


Figure 13: Q-Q plot of student measures from "Vapour pressure" in 2012

In order to investigate the possibility that the week in which experiments were conducted influenced the results obtained, the Rasch model for this experiment was examined for differential item functioning (DIF). Three person groups were compared: those who completed the experiment in Week 6, those who completed the experiment in Week 8, and those who completed the experiment in Week 10. This comprises the full data set of responses received for this experiment. Results detailed below in Table 4 show that no survey item's "difficulty" (δ) measure significantly differed based on the week the student conducted the experiment. The item closest to the occurrence of this was question 12, however the fit statistics show that

this is well within expectation based on the Rasch model generated for this data. It does appear that an unexpected degree of similarity between weeks exists for questions 1 and 2, however the reason for this is unclear. This is not problematic, as it represents an unexpected lack of variation in responses received, rather than an unexpected degree of variation and error.

Table 4: DIF between weeks for the 2012 "Vapour pressure" experiment

Survey Item	Summary DIF			Between group fit statistics	
	χ^2	d.f.	p	Mean-square	t=ZSTD
Q1	0.000	2	1.000	0.000	-2.460
Q2	0.035	2	0.984	0.006	-2.119
Q3	2.794	2	0.244	0.487	-0.307
Q4	1.642	2	0.437	0.320	-0.615
Q5	0.435	2	0.805	0.055	-1.524
Q6	2.191	2	0.331	0.423	-0.415
Q7	0.219	2	0.898	0.028	-1.753
Q8	1.343	2	0.508	0.201	-0.908
Q9	0.441	2	0.802	0.074	-1.408
Q10	4.035	2	0.131	0.558	-0.197
Q11	3.230	2	0.196	0.513	-0.266
Q12	5.839	2	0.053	1.045	0.378
Q13	0.439	2	0.803	0.060	-1.492
Q14	0.828	2	0.659	0.113	-1.216

ANOVA was used to test the equality of the distribution of student measures (biases) between these weeks, with a significant difference detected ($F(2,99) = 4.067$, $p = 0.020$). However, the descriptive statistics show that student measures in Week 10 were not normally distributed, displaying significant kurtosis (Table 5). This could be due to the small number of students sampled in that week.

Table 5: 2012 Vapour pressure experiment: Distribution of student measures (biases)

Week conducted	Mean	N	Standard Deviation	Kurtosis		Skewness	
				value	Std. Error	value	Std. Error
6	1.6143	30	1.04016	-.685	.833	.374	.427
8	1.5745	51	1.20966	.252	.656	.555	.333
10	.7124	21	1.61062	5.294	.972	-.550	.501
Total	1.4087	102	1.29460	2.929	.474	-.197	.239

The non-normality of the overall distribution appears to arise because of kurtosis, generated in part by differences in the distribution of student biases collected in week 10 compared to data collected in other weeks. The distribution of student biases corresponding to those sampled in Week 10 appears to be centred more towards negative responses than for the other two weeks. Evident in the histogram displaying the full distribution of student biases (Figure 12) is an unexpectedly high frequency of responses at student measure of approximately 0.7, corresponding to the location of the mean of the Week 10 samples (see Table 5). Were the different weeks sampled in equal proportion, the Week 10 responses would appear more frequently and the mean student bias would shift further in favour of negative response.

One data point was identified as an outlier to the overall distribution (outliers are depicted as stars in Figure 14 below). However, it does not appear that removal of this outlying data point would alter the conclusions of the data above, as the Week 10 samples still have a visibly lower median value when viewed separately.

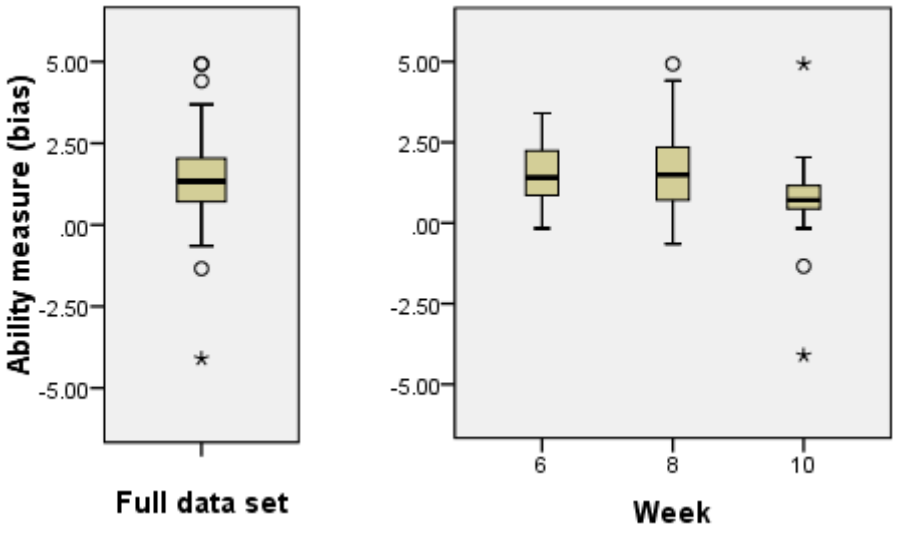


Figure 14: Identification of outlying student measures in the 2012 "Vapour pressure" experiment

(left: full data set, right: week specific data)

Even with this outlying negatively biased respondent removed from consideration, the overall distribution still appears non-normal. The non-normality, however, now manifests itself as significant skewness rather than kurtosis (Table 6), and is less evident in the Shapiro-Wilk test of normality (W = 0.967, df = 101, p = 0.012). It is, however, slightly more evident judging by the Kolmogorov-Smirnov test (D = 0.115, df = 101,p = 0.002).

Table 6: Effect of removing the outlier from the 2012 vapour pressure experiment data

Data set	Mean	N	Standard Deviation	Kurtosis		Skewness	
				value	Std. Error	value	Std. Error
Outlier removed	1.4633	101	1.17741	.575	.476	.639	.240
Outlier included	1.4087	102	1.29460	2.929	.474	-.197	.239

Overall, the data show that the 2012 sample taken from the vapour pressure experiment shows some evidence of potentially misrepresenting the broader population. A better representation of responses generated in Week 10 would have produced a higher propensity for students to respond negatively, and by implication, the current sample appears biased towards positive responses from early in the semester. Depending on the magnitude of this effect, this could theoretically alter the conclusions of the initial score-based study.

Given the parameters of the Rasch model generated, it is possible to computationally simulate the expected frequency of responses under a theoretical, more negatively biased sample. This is useful for determining the likely magnitude of the difference between the positively biased sample collected and a more equally representative sample. Figure 15 shows the "item characteristic curve" (ICC) corresponding to the Rasch model generated for this data. It displays the expected (average) response score as a function of student measure minus item measure (displayed on the plot as the "latent trait measure"). The unexpectedly high empirical

value at a latent trait measure of -6 Logits is due to the very low number of data points available at this extreme.

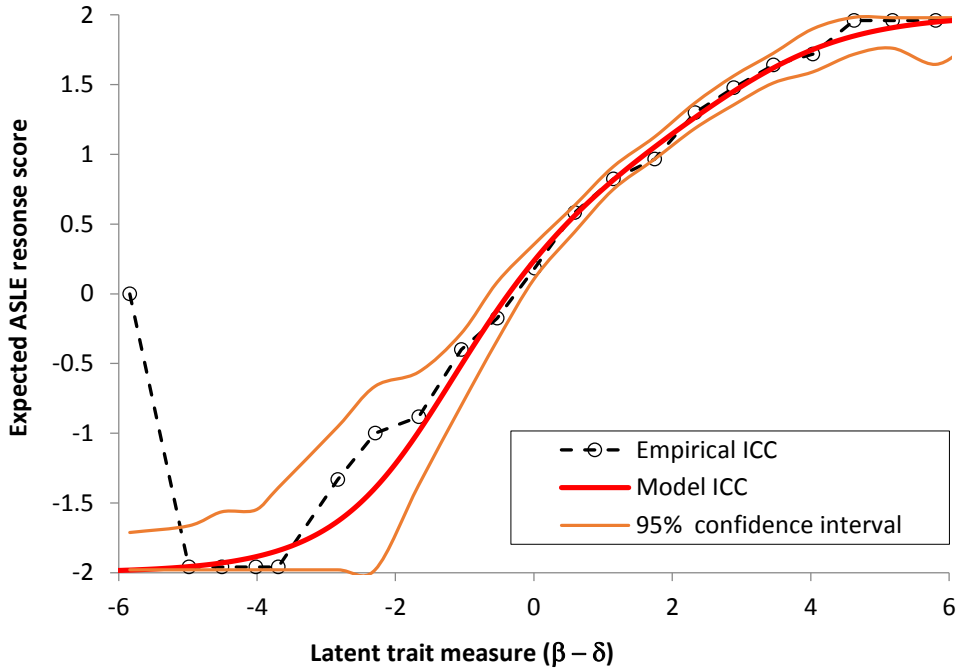


Figure 15: Item characteristic curve (ICC) for the 2012 Vapour pressure experiment Rasch model

Using this plot, it is possible to roughly estimate the magnitude of the expected change in average scored response, were the distribution of student biases (β values) shifted. It is expected that under a more equal representation of each of the three weeks sampled, the population mean bias could not possibly be lower than the current observed Week 10 mean bias (rather, it would be somewhere between the low Week 10 mean bias and the higher Week 6 and 8 mean bias values).

The Week 10 average student measure (β) is 0.7124 logits, whilst the observed sample overall has an average measure of 1.4087 logits (both including the outlier previously discussed). Assuming a model unbiased sample to have an average student measure intermediate between these two values, an unbiased sample would have an average student measure of approximately 1.06 logits; only approximately 0.35 logits lower than the observed sample. Using the ICC plot, a change in student ability measure of only 0.35 logits (from 1.41 to 1.06) corresponds to a change in average scored response of only approximately 0.15 score units. This gives a rough approximation to the size of the difference that would be expected in observed average response scores if the sample reflected Week 10 responses as opposed to the observed, positively biased sample.

Displayed in Table 7 is a replication of the statistical comparisons made between the observed data for the 2011 and 2012 versions of the vapour pressure experiment, and alongside, the comparisons corresponding to a case in which the mean scores of 2012 items were shifted 0.15 units lower.

Table 7: Observed bias in 2012 Vapour pressure sample - Magnitude of the impact on responses

	Observed 2011 sample compared to observed 2012 sample				Observed 2011 sample compared to hypothetical unbiased 2012 sample			
	Δmean	df	t	p	Δmean	df	t	p
Q1	0.70	131.65	-4.90	0.000003	0.55	131.65	-3.85	0.000181
Q2	0.55	137.85	-3.64	0.000388	0.4	137.85	-2.64	0.009209
Q3	0.61	161.51	-3.53	0.000533	0.46	161.51	-2.67	0.008430
Q4	0.40	169.52	-2.62	0.009497	0.25	169.52	-1.64	0.103801
Q5	0.64	138.36	-4.15	0.000058	0.49	138.36	-3.17	0.001870
Q6	0.65	142.86	-4.21	0.000046	0.5	142.86	-3.23	0.001531
Q7	0.29	167.99	-1.94	0.054108	0.14	167.99	-0.92	0.357564
Q8	0.14	169.00	-1.07	0.284552	-0.01	169.00	0.08	0.936486
Q9	0.39	179.09	-2.34	0.020224	0.24	179.09	-1.45	0.150128
Q10	0.14	162.76	-0.95	0.341206	-0.01	162.76	0.10	0.922074
Q11	0.04	172.69	-0.28	0.783503	-0.11	172.69	0.90	0.367410
Q12	0.43	147.63	-3.27	0.001352	0.28	147.63	-2.13	0.034877
Q13	0.49	138.19	-4.89	0.000003	0.34	138.19	-3.38	0.000929
Q14	0.71	149.03	-5.06	0.000001	0.56	149.03	-4.00	0.000101

Shaded red are the cells containing probability values deemed statistically significant, accounting for the multiple comparisons correction used in the original comparative study. As can be seen above, in a hypothetical sample without the bias imparted by an underrepresentation of students who conducted the vapour pressure experiment in Week 10, significant differences in responses to survey questions 5, 6 and 13 are no longer evident. Significant differences in responses to questions 1 and 14, however, remain.

Based on this rough analysis, it is expected that the differences in scored responses to survey items 5, 6 and 13 for the vapour pressure experiment could feasibly be attributed to bias in the sample, introduced by underrepresentation of students conducting the 2012 experiment in Week 10. Some indication of difference in these items still exists after accounting for this bias, however the differences are small enough to possibly be attributed to the issue of multiple comparisons.

To further corroborate this conclusion, the 2011 and 2012 "Vapour pressure" samples were merged and entered into a single rating scale model Rasch analysis. Each item of the survey was tested for DIF between the two years. Since DIF compares the "difficulty" (δ) of responding positively to an item independent of student propensity to respond positively (β), it provides a means of testing for significant differences independent of student biases. As stated in the introductory material to this study, no single student provided responses in both years for this same experiment, and therefore no common point of reference exists to establish the relative 'central location' of the measures for each group (β or δ). DIF analysis can, however, still be used to reveal differences in the relative locations of each of the δ measures with respect to the others within each group since, as mentioned in the introductory material, DIF is robust against large differences in student propensity to respond positively when coupled with Rasch analysis.[299] Entering the two disconnected subsets into a single Rasch model, it can be initially assumed that the δ measures are equivalent for the two groups in order to "connect" the data. Significant DIF detected would indicate falsity of this

assumption, though would not clearly indicate which items differed between groups specifically. The results of such an analysis, performed on a merged model of the 2011 and 2012 vapour pressure experiment data, are presented in Table 8.

Table 8: DIF between the 2011 and 2012 forms of the Vapour pressure experiment

Survey Item	Summary DIF			Between group fit statistics	
	χ^2	d.f.	p	Mean-square	t=ZSTD
Q1	6.8875	1	0.0087	3.5327	1.5809
Q2	1.2245	1	0.2685	0.6179	0.1569
Q3	0.9657	1	0.3258	0.4847	0.0164
Q4	0.4314	1	0.5113	0.2167	-0.3757
Q5	2.5762	1	0.1085	1.3049	0.6682
Q6	3.7287	1	0.0535	1.895	0.9752
Q7	1.7904	1	0.1809	0.9047	0.4018
Q8	4.1944	1	0.0406	2.1367	1.0823
Q9	1.0124	1	0.3143	0.5085	0.0432
Q10	8.1537	1	0.0043	4.1969	1.7718
Q11	10.4717	1	0.0012	5.4435	2.0817
Q12	0	1	1	0.0114	-1.173
Q13	0	1	1	0.015	-1.1268
Q14	3.9268	1	0.0475	1.9938	1.02

The data clearly evidences cases where item difficulty measures significantly differ ($p < 0.05$ highlighted) between the two cohorts. That is, after accounting for student propensity to respond positively, it appears some items appear either more or less difficult to answer positively in the different years. Questions 1, 8, 10, 11 and 14 show significant DIF, with question 11 so different as to be outside expectations of this merged data Rasch model. The between group fit statistics for other items showing significant DIF do not reveal confident difference, but do appear slightly elevated (notably for questions 1 and 10). It is important to note that items exhibiting significant DIF are *not* necessarily the specific items which have changed between years.

Person (bias) measures estimated in this merged analysis were, by necessity, estimated under the initial presumption that all item measures (δ) were equivalent for the two years. The results presented above demonstrate this presumption to be a false one. Differences in measures of experiment quality (δ) appear to exist between the two years, independent of any student dependent contributions. It appears that significant differences between the data logger form of the experiment and the laptop form of the experiment exist independent of the fact the 2012 sample is likely to be positively biased.

3.2.3.3 2011 Biological buffers experiment

Two extreme responses, where the student provided the extreme positive response for all survey items, were removed from consideration in the following analysis. These two students, being a small fraction of the broader sample, are not expected to significantly impact on any conclusions.

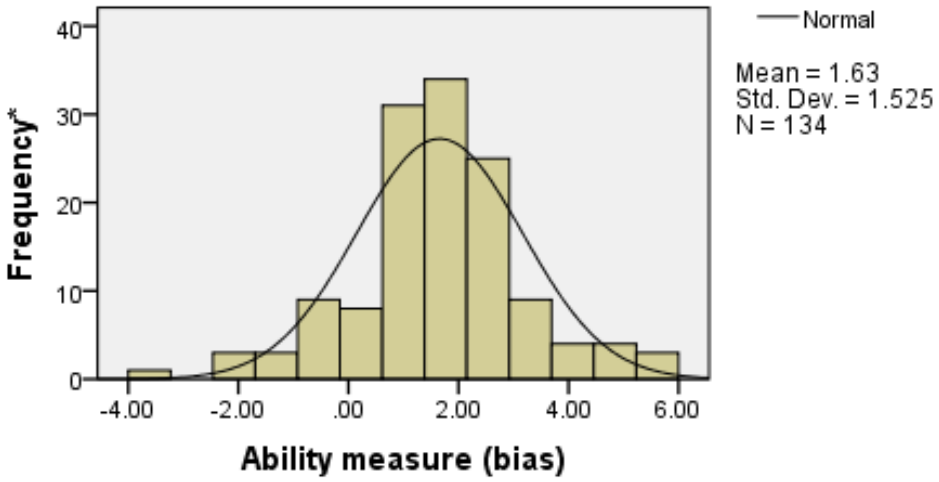


Figure 16: Person measures obtained for the Biological Buffers experiment in 2011

Results demonstrate that the 2011 sample from the biological buffers experiment contains a non-normal distribution of student biases, arising from a significant degree of kurtosis. Both the Kolmogorov-Smirnov (D = 0.106, df = 134, p = 0.001) and Shapiro-Wilk (W = 0.965, df = 134, p = 0.002) tests reject the null-hypothesis of normality. Kurtosis is significant, yet skewness remains acceptable (Table 9, presented on page 72). Kurtosis is easily visible upon visual inspection of the distribution (Figure 16). There also appears to be an outlying negatively biased student, as displayed in the box plot below (shown using a star, Figure 17) and also visible to the far left of the Q-Q plot (Figure 18).

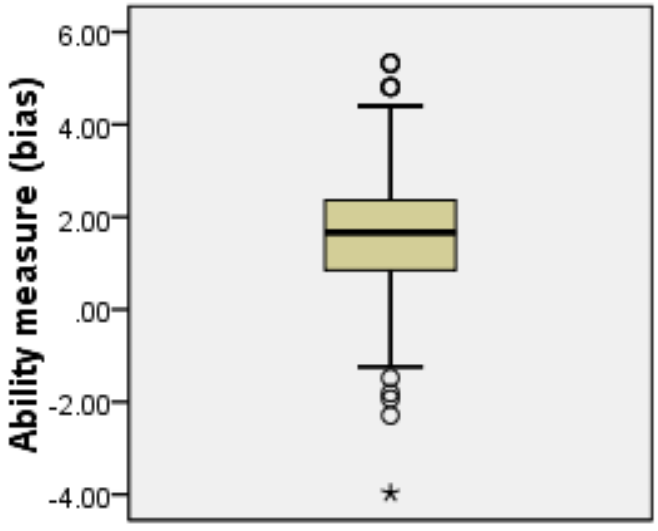
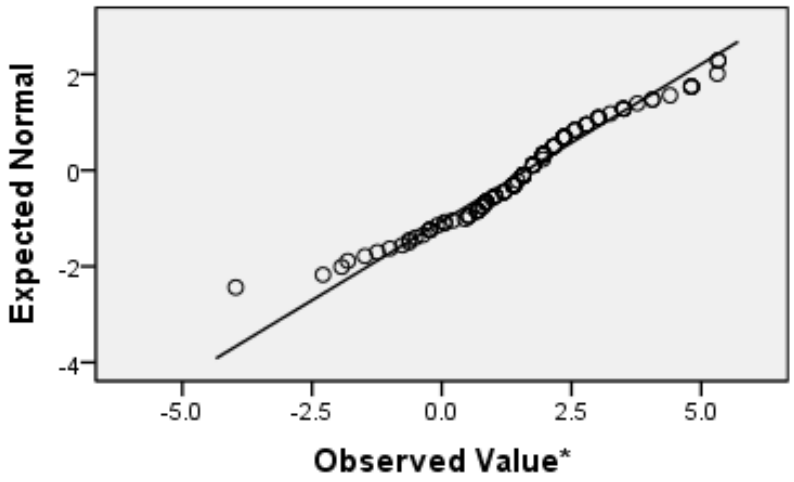


Figure 17: Outlier identification for the 2011 Biological buffers experiment student measure data



* two extreme responses removed from consideration

Figure 18: Q-Q plot of student measures for the 2011 Biological Buffers experiment

Removal of this outlying student from consideration does not alter any conclusions, as the distribution remains non-normal as judged by the Kolmogorov-Smirnov (D = 0.106, df = 133, p = 0.001) and Shapiro-Wilk (W = 0.972, df = 133, p = 0.008) tests. The degree of kurtosis does appear less significant, however (Table 9).

Table 9: Effect of removing the outlier from the 2011 Biological Buffers experiment data

Data set	Mean	N	Standard Deviation	Kurtosis		Skewness	
				value	Std. Error	value	Std. Error
Outlier included	1.6254	134	1.52497	1.495	.416	-.224	.209
Outlier removed	1.6674	133	1.45059	0.777	0.417	0.090	0.210

Overall, the distribution of student biases in this sample does not show evidence of overall bias in either the positive or negative direction. However, its shape (judging by kurtosis) may not be representative of the broader population. It is not expected that this would influence the results of the main study, as the distribution does not appear to be significantly skewed. That is, there does not appear to be a disproportionate number of students responding positively or responding negatively. There is simply an overrepresentation of the students with more extreme biases (both positive and negative, equally so), or an overrepresentation of those with average bias, or both.

3.2.3.4 2012 Biological buffers experiment

The Kolmogorov-Smirnov test for normality marginally rejects the hypothesis of normality for this sample (D = 0.100, df = 80, p = 0.046), whilst the Shapiro-Wilk test does not (W = 0.975, df = 80, p = 0.116). Given the lack of skew or kurtosis, as well as the reasonable appearance of the histogram (Figure 19) and Q-Q plot (Figure 20), the majority of evidence suggests that the distribution is sufficiently close to normality, and the sample does not appear to represent positively and negatively biased students in unequal proportion. Any departures from normality appear minimal, and therefore would be unlikely to influence conclusions drawn.

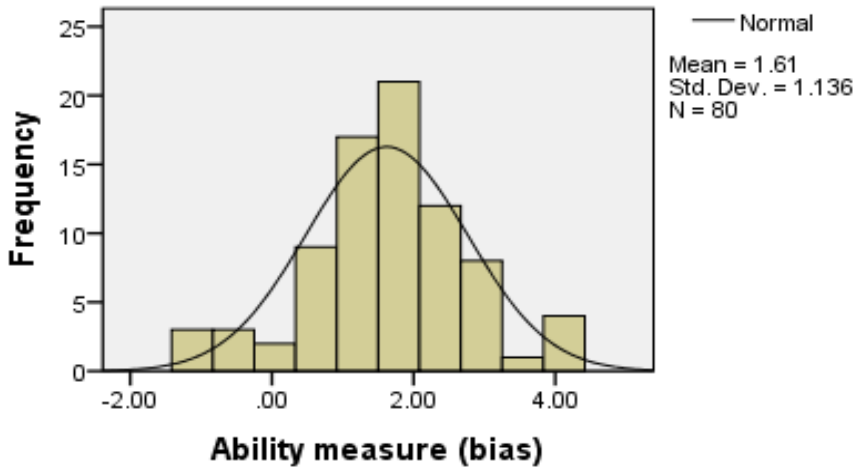


Figure 19: Distribution of person measures for the 2012 Biological buffers experiment

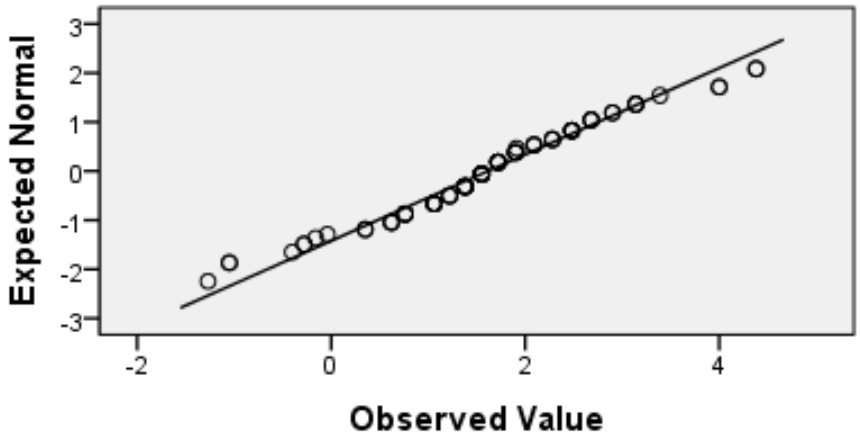


Figure 20: Q-Q plot of student measures for the 2012 Biological buffers experiment

In order to investigate the possibility that the week experiments were conducted influenced the results obtained, the Rasch model for this experiment was examined for differential item functioning (Table 10). Four person groups were compared: those who completed the experiment in Week 4, those in Week 6, those in Week 8 and those in Week 10. This comprises the full data set from this sample. No DIF was detected, indicating no single item appeared to be more or less “difficult” to answer positively in the different weeks sampled.

Table 10: DIF between weeks for the 2012 Biological buffers experiment

Survey Item	Summary DIF			Between group fit statistics	
	χ^2	d.f.	p	Mean-square	t=ZSTD
Q1	3.309	3	0.345	0.264	-1.044
Q2	3.348	3	0.340	0.291	-0.966
Q3	4.568	3	0.205	0.387	-0.725
Q4	2.133	3	0.544	0.206	-1.232
Q5	3.022	3	0.387	0.228	-1.158
Q6	1.726	3	0.630	0.159	-1.413
Q7	1.326	3	0.723	0.117	-1.603
Q8	2.316	3	0.508	0.208	-1.225
Q9	4.199	3	0.240	0.387	-0.725
Q10	2.837	3	0.416	0.223	-1.175
Q11	3.642	3	0.302	0.337	-0.846
Q12	2.333	3	0.505	0.174	-1.352
Q13	0.404	3	0.940	0.035	-2.199
Q14	1.600	3	0.659	0.150	-1.452

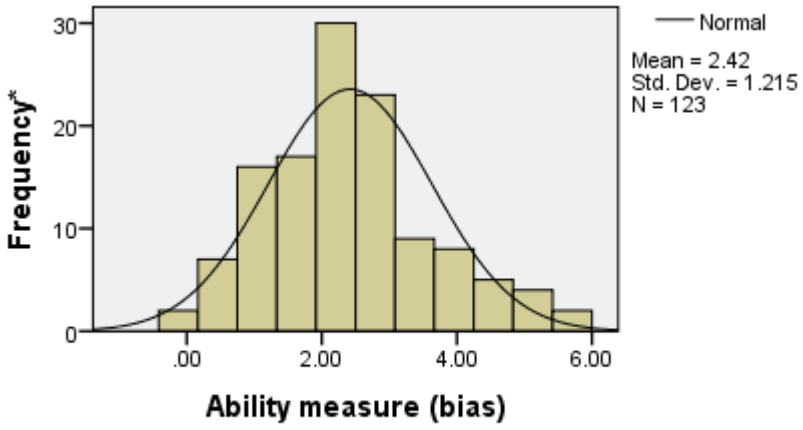
Student "ability" measures (bias) were compared between weeks by ANOVA, with no significant difference detected ($F(3,76) = 0.136$, $p = 0.938$). The distribution of biases within each week sampled appears normally distributed, with the exception that the Week 10 samples exhibit a significant degree of kurtosis. This could be attributed to the small number of samples when considering each week separately. The distribution of biases within the total sample for the 2012 Biological buffers experiment appears not to have a significant degree of skew or kurtosis, and generally appears normal (Table 11).

Table 11: Distribution of ability measures (biases) for the 2012 Biological buffers experiment

Week conducted	Mean	N	Standard Deviation	Kurtosis		Skewness	
				value	Std. Error	value	Std. Error
4	1.6947	15	.84574	-.978	1.121	.090	.580
6	1.6579	24	1.09054	-.104	.918	.002	.472
8	1.6419	21	1.40130	.559	.972	-.262	.501
10	1.4755	20	1.14374	2.163	.992	.306	.512
Total	1.6150	80	1.13650	.664	.532	-.052	.269

3.2.3.5 2011 Copper(II) ion concentration experiment

The distribution of student biases in this case appears marginally non-normal, arising from a degree of skewness (skewness = 0.439, S.E. = 0.218; kurtosis = 0.195, S.E. = 0.433), visible most prominently in Figure 21. This is suggestive of slight bias in the sample. Whether this bias is towards more positive responses or more negative responses is unknown, as it is not clear whether the skewness arises from an underrepresentation of responses at one extreme, or an overrepresentation of responses slightly off-centre of the distribution.

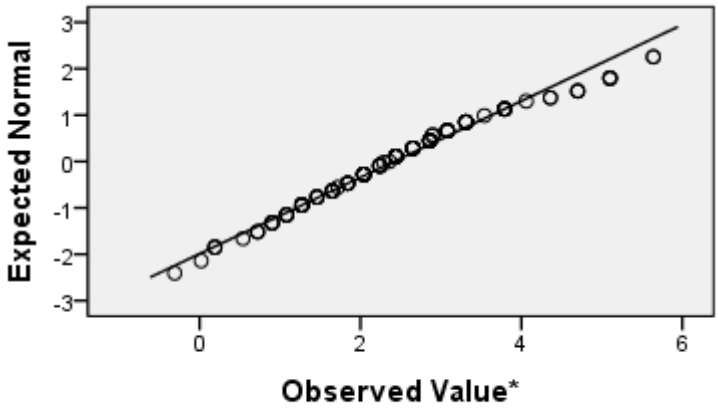


* three extreme responses removed from consideration

Figure 21: Distribution of student measures for the 2011 Copper(II) ion concentration experiment

Three extreme positive responses were excluded from consideration in these analyses. Were they included and were their contribution substantial, the overall distribution of biases would be more positive. This would not impact on the conclusions of the initial score-based study, as this effect would make improvements from 2011 to 2012 less evident, as opposed to more.

The non-normality of the distribution obtained is only slight, with the Q-Q-plot appearing largely reasonable (Figure 22), and tests of normality rejecting only marginally if at all (Kolmogorov-Smirnov: D = 0.078, df = 123, p = 0.064 ; Shapiro-Wilk: W = 0.979, df = 123, p = 0.048). Any overall bias is therefore likely to be only slight and of little consequence.



* three extreme responses removed from consideration

Figure 22: Q-Q plot of student measures for the 2011 Copper(II) ion concentration experiment

3.2.3.6 2012 Copper(II) ion concentration experiment

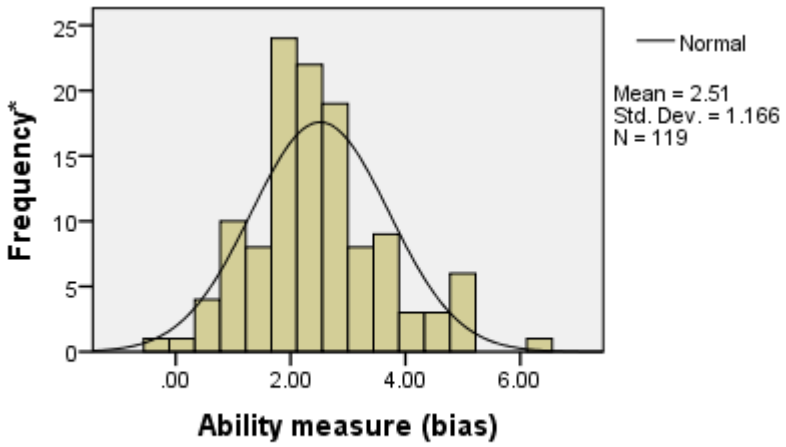
In order to investigate the possibility that the week in which experiments were conducted influenced the results obtained, the Rasch model for this experiment was examined for differential item functioning (DIF). Three person groups were compared: those who completed the experiment in week 6, those in week 8 and those in week 10. This comprises the full set of

responses collected for this sample. No DIF was detected, indicating no single item appeared to be more or less difficult to answer positively in the different weeks sampled.

Table 12: DIF between weeks for the 2012 Copper(II) ion concentration experiment

Survey Item	Summary DIF			Between group fit statistics	
	χ^2	d.f.	p	Mean-square	t=ZSTD
Q1	0.7459	2	0.6873	0.1227	-1.176
Q2	3.0938	2	0.2098	0.5578	-0.1972
Q3	3.099	2	0.2093	0.5609	-0.1926
Q4	3.2753	2	0.1915	0.5431	-0.2191
Q5	0.144	2	0.9324	0.0253	-1.7865
Q6	3.0705	2	0.2123	0.5317	-0.2362
Q7	0.3802	2	0.8275	0.063	-1.4732
Q8	5.1171	2	0.0759	0.8591	0.1852
Q9	0.3498	2	0.8403	0.0546	-1.5286
Q10	1.6299	2	0.4391	0.291	-0.6786
Q11	4.4028	2	0.1086	0.7655	0.0777
Q12	3.8376	2	0.1443	0.6571	-0.0585
Q13	1.7273	2	0.418	0.306	-0.645
Q14	0.7195	2	0.6965	0.1231	-1.1743

There is an extreme response used to generate the Rasch model used to conduct the above analysis; however their inclusion would not impact upon the results of the DIF analysis. As extreme responses necessarily fit the Rasch model optimally, they do not contribute to DIF statistics and are excluded from the analysis. This extreme response was excluded from consideration in the subsequent analyses for reasons described in the introduction (see section 2.3.3).



* one extreme response removed from consideration

Figure 23: Distribution of student measures for the 2012 Copper(II) ion concentration experiment

Student propensity to answer positively (bias) was also examined. The measures appear not to be normally distributed (Kolmogorov-Smirnov: D = 0.125, df = 119, p < 0.001 ; Shapiro-Wilk: W = 0.967, df = 119, p = 0.005), primarily due to a degree of skewness (see Table 13, presented

on page 78), visible in the distribution itself (Figure 23) as well as to some extent in the Q-Q plot (Figure 24).

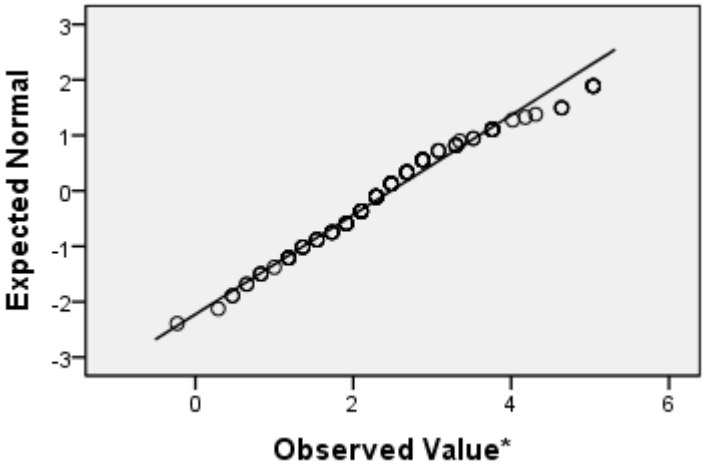


Figure 24: Q-Q plot of student measures for the 2012 Copper(II) ion concentration experiment

The samples from different weeks appear not to significantly differ as judged by ANOVA ($F(2,116) = 2.379$, $p = 0.098$) and the additional removal of an identified outlier (indicated as a star in Figure 25) does not alter this conclusion ($F(2,115) = 2.733$, $p = 0.069$). However, the data from Week 6 appear to exhibit significant skew and kurtosis when the outlying response is included (but not otherwise), and this may influence the validity of the ANOVA test (see details in Table 13).

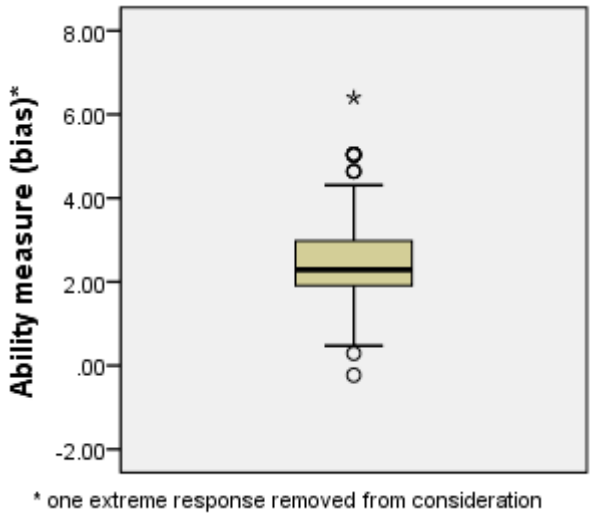


Figure 25: Outlying student identification for the 2012 Copper(II) ion concentration experiment

The mean "ability" measure (bias) of the total sample from this experiment appears roughly at the midpoint between the three mean measures taken from each sampled week. As such, uneven sampling between weeks seems to 'balance out' in terms of the central location of the distribution of biases. However, skew of the distribution of all results remains, possibly originating primarily from the Week 6 data, likely generating the observed non-normality. The

skewness is far less significant upon removal of the outlier, however the overall distribution remains non-normal as judged using statistical tests. (Kolmogorov-Smirnov: D = 0.114, df = 118, p = 0.001 ; Shapiro-Wilk: W = 0.970, df = 118, p = 0.009).

Table 13: Week-specific bias statistics for the 2012 Copper(II) ion concentration experiment

Extreme response removed only

Week conducted	Mean	N	Standard Deviation	Kurtosis		Skewness	
				value	Std. Error	value	Std. Error
6	2.5057	44	1.10600	2.380	.702	1.063	.357
8	2.8119	37	1.13227	-.035	.759	.299	.388
10	2.2321	38	1.22410	.780	.750	.649	.383
Total	2.5135	119	1.16627	.686	.440	.592	.222

Extreme response and outlier removed

Week conducted	Mean	N	Standard Deviation	Kurtosis		Skewness	
				value	Std. Error	value	Std. Error
6	2.4151	43	.93962	-.223	.709	.303	.361
8	2.8119	37	1.13227	-.035	.759	.299	.388
10	2.2321	38	1.22410	.780	.750	.649	.383
Total	2.4806	118	1.11428	.192	.442	.406	.223

Overall, this sample appears to be biased. Positively biased and negatively biased students have been sampled disproportionately in Week 6, notably due to a substantially outlying response, and this has translated into skew of the overall distribution of biases in the total sample. However, even the removal of this outlier does not entirely resolve the issue. It is unclear which students, those positively biased or those negatively biased, were sampled in inappropriate proportion.

Based on the analysis of both the 2011 and 2012 samples from the Copper(II) ion concentration experiment, sampling bias may perturb results obtained. It is uncertain whether this would over-exaggerate or under-exaggerate any differences in student perception of learning experience studied, though if the outlying student in 2012 is the exception to the general trend, scored results calculated with its inclusion may report more positive perception that warranted. This would exaggerate reported improvements, though not to a large degree provided the rest of the sample may be considered representative. Given very few differences were actually observed for this experiment in the initial score-based study, the apparent bias of these samples does not impact the broader initial conclusions.

3.2.3.7 Analysis of spinach extracts as a "negative control"

In order to establish comparability between the 2011 and 2012 student cohorts, a Rasch model was generated for the Analysis of spinach extracts experiment; an experiment which remained unchanged between the two years, and from which a relatively large sample was gathered (144 in 2011 and 77 in 2012). The data from both years was grouped into the same Rasch model, testing for DIF between the item measures estimated from each year's responses.

Table 14: Analysis of spinach extracts item measures compared between years

Survey Item	Summary DIF[a]			Between group fit statistics[b]	
	χ^2	d.f.	p	Mean-square	t=ZSTD
Q1	3.8564	1	0.0496	1.7746	0.9183
Q2	1.303	1	0.2537	0.5816	0.1208
Q3	0.7015	1	0.4023	0.3118	-0.2115
Q4	0.6171	1	0.4321	0.2747	-0.271
Q5	3.7627	1	0.0524	1.7378	0.9005
Q6	0	1	1	0.0011	-1.4323
Q7	2.3871	1	0.1223	1.0778	0.5251
Q8	8.3175	1	0.0039	3.84	1.672
Q9	2.9865	1	0.084	1.3614	0.7012
Q10	0.6448	1	0.422	0.2975	-0.2338
Q11	19.1133	1	<0.0001	9.6979	2.8738
Q12	0.1138	1	0.7358	0.0543	-0.8468
Q13	0.499	1	0.48	0.2294	-0.3514
Q14	0.1031	1	0.7481	0.0446	-0.8976

The DIF analysis shows that a degree of differential item functioning exists. Large, significant differences exist for questions 8 and 11, with question 11 showing DIF outside of Rasch model expectations. Questions 1 and 5 also show difference of marginal significance.

Question 11 asks students to rate whether working in a team to complete the experiment was beneficial. In the case of this experiment, students worked individually, rather than in groups or pairs. As a consequence, this survey item was often not answered at all, and was not responded to in any consistent manner. Apparent DIF for this survey item's responses is therefore neither unexpected nor problematic. Question 8 asks students about the efficacy of their practical demonstrator's supervision and guidance. This is the one aspect of the laboratory environment which changed between the two years, and again it is unsurprising that this specific item may show significant difference. It is possible that the two items showing marginal difference are cases of "artificial DIF" induced by the presence of other survey items with a large and significant genuine DIF, such as item 11. The issue of multiple comparisons (see section 2.4.1) may also play a part in the presence of some significant differences.

The majority of items report no substantial DIF, and it is reasonable to presume that the two sets of item measures for the two years are broadly similar in their location. As such, little error in the location of student measures estimated would be introduced, making a comparison of student biases between the two years viable. The distribution of student biases sampled in 2011 appears not to be normally distributed (Kolmogorov-Smirnov: D = 0.090, df = 141, p = 0.008 ; Shapiro-Wilk: W = 0.971, df = 141, p = 0.005). The 2012 data, however, appears to have a normal distribution of sampled student biases (Kolmogorov-Smirnov: D = 0.098, df = 77, p = 0.064 ; Shapiro-Wilk: W = 0.976, df = 77, p = 0.150). The significant non-normality of the 2011 data appears to arise because of a significant degree of kurtosis, and skew in the distribution appears not to be an issue (Table 15). The centre and extremities of the distribution of student biases therefore appear to have been sampled disproportionately in 2011, however, this should not influence the location of the centre of the distribution. The three extreme responses in 2011 removed from consideration in these analyses correspond to persons responding in the extreme positive category in every case. Even if these three

students were a significant enough proportion to generate an overall bias in the sample taken as a whole, this bias would be a positive one, masking significant improvements in 2012 rather than exaggerating them. The results of the main study (section 3.1) would therefore not be compromised.

Table 15: Student bias distribution statistics for the Analysis of spinach extracts experiment

year	Median	Mean	N	Std. Deviation	Kurtosis		Skewness	
					value	Std. Error	Skewness	Std. Error
2011	1.46	1.5653	141	1.15439	1.014	.406	.386	.204
2012	1.61	1.7962	77	1.12784	-.539	.541	.308	.274

Because the distribution of the 2011 data appears significantly non-normal, the distributions of student "ability" measures were compared using non-parametric statistical tests. Both the mean "ability" measure and the shape of the overall distribution of these measures were compared. The data clearly show that the median and distribution of student biases may be considered equivalent between the two years, as judged by Mood's median test (grand median = 1.61, $\chi^2 = 0.432$, *df* = 1, p = 0.5108) and the Mann-Whitney U test (standardised test statistic = 1.285, p = 0.1988). Fitted normal distributions are displayed in Figure 26 using black lines.

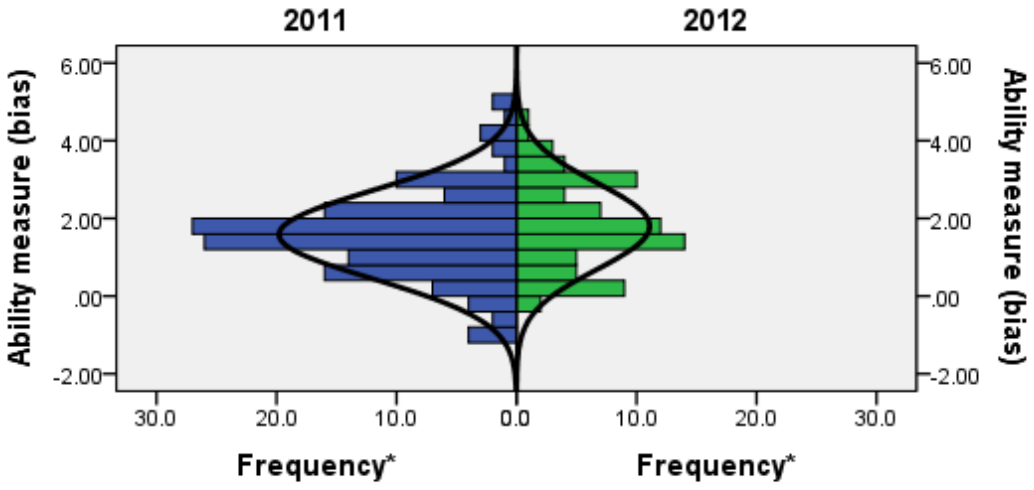


Figure 26: Student bias distributions for the Analysis of spinach extracts experiment

Overall, the distributions of broad-scale student biases for samples taken from the two different years appear equal, as judged using responses to this unchanged experiment. Though some items appear to be more "difficult" to answer positively for the 2012 cohort, potentially indicating a finer level of bias applicable to these questions only, this would mask significant improvements rather than exaggerate them. For this reason, the conclusions drawn in the main study appear valid.

3.2.4 Discussion

3.2.4.1 Summary of sample adequacy implications

A major outcome of this investigation is the evaluation of the adequacy of samples gathered for the initial score-based study (section 3.1), with particular focus on their representative nature or otherwise and their comparability between the two years. These two features are

required for the conclusions of the initial study to be valid and as such are summarised for each studied experiment individually below.

Vapour Pressure

The 2011 data appear unbiased, with the exception of one response which may be a 'donkey vote'. This response is not expected to influence results significantly.

The 2012 data appear significantly biased towards positive responses. This appears to have arisen because responses collected from Week 6 and Week 8 of the semester appear more positive than data from Week 10, and these were more frequently represented in the sample. It has been shown, however, that the magnitude of the bias present in the 2012 distribution is not sufficient to account for all differences observed in the main study data comparisons. It was also shown that the difficulty of responding positively to some (unidentified) items of the survey significantly differs between the two years, independent of student biases.

Biological Buffers

The 2011 data appear unbiased towards more positive or negative responses overall. Whilst the distribution of student biases in this sample does not show evidence of overall bias in either the positive or negative direction, its shape may not be representative of the broader population, as some kurtosis is evident.

The 2012 data appear unbiased. Neither student propensity to respond positively, nor the difficulty of responding to any item positively independent of bias, changes between the weeks in which the experiment may have been conducted. The majority of evidence suggests that the distribution of student biases sampled for this experiment approximates normality.

Copper (II) Ion Concentration

The 2011 data appear biased, but it is unknown whether the bias is positive or negative in nature.

The 2012 data appear biased, caused by skew in the distribution of biases sampled from Week 6 of the semester. It is unknown whether this bias is in the positive or negative direction. The "difficulty" of responding to any given item positively, independent of student bias, did not appear to differ between weeks of the semester.

The significant bias in the samples taken from the two cohorts for this experiment may potentially explain why no significant differences were detected in survey responses between the two years for this experiment in the main study, despite the fact clear differences between years are apparent in the case of the other two experiments.

Negative control: Analysis of spinach extracts

The student cohorts from 2011 and 2012 appear comparable. The distributions of student biases sampled from each year are equivalent. Some items appear more "difficult" to respond to positively for the 2012 cohort, potentially masking improvements in 2012 as opposed to exaggerating them.

3.2.4.2 Scores versus measures: incongruent results

As can be seen from the summary presented above, consideration of student bias effects has the potential to perturb the conclusions drawn from a score-based study. Whilst in this specific case the student samples obtained appear roughly acceptable, it needed to be confirmed that various features observed were not problematic numerous times over the course of the analysis. Even with typically greater than 50 observations for each sample, the distribution of biases often appeared skewed, opening the possibility that this bias could have perturbed the scored results obtained.

The case of the vapour pressure experiment highlights this point: scored results yielded an exaggeration of the actual cohort difference, thus it needed to be confirmed using Rasch measurement techniques that any genuine differences in experiment quality persisted after taking this into account. Once Rasch measurement was used to estimate the impact of student bias differences, only items 1 (concerning data interpretation skills development) and 14 (concerning overall learning experience) are confidently reported to differ between years. "False positives" were reported for three other survey items by the scored results prior to this amendment. It is conceivable that shifts in the distribution of student biases could similarly produce "false negatives" in other analyses, counteracting genuine differences in experiment quality which would otherwise be evident if using student independent measurements (i.e. Rasch measures).

Based on these observations, it appears that scored results do not necessarily yield the same conclusions as sample independent measures (Rasch measures) would. The fact that scores conflate student dependent and student independent effects allows the biases of individual student responses to perturb the results observed. A substantial problem is faced by researchers who observe a difference in scored results: it could be that experiment quality has changed or it could be that the students sampled have a different bias, but it is not a simple matter to tell the difference between the two. Further, when using scoring techniques, shifts in student bias could mask genuine differences in experiment quality which would otherwise be evident. The use of the traditional integer scoring method, therefore, places the researcher at risk of both type I and type II errors in absence of demonstrably comparable student samples.

Fortunately in this specific case, open response data was also available to affirm or refute the conclusions implied by the scored Likert-type data. It is unfortunate that this particular study could not make use of student identification numbers in order to connect subsets of the data gathered in different years, in order to make the same comparisons using Rasch measurement instead of scores. The fact remains, however, that genuine student independent measures and scores do not necessarily yield the same conclusions. Making more accurate comparisons requires student biases to be estimated from other data, such that the locations of student independent measures for each group compared can be estimated relative to them. Achieving this requires the identification of cases where the same student (presumably of the same bias) responds on multiple different occasions, allowing any differences in the 'bulk' distribution of student biases to be identifiable and able to be controlled for. Subsequent studies presented in this thesis will make use of student identification numbers to achieve this. Additionally, it remains unknown from this study precisely how much student bias varies. It may be the case that the observations of this study are a rare exception, whereas student biases typically may remain reasonably invariant overall. A more generalised approach to this topic is therefore needed, beyond this specific case. A general relationship between the Rasch measure for a

survey item and the corresponding mean scored response expected from the student population is needed to shed light on this issue.

3.2.5 Conclusion

Data presented here reveals the potential for scored ASLE survey Likert-type data to yield inaccurate conclusions. Whilst the broader scale conclusions of the previously presented research (section 3.1) appear to hold, the results of the Likert-type item comparisons need some degree of revision. Commonly utilised mean scoring techniques appear susceptible to the effects of student biases, conflating these effects with any changes in sample-independent factors, such as the "objective" quality of the experiment. Following the observation of this specific case of scores yielding subtly different conclusions to genuine interval scale measures, a more generalised investigation into the relationship between scored results and sample independent measures is needed. The extent to which student biases confound scored results in a general sense remains an open question, and achieving a more valid means to contrast experiments evaluated using the ASLE survey remains as a future goal. These two points are addressed in subsequent sections of this thesis.

3.3 Scoring responses to individual Likert-type items on the ASLE survey

3.3.1 Outline

Following the previous study's illustration that student biases may confound the results of comparisons made using integer scoring methods, the extent to which these effects are likely to be problematic in a more general sense became a pressing question. The research presented in this section now explores the relationship between traditional scores and corresponding Rasch measures far more explicitly, investigating the precise mathematical relationships between the two. This topic will be explored both at the level of individual response data, as well as mean scores obtained from entire samples. Through these analyses, the relative contribution of student bias effects on the variance in observed survey responses can be known, shedding light on whether mean scores are likely to be a valid means of indicating change in the "objective" quality of experiments evaluated. Knowledge of the relationship between scores and sample independent, interval scale (Rasch) measures will be used to explore the validity of the statistical treatment of ASLE response data, including the application of the integer scoring technique as well as parametric statistical methods.

At this stage, it will still be presumed that "objective" experiment quality can validly be attributed a numerical value. That is, it is presumed that valid interval scale measures of student independent experiment quality exist. The accuracy or otherwise of this supposition is necessary for these preliminary investigations, but will be addressed subsequently in this thesis (see section 4.1). Working under this presumption, Rasch models can be generated describing the relationship between score and measure for each item of the ASLE survey and from this, population level expectations can be extrapolated. In this way the relationship between Rasch measure and expected score can be known, more conclusively answering the question of whether integer scoring methods are a valid means of measuring experiment quality in general, presuming such a thing can be measured. The first primary hypothesis of this thesis:

> *Conclusions drawn from the ASLE survey data using typical scoring techniques resemble conclusions drawn using sample independent, interval scale measures extracted from the same data.*

will thus be supported or refuted at the conclusion of this study. It will also be determined whether parametric statistical techniques are appropriate for data treated using integer scoring techniques, if traditional integer scoring techniques are valid. Broadly speaking, this study evaluates the adequacy of the current methods used to extract measurements from survey responses. In generating Rasch models for the ASLE data based on current assumptions of the way ASLE surveys function, the presumed "measurement mechanism" by which observed responses are related to underlying latent traits may also begin to be explored.

3.3.2 Specific methods

3.3.2.1 Assumptions of ASLE survey use and Rasch model construction

Rasch models estimated during the course of this study were constructed based on the assumptions required for the usual uses of the ASLE survey data. These assumptions and their implications for the construction of Rasch models for the data are outlined below.

Assumption 1: Responses given by the same student are comparable between different occasions, and between different experiments.

For example, if a student responds "agree" to item three in Experiment A, and also responds "agree" to the same survey item in Experiment B, these responses may be interpreted equivalently in terms of what they imply regarding experiment quality. In terms of constructing a Rasch model of the data, this implies student measures; each student's propensity to provide positive response, must remain constant for all occasions. This assumption also requires that the structure of the rating scale remain constant for all occasions (experiments).

Assumption 2: Different survey items concern different topics, and responses to them therefore reflect different latent variables.

This implies two things for the construction of Rasch models. Firstly, if student independent factors are presumed to exist, a different measure must exist for each of the fourteen survey items. Secondly, students must each have fourteen different measures; one contributing to each of the fourteen different question specific latent variables. For example, a student's tendency to find experiments interesting is presumed independent to their tendency to report that the experiment provided them with the opportunity to take responsibility for their own learning, as these are fundamentally different topics. In general, any given student may (but not necessarily) have a different tendency to respond positively to a different survey item for similar reasons, and must therefore be assigned a different Rasch measure for each.

Assumption 3: Student independent measures of experiment quality exist and influence ASLE survey responses for the relevant experiment.

This assumption requires that there is some student independent component to observed responses: a measure of experiment quality, generally true for all students. Coupling this with the requirements of assumption 2 above, these measures must be specific to the question asked as well as the experiment being evaluated. Each experiment evaluated is thus presumed to have 14 different quality measures associated with it: one targeted by each Likert-type survey item.

Overall, these assumptions necessitate that a Rasch model of the ASLE data be constructed in the following way:

$$ln\left[\frac{P_{q,n,i}(X=x_k)}{P_{q,n,i}(X=x_{k-1})}\right]=\beta_{q,n}-\delta_{q,i}-\tau_{q,k} \qquad \textbf{32}$$

where the subscripts q, n and i denote specificity to the qth survey question, nth student responding and ith experiment evaluated respectively. The variable X is the observed response, equal to one of the five available response categories x_1 to x_5 and the values β and δ correspond to student dependent (bias/ "ability") and student independent (experiment quality/"difficulty") measures respectively. The τ parameters define the rating scale structure.

An interesting feature of this model is that no variable is independent of the question asked. From a measurement perspective, this implies there is no common point of reference for which measures specific to different questions may be assigned values relative to one another and relative to the common reference. This means that data obtained for different survey questions is necessarily disconnected (see section 2.3.3) and not directly comparable. For this

reason, fourteen distinct question specific Rasch models exist under these presumptions, and this is therefore how the data were modelled in the investigation discussed.

This model of the data is also intuitively sensible, given the deliberate multidimensionality of the ASLE survey instrument; each Likert-type item, by design, targets a different dimension of the laboratory learning experience. Here, a different unidimensional Rasch model is generated for the measurement of each distinct dimension of the survey, each with student "ability" measures and experiment "difficulty" measures specific to that dimension. The result is one Rasch model for each of the fourteen survey questions asked: the "items" within each Rasch model are simply the different circumstances (experiments) in which the identical question was asked of the students, with each circumstance having a different "difficulty" of providing positive response to that same question.

3.3.2.2 Data collection

Responses to ASLE surveys (Table 1) utilised in this investigation were gathered from first year undergraduate chemistry students at the University of Adelaide nearing the end of their laboratory sessions from late 2012 through to the end of 2013. These surveys were presented as optional, and allowed for the voluntary inclusion of the student's identification number. Responses from the 1127 different students who freely chose to provide these numbers were collated and used to estimate Rasch models for each of the ASLE survey's fourteen Likert-type questions. Those without identification were excluded out of practical necessity. A total of 33 experiments were evaluated using survey responses from these students, listed in the supporting information (Table S 40, section 7.3.1). Surveys were gathered from both the Chemistry IA/B and Foundations of Chemistry IA/B cohorts.

Experiments of the same title conducted by the two different student cohorts were not necessarily identical, with the Foundations of Chemistry forms of the experiments having been modified to suit the different student cohort in some cases (see section 2.1.3). Though the same experiments were present in both of the two years, small changes had been made in some cases and therefore each was considered a different experiment (with a different set of "difficulty" measures) for the purposes of the Rasch analyses. This data is consolidated in subsequent investigations, only after establishing it is valid to do so (see section 4).

The number of responses received from each experiment contributing to the analyses in this study is detailed in the supporting information (Table S 40, section 7.3.1) as well as the total number of responses received, not all of which could be used for these purposes owing to the lack of provision of a student identification number. Far more students provided responses in 2013, and of those, a higher proportion chose to provide their identification numbers.

3.3.2.3 Data cleaning process

Each separate question of the ASLE survey was analysed in isolation from the others, yielding a separate two-facet (experiments, students) rating scale Rasch model specific to each Likert-type survey question. Rasch models were generated using the Winsteps Rasch measurement software.[153] For each question separately, survey responses received which included a student identification number were collated and entered into the Winsteps software. Persons or items (experiments) which were reported by the software to be disconnected from the bulk of the data and present in isolated "subsets" were removed, as were students providing all responses in the extreme positive category or all responses in the extreme negative category (for

justification of these procedures, see section 2.3.3). This procedure was repeated until all remaining data points appeared connected as reported by Winsteps.

At this point, two different Rasch models were generated for the data remaining: one model in which the rating scale structure was constant for all experiments, and a second model in which the rating scale structure was allowed to differ between Chemistry IA/B experiments and Foundations of Chemistry IA/B experiments. The model which best explained the data as judged by the corrected Akaike Information Criterion (see section 2.5.4.2) was then used for subsequent data preparation and analysis.

In order to achieve the best estimates of the Rasch model, students significantly misfitting the model were, at this point, removed from the analysis (see section 2.3.2 for justification). Students selected to be removed were those for which the infit or outfit z value was outside of the +2 to -2 range. The z value was used to identify misfit as opposed to the mean square value due to the mean square's insensitivity to variance in the measures. It is acknowledged, therefore, that whilst those students removed misfit the model to a statistically significant degree, the actual magnitude of their misfit was not necessarily large. For this reason more students may have been removed than was necessary due to this conservative methodology. However, this is unlikely to cause significant issues as Rasch modelling is useful even with small sample sizes, preferably a minimum of 10 observations per response category.[239, 300]

Following this removal of misfitting students, the data were further examined for connectivity, and persons or items (experiments) appearing to be present in separate subsets to the rest of the data were removed. Extreme responding students were also again removed at this stage (see section 2.3.3 for information on both disconnected and extreme responses). Rasch models generated from the data remaining were used for the subsequent analyses presented. Details of the results of these models, as well as the numbers removed during data preparation steps previously described, are available in the supporting information (sections 7.3.4 - 7.3.17).

3.3.2.4 Generating score to measure relationships

For each Rasch model (one for each Likert-type survey question), the probability of observing each possible response (of score x) as reported by a student of measure β in response to an experiment of measure δ is calculated directly from the Rasch model. By modelling student measures as being normally distributed, an assumption made by initial estimation methods of the Winsteps software (see section 2.3.1), the probability P* of observing each given response (x_k) when sampling randomly from the whole student population, all of whom are responding to an experiment of quality measure δ with respect to the question being asked, was derived using the law of total probability.[301]

$$P^*(x_k, \delta_{i,m}) = \int_{-\infty}^{\infty} P(X = x_k) \times P(\beta) \,.\, d\beta \quad ; \quad P(\beta) \sim \mathcal{N}\left(\bar{\beta}, \sigma_\beta^2\right) \qquad \textbf{33}$$

Equation 33 results from taking the probability that a specific student of measure β will provide response x_k (obtained directly from the Rasch model, see Equation 1), multiplying that value by the probability of sampling a student of that specific β measure, then summating across the entire distribution of possible β measures. These functions, like the Rasch model equations from which they are derived, are specific to the survey item posed.

From these functions, evaluated using the Matlab software,[302] the distribution of expected mean ASELL scores as a function of the experiment measure δ were derived for each of the 14 Likert-type items of the ASLE survey. The population level probabilities of observing a response in each category, computed via Equation 33, were taken to approximate the expected population level proportion of responses in each category (the observed count c for the category divided by the total number of responses N).

$$\frac{c_k}{N} \approx P^*(x_k, \delta) \qquad \textbf{34}$$

These values were then used to compute the mean ASELL score in the usual way (see Equation 28). The expected mean ASELL score provided by the average student (the "fair average" associated with the relevant δ value) was also calculated, the mathematics of which has previously been described (see section 2.5.1). Standard error values (and subsequently 95% confidence intervals) for the expected population level mean ASELL scores were calculated at various sample sizes using standard statistical formulae previously discussed (see sections 2.4.1 and 2.4.3), again utilising the approximation shown in Equation 34.

3.3.2.5 Simulation of population level distributions

Population level probabilities generated using Equation 33 were used to simulate 5000 samples of 100 observations each for item 2 of the ASLE survey: "This experiment helped me to develop my laboratory skills", selected from the fourteen Likert-type items of the survey for reasons detailed in the subsequent discussion (see section 3.3.3.3). This simulation study was conducted in order to test the assumption that the mean scores obtained from random samples appear normally distributed about the true population mean; a requirement of parametric statistical methods. The population level proportion of responses in each of the five response categories for this item were evaluated at experiment measure δ=0, as this is by definition the measure of the average experiment in the sample. Random numbers ranging from zero to one were generated using Microsoft Excel's RAND() function, and from these random numbers, simulated responses in one of the five categories were assigned based on the random number values. The range of random number values corresponding to each assigned response was, in each case, selected such that the size of the random number range corresponded to the relative probability of that response category being observed.

Given the central limit theorem, sample mean scores will more closely approximate a normal distribution as sample size increases (see section 2.4.2, Equation 11). The task is therefore simply to evaluate the sample size at which the approximation is sufficiently close, and this was done based on Muthén and Muthén's criteria[303] for deciding sufficient sample size using randomly simulated data. Accordingly, the sample size was deemed to be sufficient once coverage (the proportion of the data falling within the expected 95% confidence interval) was consistently between 0.91 and 0.98, and the bias in the estimated standard error in the mean score was less than 5% in magnitude. Muthén and Muthén also recommend that bias in any parameters estimated (in this case the mean score) should not exceed 10%, however given that the location of the zero point of the scale influences this bias value and that the zero point on the ASELL scoring scale is arbitrary, this was not investigated. The estimated and true population means were, however, still compared.

3.3.3 Results

3.3.3.1 Features of models generated

Statistical details of the fourteen rating scale Rasch models generated, one model for each Likert-type survey item, are available in the supporting information (sections 7.3.4 - 7.3.17). This information includes the number of data points received in each response category, separation and reliability values for both persons and experiments, the raw variance in observed responses explained by the person and experiment measures as evaluated by Winsteps' variance decomposition, fit statistics for the response categories and for each model globally, measures and ranges for response categories, estimated Rasch-Andrich thresholds with associated standard errors as well as Rasch-Thurstone thresholds and estimated local discrimination values for each pair of adjacent response categories, histograms displaying distributions of estimated person and item measures and figures displaying empirical and observed response category probability curves. The interpretation of these statistics is described extensively in section 2.5.2: "Rasch model fit statistics and descriptive values".

Person measure reliability values in all fourteen models are very low. This is most likely a consequence of the fact that any given student only responded to a small number of experiments, making estimation of their Rasch measure imprecise. Experiment measures show better reliability values, though these vary broadly from high values near 0.95 through to low values closer to 0.6 in some cases. This variation may have arisen for a broad array of reasons. Regardless, this should not be problematic for this study, as the main focus here is the response scale structure. Sample sizes achieved appear adequate for the most part, however the most negative category received less than the recommended 10 data points for a number of survey items. The practical implication of this is that score to measure relationships presented here are imprecise for the extreme negative end of the response scales presented. This imprecision is reflected in large standard error values seen for the lowest Rasch-Andrich threshold parameter in each model. Poor alignment between person measures and experiment measures (poor targeting) in most cases may have contributed to this issue as well as the issues with reliability previously described. Better survey targeting could yield more accurate results than presented here, notably by including survey responses gathered from experiments which illicit more negative responses than the experiments studied in this research.

Poor survey targeting is also a likely contributor to the low proportions of observed variance explained by the experiment measures. Variance explained by experiment measures is under 10% for almost all survey items; a small fraction of the total variance explained by both student and experiment measures combined, often near 50% (which is acceptable). Most experiments have measures with substantially lower values than most person measures, making positive responses far more likely and therefore reducing the amount of variation in observed data attributable to measurement differences. This makes the relative contribution of random effects (such as the inherent imprecision of the five-point response scale) comparatively larger, reducing the precision and reliability of measurements achieved. Additionally, the distribution of student measures in each case appears much broader than the distribution of experiment measures, reducing the relative contribution of experiment-specific factors to the observed responses and depleting the variance explained by experiment measures. This could potentially mean that large samples need to be gathered in order to 'average out' these student factors and random effects when using scoring methods. The need for further investigation into the nature of these undesirable contributions to the ASLE survey

responses therefore warrants subsequent research, discussed later in this thesis (see section 4.1).

Despite these effects, fit statistics appear acceptable and the numbers of students removed due to misfit were not excessive, meaning that Rasch models generated likely provide a good representation of expectations for the bulk of the student population. The narrow range of experiment measures and the poor targeting achieved here is a property of this specific sample and not necessarily to be expected in all cases. Category structures and threshold parameters estimated should still be generalizable, despite the fact their associated errors could have been reduced if the experiments surveyed were more varied and better aligned with the distribution of student measures.

3.3.3.2 Relationship between mean ASELL scores and Rasch measures

Estimation of Rasch models for each survey item revealed the rating scale category structure for each item of the survey. An example is displayed below in Figure 27, and similar graphs are available for all items of the ASLE survey in the supporting information (sections 7.3.4 - 7.3.17).

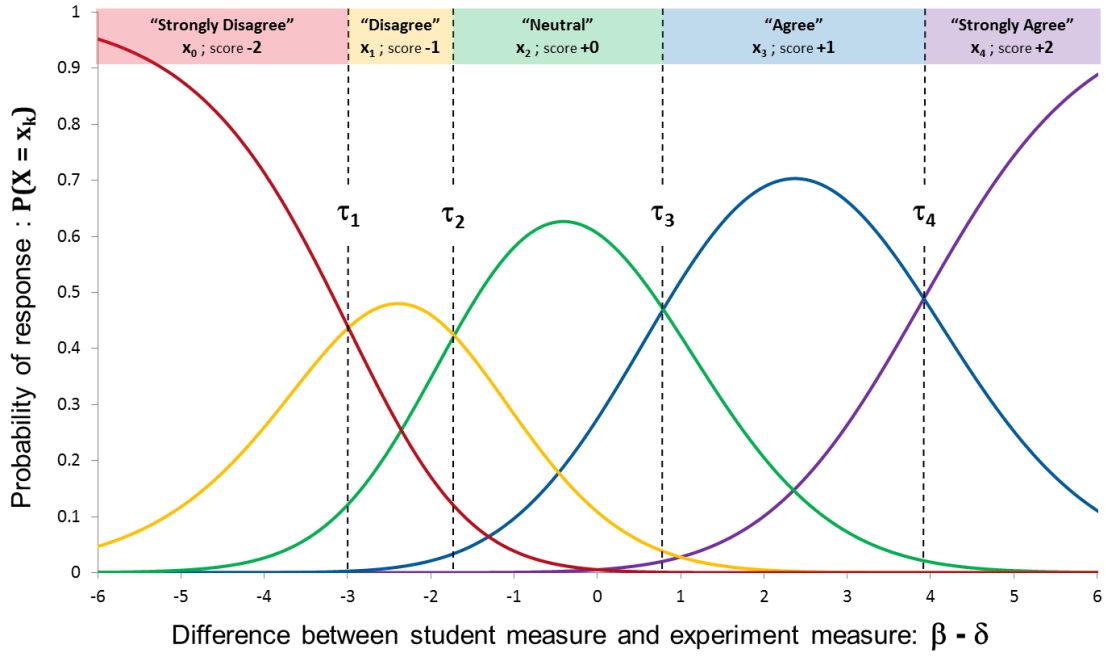


Figure 27: Response structure for item 3: "I found this to be an interesting experiment"

Colours of red, yellow, green, blue and purple have been used to illustrate the probability that a given student will provide response in the five available response categories (strongly disagree, disagree, neutral, agree and strongly agree respectively). The category threshold parameters of the Rasch model (τ_k) mark the boundaries between most probable response categories for the given latent trait measure ($\beta - \delta$ on the horizontal axis). Either higher student measure (β, reflecting bias) or lower experiment measure (δ, reflecting "difficulty" of providing positive response) imply that more positive responses become more probable.

A key feature of most rating scale category structures is that the categories do not appear "equidistant": more positive response categories typically gather a broader range of perceptions. The lack of consistency in the width of each category explains why the more positive categories are often associated with higher coherence values; it is more often accurate to infer a measure of perception from the category observed (and vice versa) in these cases, because in these ranges of perception the probability of responding in other categories

remains relatively low. By contrast, drawing inferences between observed response category and measure of perception may be inaccurate for the lower categories, as a high degree of overlap exists between category probability values. Shown below in Figure 27, for example, even at the most probable point of responding in the "disagree" category ($\beta - \delta \cong -2.5$), there is still only a 50% chance the student will respond with "disagree" rather than the adjacent response options. For reasons such as these, the frequency of accurate inference from response category to measure of perception often drops below 50%, with the exception of the second to most positive response category. This category, except in the case of item 13, universally gathers the most responses and covers the broadest range of perceptions, making inference from observed category to measure of perception accurate typically more than 70% of the time for this category.

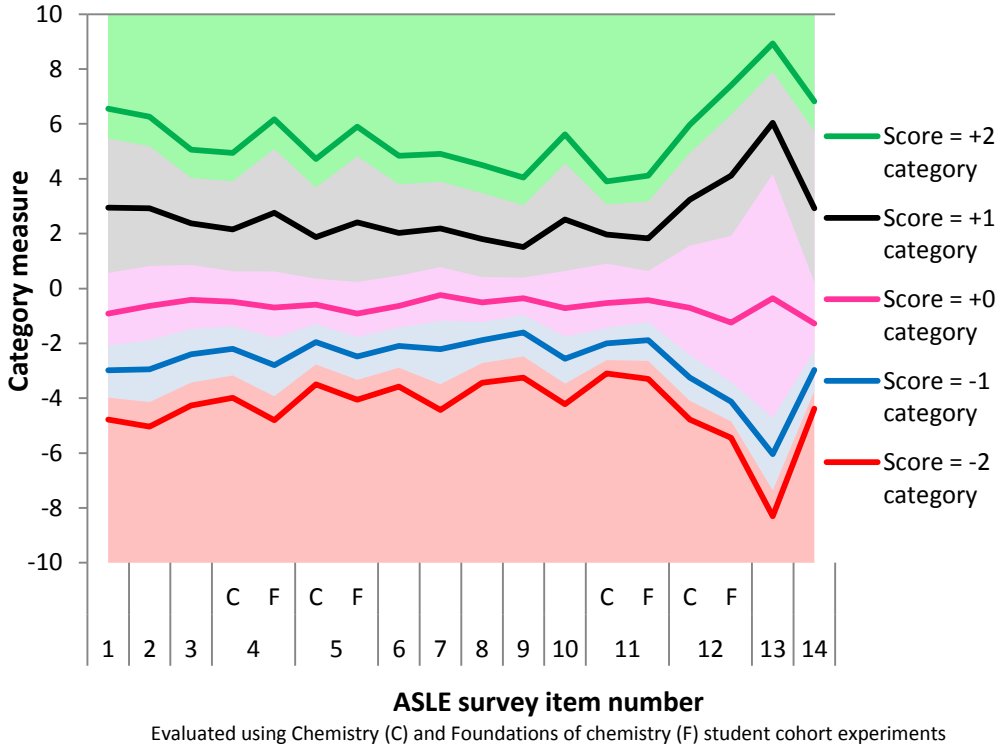


Figure 28: Category measures and ranges for the ASLE survey items

Solid lines indicate category measures: the measure ($\beta - \delta$, see Figure 27) at which responses observed would be expected to receive an average score of the value displayed in the legend. Boundaries between shaded regions reflect half-point thresholds: the measures ($\beta - \delta$) at which the expected average scored response is mid-way between integer value scores assigned to the response categories.

Category measures and ranges reported by *Winsteps* for the models generated were obtained for each item of the survey (Figure 28). Survey items 4, 5, 11 and 12 were best modelled as having a different response scale structure for the two different student cohorts, usually with slightly wider categories for the Foundations of Chemistry student cohort model. One key feature of the response scales highlighted by these measures is the fact that the measure of the centre category is typically a negative value. The zero-point of the category structure is the point of equal probability of responding in the two most extreme categories, and in all cases observed here, the centre category appears lower than this point. Figure 28 above again displays the lack of consistency in the width of the five available response categories, for each item of the survey. The relationship between Rasch measures and corresponding score values

for each category shows a clear lack of equivalence between the magnitude of a change in score and the magnitude of a chance in perception. The category structure also appears to vary between different items of the survey instrument. For example, item 13 shows a category structure with a very wide centre category compared to the other items.

The impact of this variation and the inequality of the spacing between categories of response on the generation of mean ASELL scores was investigated further, by generating the relationships between experiment measure (δ) and expected mean score. An example of this is displayed in Figure 29.

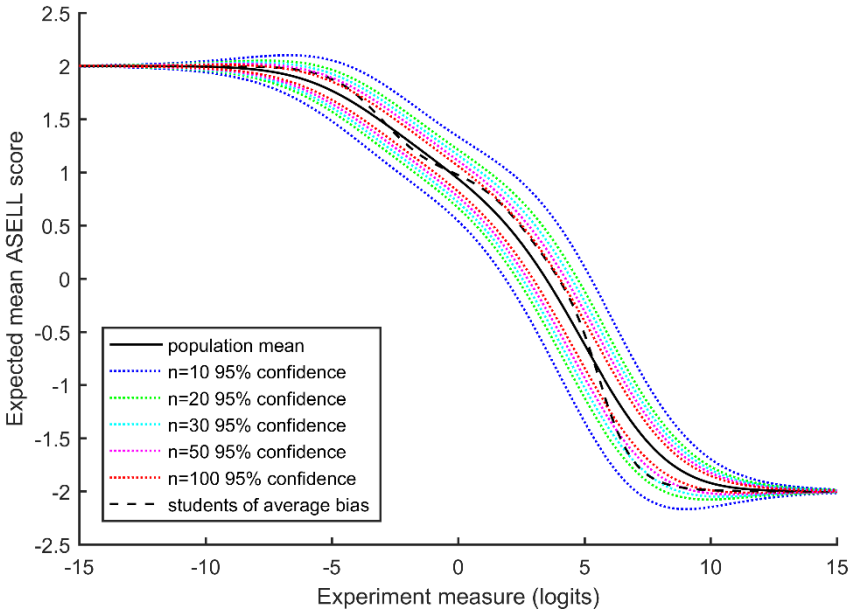


Figure 29: Expected mean ASELL scores for item 14: "Overall, as a learning experience, I would rate this experiment as"

The "Experiment measure" corresponds to the δ value referenced continually throughout this study: the student independent measure of "difficulty" in providing a positive response to the item. The plot shows a non-linear relationship between this truly interval-scale measure of objective experiment quality and the mean score value expected to be observed when following typical ASELL data treatment methods. Dotted lines display the variation in expected mean scores at a variety of sample sizes, with less variation observed when more samples are gathered.

The change in mean ASELL score as a function of the Rasch experiment measure does not appear to be strictly linear, though appears roughly so in the region in which non-extreme ASELL scores (-1.5 to 1.5) would be received. A roughly sigmoidal curve is observed in most cases, with the maximum and minimum mean ASELL scores of plus and minus two observed beyond approximately -5 or +10 logits respectively from the measure of this sample's average experiment (experiment measure $\delta = 0$). Item 13, regarding the time available, appears different; the expected mean ASELL score plateaus and does not substantially change from zero for approximately 3 logits either side of the average experiment's measure (Figure S 48, p.259). The standard deviation in experiment measures for this item is roughly two logits based on this sample, meaning that ASELL scores for item 13 are extremely insensitive to changes. This is a limitation of the scoring system when applied to this item. Non-linearity in the relationship between ASELL score and measure is not unique to item 13, however. The expected mean ASELL score received from the average student does not always align with the

expected mean score received from the whole student population, suggesting possible skew in the distribution of expected mean ASELL scores at locations where these values differ. This may be important for obtaining the sample sizes required. Regions of the scale which produce a more skewed distribution of mean ASELL scores would require larger sample sizes for the purposes of statistical comparisons.[173]

In order to investigate the degree to which the non-linear relationship between score and measure impacts the validity of the use of scores, the correlation between mean ASELL score and estimated Rasch measure was calculated (Figure 30). Scores were calculated using both the traditional ASELL integer scores and by using the category measures obtained by Rasch modelling as "optimised" scores. This was conducted using data points used for the Rasch measure estimations and repeated using all data points collected prior to the data cleaning process. Because of the small number of responses received in 2012, correlations between scores measures were also separately evaluated for the 2013 data only.

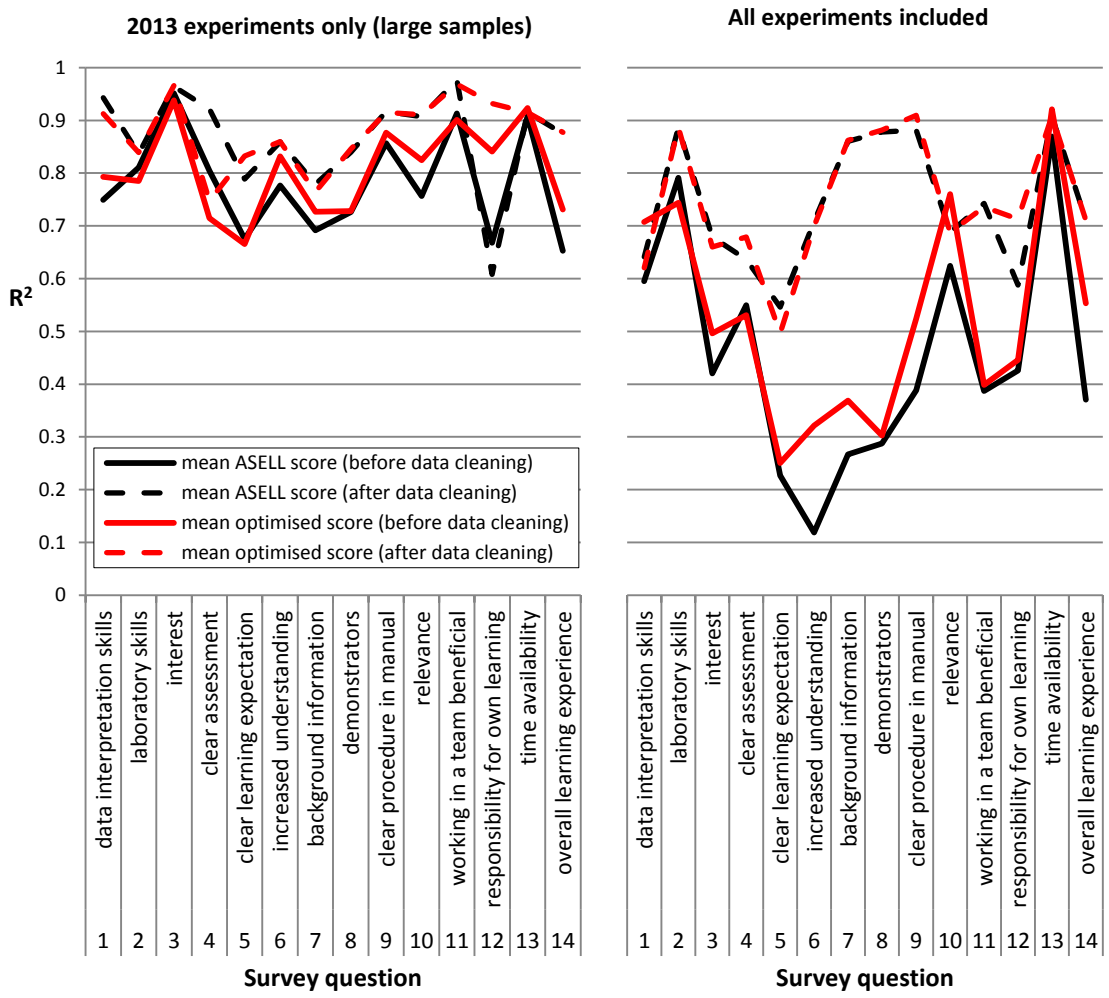


Figure 30: Observed similarity between mean score values and estimated Rasch measures

Data cleaning processes noted in the legend include removal of extreme, disconnected and substantially misfitting responses. Data points remaining after these cleaning processes reflect those which may be adequately described by Rasch models (thus presumed not to be "donkey votes" or inconsistent with the majority). The "optimised" scores referred to in the figure utilise category measures (see Figure 28) in place of the usual integer score values assigned in order to better reflect the true magnitude of difference between the five rating scale options.

Results show that scores generally correlate reasonably strongly to the interval scale Rasch measures, with little difference between the integer scoring system and the optimal scores obtained from using category measures. The data cleaning process does not appear to have altered this. However, small sample sizes appear to severely deteriorate this correlation between score and measure. This can also be seen in the broadening of the 95% confidence intervals of expected mean ASELL scores in the score to measure relationship plots for each item's Rasch model (such as the example shown in Figure 29). Data misfitting the Rasch model, such as "donkey votes" and students dissenting from the majority, appear to severely enhance the lack of correlation for small samples.

3.3.3.3 Simulation study: the appropriateness of parametric statistics

A simulated distribution of mean ASELL scores received for an average experiment ($\delta = 0$) in response to item 2 of the ASLE survey: "this experiment allowed me to develop my laboratory skills" is shown in Figure 31. Simulations have been repeated for a range of sample sizes to obtain the distributions shown (5000 simulated samples of the relevant size in each case).

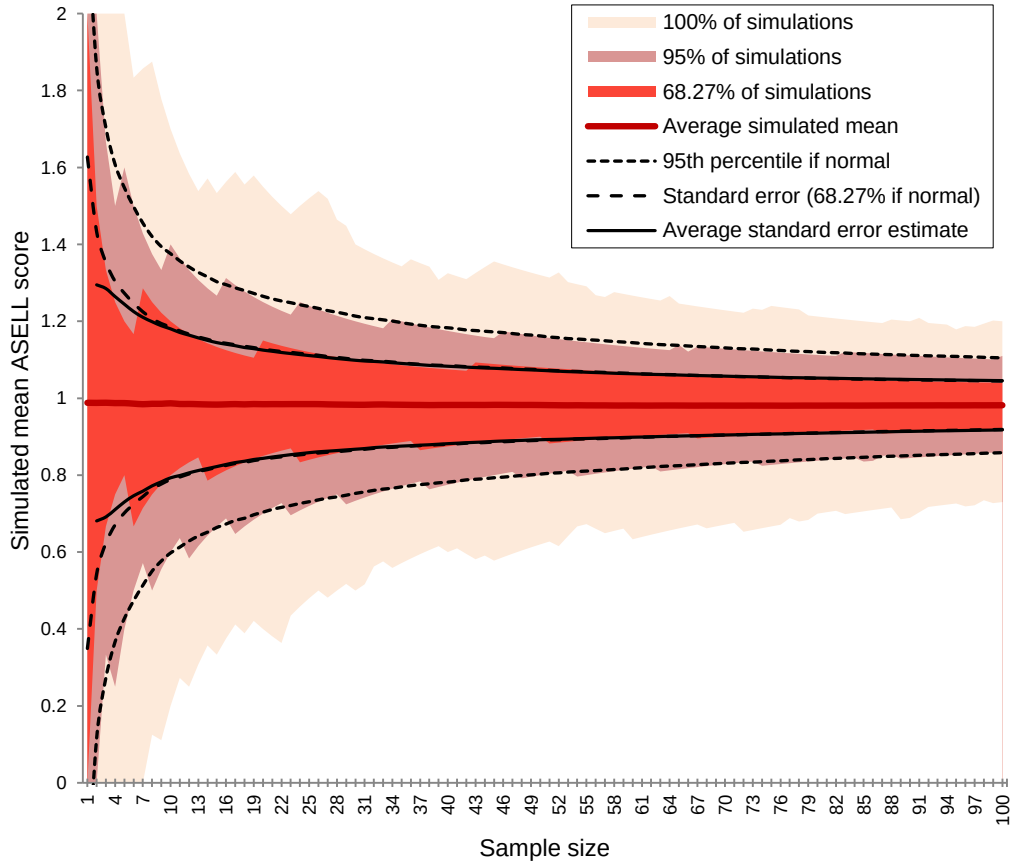


Figure 31: Distribution of mean ASELL scores obtained from samples of various size, in response to item 2 of the ASLE survey

Alignment between the observed simulated distributions (shaded red and orange) and the predictions of a normally distributed data set (dotted lines) indicate that the presumptions of parametric statistical methods are closely met. Results were generated using parameters estimated for item 2 of the ASLE survey: "This experiment helped me to develop my laboratory skills", detailed in Table S 42 of the supporting information (section 7.3.5).

Survey item 2 was chosen for this simulation study for a number of reasons. Empirically, this question shows the highest variance in observed responses explained by the experiment measures (as opposed to student factors) of all questions examined (23.8%), and also one of the highest total proportions of variance explained by the Rasch model as a whole (55.2%). The person measures also appear at least roughly normally distributed, meaning the normal distribution approximation used to generate the measure to expected mean ASELL score relationships is more likely accurate. The category structure also appears to have good fit statistics, with a sufficient number of samples gathered in each category to give an acceptable approximation of the category parameters. This survey item also shows one of the highest experiment measure reliability values (0.94).

The average mean ASELL score obtained in the simulations closely aligns with the true population mean score of 0.98 for all sample sizes. The average estimate of the standard error, however, appears smaller than the true standard error for small sample sizes. The bias in the average standard error estimate reaches acceptable criteria (below 5% bias) by a sample size of eight (precise values not shown). The 95% confidence interval assumed by parametric statistics (assuming normal distribution of mean scores) aligns well with the empirical region containing 95% of all simulated means obtained for all sample sizes, improving as sample size increases as expected. The proportion of simulated means falling within the 95% confidence interval assumed by parametric statistics reaches acceptable levels (between 0.91 and 0.98) after only two samples. Methods of calculating the sample size required for a normal distribution in population means based on the skewness of the distribution of observed data, such as Boos and Hughes-Oliver's[173] suggestion of $(5.66 \times skew)^2$ for a two-sided test or Cochran's rough guideline[172] of $25 \times skew^2$ appear roughly accurate, recommending sample sizes greater than approximately 5 and 4 respectively. Criteria recommended by Boos and Hughes-Oliver[173] are suggested based on achieving coverage of at least 0.94, and this is achieved in all but few cases beyond the sample size recommended by their method.

3.3.4 Discussion

The structure of the response scale categories yielded by Rasch models reveals a number of important results. Firstly, the variation in category structure between the different survey items removes the ability to interpret scores in the same way for different items of the survey. For example, a mean score of +1.2 on item 3 (interest) and mean score of 0.8 on item 6 (understanding) does not necessarily imply that students responded more positively regarding their interest than they did regarding their understanding. It could be that the position of a truly "neutral" perception on each item's response scale differs, or that the five response categories for one item gather different ranges of student perceptions than is the case for other survey items. The fact that response categories for all items are not equidistant also implies that at the level of individual responses the integer scoring system is strictly inaccurate. Given the variation in category structure between different survey items, the most accurate scoring system for one survey item is also not necessarily the same as for another.

The relationship between mean ASELL score and the interval scale Rasch measure is not linear, implying that changes in mean ASELL scores reflect different sized changes in the variable of interest depending on the scale location. This reflects the non-additive nature of scores as opposed to measures often referenced in the literature advocating Rasch modelling over scoring systems. Mean ASELL scores do, however, strictly increase as the latent variable of interest increases, and for this reason are of practical use as an indicator of change. Changes in an experiment's tendency to illicit positive responses can validly be inferred from observing a

change in mean ASELL score, but the magnitude of the change in ASELL score cannot be used to infer the magnitude of the change in the latent variable of interest.

However, despite the strictly inaccurate nature of the integer scoring system, taking the average ASELL score still appears to be useful. Using ideal category measures in place of the usual integer values does not appear to greatly improve the validity of using scores as measures, and scores generally tend to correlate highly to the measures of interest provided sample size is sufficient. This congruence between Rasch measure and observed mean score aligns with the pre-existing statistics literature showing good agreement between item response theory based measures and scores assigned using classical test theory.

Scored individual student responses, however, appear to have highly variable meaning. One student's response scoring +1 could feasibly correspond to a less positive perception of the experiment compared to another student's response scoring +0, based on the high degree of overlap in the rating scale categories (see Figure 27 previously). It is unclear how much this compromises the validity of rank-based non-parametric statistical tests, which have previously been used to avoid the controversies of using parametric methods on ordered categorical data. It appears that at the level of individual responses, the data may not even classify as being consistently ordered.

The validity of using scores as if they were a true measure of the variable of interest appears to be the key problem with small sample sizes, rather than the validity of using parametric statistics. Correlations between scores and Rasch measures drop sharply if including data points of small sample size, likely due to the large variation in student measures which would otherwise be 'averaged out' over a large sample. Parametric statistics, however, appear entirely appropriate for mean ASELL scores from very low samples, as the assumption of normally distributed mean values is met. Mean ASELL scores are therefore fit for t-tests, ANOVA and other parametric techniques. Again, however, it should be emphasised that change in mean ASELL scores cannot be used to infer the actual magnitude of the change in the variable of interest. The non-linear relationship between scores and measures also means some differences in the trait of interest may not be detected in the mean scores. Using scores instead of measures therefore appears to put the researcher at risk of "type two errors": inferring no difference when in fact there is one.

3.3.5 Conclusion

Despite common criticisms of using successive integer score values for the analysis of individual Likert-type items, it has been shown here that using mean scores in the case of the ASLE survey provides some practical use. Presence of differences in the latent variable of interest may be validly inferred from differences in mean scores, and mean scores appear to be fit for parametric statistical methods such as t-tests and ANOVA. Researchers using scores rather than more sophisticated measures such as those obtained through Rasch modelling should, however, be conscious of the non-linear relationship between mean scores and the underlying variable of interest they are intended to reflect. Mean scores generally correlate well to Rasch measures if sample size is large, however the magnitude of mean score differences is not informative of the true magnitude of any difference in the underlying variable. The presence of experiment independent factors such as a wide variety of student biases appears to threaten the validity of using mean scores as experiment-specific measurements at small sample sizes, warranting further investigation into the nature of these effects.

4 *Qualitative interpretations and the ASLE survey data*

In this section, the appropriate interpretation of ASLE survey data is explored in a more qualitative sense, exploring questions regarding the connection between survey data and any measures of experiment quality which may be considered largely student independent. These questions are investigated by refining different models of survey response, seeking the best explanatory model of the observed data. In so doing, this section tests the second and third primary hypotheses of this thesis:

> ***Hypothesis 2:***
> *Student independent contributions to the ASLE survey responses exist and are measurable.*
>
> ***Hypothesis 3:***
> *Student independent measures obtained from ASLE survey data reflect qualities of the experiment evaluated.*

Section 4.1 considers an array of possible different interactions between students and the ASLE surveys, encapsulating each as a different Rasch model. The observed data are then fit to each Rasch model, using both fit to the model and parsimony of the model to determine the best explanation of the observed data. A principal question answered within this section is whether any student independent qualities, emergent from experiment design, can be said to contribute.

Section 4.2 makes use of the best explanatory model of the ASLE data determined in the previous study, comparing the perceptions of male and female students. Whilst this section does not directly test either of the hypotheses reproduced above, it was considered worthwhile to answer questions regarding gender equality in the course given the opportunity. The study exemplifies the ability of Rasch methodology to investigate more deeply than scored analyses are capable, separately investigating student dependent and student independent measures.

Section 4.3 probes the relationship between student independent measures associated with ASLE survey data and factors of the learning experience. A technique is developed and implemented to explain ASLE survey student independent measures as a function of more basic elements of the laboratory experience, encapsulating the relationships within a Linear Logistic Test Model (LLTM). Notably, the LLTM structure is estimated from observed data rather than stipulated a priori, as would usually be the case.

Section 4.4 examines the relationships between different facets of the laboratory learning experience, as revealed by the LLTM derived in the previous section. Particular focus is given to the connection between generalisable measures of quality and the design features of the learning activity, addressing Hypothesis 3 above. Relationships uncovered reveal important conclusions for the pedagogy of science in the laboratory setting.

4.1 Valid measurement of experiment quality using the ASELL project surveys

4.1.1 Outline

In the previous study, the ASLE surveys were analysed using Rasch measurement techniques, presuming the validity of assigning interval scale measures. Estimates of experiment and question specific measures were found to contribute very little to the variance in observed responses, whilst a broad range of different student specific effects was the dominant source of variation. Whilst gathering large student samples may average out these student variations, the small contribution of the assumed experiment and question specific factors raises the question of whether it would serve as a better explanation of the data to disregard these factors entirely. This study therefore employs a number of Rasch measurement techniques to test a critical assumption of the ASLE surveys: that survey responses gathered from a particular experiment genuinely reflect some quality of that specific experiment. Rather than making assumptions regarding what the survey responses reflect and regarding their comparability between different survey occasions, a broader data set of ASLE responses is modelled under an array of different interpretations, varying the question specific and/or experiment specific nature of student dependent and student independent factors. Accounting for both the fit and parsimony of these different models, the Rasch model which serves as the best explanation of the ASLE survey responses is determined. The best model established here provides a foundation to begin more detailed investigations into why experiment measures take on the values that they do, allowing for the development and refinement of a specification equation deriving experiment quality measures directly from empirical features of laboratory activity design. Results obtained in this study also reveal information about the comparability of data obtained on different occasions, and the errors likely to be introduced by these effects are discussed.

4.1.2 Specific methods

4.1.2.1 Data collection and initial treatment

The same data set used to conduct the statistical techniques validation study (section 3.3) was used to generate a variety of separate two-facet and three-facet Rasch models, each explaining the observed responses in a different way. These models were then statistically contrasted in order to determine which model served as the best explanation of the data available (using techniques to be discussed). In order to allow for comparative statistical tests between these models, the same exact data points must be used for each. Data points which contributed to an ‘extreme’ measure in any of these models were therefore removed from consideration by practical necessity, as these responses have indefinite associated statistics (see section 2.3.3). Additionally, it was ensured that all data points used did not appear disconnected from the bulk of the other data gathered for the same survey item, or from the same experiment (unless this was an artefact of the model used). This resulted in a total of 45,641 data points being used to generate each of these Rasch models, gathered over a period of time from the second half of 2012 through to the end of 2013. Surveys gathered include responses from both the Chemistry IA/B and Foundations of Chemistry IA/B cohorts, some of which are associated with student identification numbers.

4.1.2.2 Facets of the explanatory models generated

The generation of ASLE responses was conceptualised as the interaction of three basic components: the student responding, the experiment conducted and the questions asked. Building on this notion, a total of six possible facets which may contribute to an explanatory model of the ASLE data were identified. Of these facets, those which take different values depending on which student is responding were deemed '*student dependent*' facets, whereas those which take the same value regardless of which student is responding were deemed '*student independent*' facets. An initial presumption was made that the best explanatory model of the data would include at least one student dependent facet and at least one student independent facet, though this presumption was subsequently tested following the generation and analyses of these models, as described in this study's results.

Table 16 and Table 17 provide a description of each facet potentially included in explanatory models explored. Displayed are generalised descriptions of the manner in which different facet element numbers were assigned, using example element numbers. Each different facet element number corresponds to a different assigned measure value for that facet under the relevant circumstances, and conversely the same facet element number implies the assignment of the same measure value. More rigorous mathematical justification of Rasch model formulations including facets which vary in this manner are provided in the supporting information (section 7.4.1). Notations used for each facet here highlight the manner in which each facet is assigned different element numbers, whilst notations used in the supporting information are selected to best highlight the mathematics underpinning the facets' derivations.

Table 16: Student independent facets

E

Specific to: **Experiment**

Independent of: **Question Student**

Experiment	Student	Question		
		q.1	q.2	q.3
Ex.A	St.1	1	1	1
Ex.A	St.2	1	1	1
Ex.B	St.1	2	2	2
Ex.B	St.2	2	2	2

Each experiment is assigned a different measure. These values remain the same regardless of which student is responding or which question is asked.

Q

Specific to: **Question**

Independent of: **Experiment Student**

Experiment	Student	Question		
		q.1	q.2	q.3
Ex.A	St.1	1	2	3
Ex.A	St.2	1	2	3
Ex.B	St.1	1	2	3
Ex.B	St.2	1	2	3

Each question is assigned a different measure. These values remain the same regardless of which student is responding or which experiment is being evaluated.

δ

Specific to: **Experiment Question**

Independent of: **Student**

Experiment	Student	Question		
		q.1	q.2	q.3
Ex.A	St.1	1	2	3
Ex.A	St.2	1	2	3
Ex.B	St.1	4	5	6
Ex.B	St.2	4	5	6

Each experiment is assigned a different measure for each different question. These values remain the same regardless of which student is responding.

Table 17: Student dependent facets

β	β_E	β_Q
Specific to: **Student** Independent of: **Experiment Question**	Specific to: **Student Experiment** Independent of: **Question**	Specific to: **Student Question** Independent of: **Experiment**
Each student is assigned a different measure. These values remain the same regardless of which experiment is being evaluated or which question is asked.	Each student is assigned a different measure for each different experiment evaluated. These values remain the same regardless of which question asked.	Each student is assigned a different measure for each different survey question. These values remain the same regardless of which experiment is being evaluated.

Experiment	Student	Question q.1	q.2	q.3
Ex.A	St.1	1	1	1
Ex.A	St.2	2	2	2
Ex.B	St.1	1	1	1
Ex.B	St.2	2	2	2

Experiment	Student	Question q.1	q.2	q.3
Ex.A	St.1	1	1	1
Ex.A	St.2	2	2	2
Ex.B	St.1	3	3	3
Ex.B	St.2	4	4	4

Experiment	Student	Question q.1	q.2	q.3
Ex.A	St.1	1	2	3
Ex.A	St.2	4	5	6
Ex.B	St.1	1	2	3
Ex.B	St.2	4	5	6

An array of nine possible two –facet models for the ASLE data was therefore determined by modelling the latent trait underpinning responses as the interaction of one student dependent facet and one student independent facet. Two other possible models were also identified; one three facet model containing all three broadly specific facets (those specific only to students, only to experiments or only to questions), as well as another three-facet model comprised of all three jointly specific facets (experiment and question specific, experiment and student specific, question and student specific). All models aside from these would contain redundancies and be reducible to one of these eleven,[ii] leaving these eleven models to be the full range of possibilities. All Rasch models generated were of the form displayed in Equation 2 (section 2.2.1), with the Andrich threshold (τ) parameters differing for different survey questions (i.e. 14 scale groups, one for each survey item). Table 18 displays the way the latent trait variable φ was modelled under each explanatory model. A Rasch model was generated for the observed data under each of these interpretations, recording the log-likelihood chi-square value and the number of free parameters estimated for each. These values were used to calculate the corrected Akaike Information Criterion (AICc) for each model, and the model with the lowest AICc value was taken to be the best explanatory model for the observed data (see section 2.5.4.2).

[ii] For example, a three-facet model β-δ-Q would be equivalent to a two-facet model β-δ, as the δ measures, jointly specific to both experiment and question, would already embody any broad scale question specific variation otherwise encompassed by the Q measures. This arises for any model in which one facet, broadly specific to one component contributing to responses, is paired with another facet already jointly specific to that same component and another (for this reason, the β_Q-Q model and the β_E-E model are actually equivalent to one-facet models β_Q and β_E respectively). Further details are available in the Rasch model derivations provided in the supporting information (section 7.4.1)

Table 18: Mathematical form of the latent trait variable φ in each explanatory model

		Student independent factors					
		Experiment and question specific		Question specific		Experiment specific	
Student dependent factors	Broad (non-specific)	$\beta - \delta$	A	$\beta - Q$	G	$\beta - E$	F
	Question specific	$\beta_Q - \delta$	I	$\beta_Q - Q$	J	$\beta_Q - E$	C
	Experiment specific	$\beta_E - \delta$	H	$\beta_E - Q$	B	$\beta_E - E$	K
Other models		$(\beta_Q + \beta_E) - \delta$ [all jointly specific facets]	E	$\beta - (E + Q)$ [all broadly specific facets]			D

Each Rasch model in the table above is derived from first principles in the supporting information (section 7.4.1). Labels shown at the right of each model formulation (for example A) also appear next to the appropriate corresponding equation in the model derivations. Notations for each facet used in the main body (those shown above) are not the same as the more complex notations used in the model derivations. Notation here is based on how the factors vary and the way in which facet element numbers are assigned, whilst notation in the model derivations is based on the mathematical structure of each facet's derivation.

4.1.2.3 Further investigation using an expanded data set

Upon determination of the best explanatory model, further investigation was conducted using an expanded data set, consisting of survey responses gathered from many experiments conducted at the University of Adelaide from times ranging from 2010 to mid-2014. A broader data set of survey responses was used for this subsequent investigation due to the fact that not only had more responses been collected by this time, but also because the best explanatory model determined allows for the use of survey responses without associated student identification (for reasons to be discussed). The use of these responses was not possible in the initial investigation due to the nature of some possible models (e.g. those where student measures remained constant throughout all experiments). This expanded data set made use of 9,287 surveys gathered after removal of extreme responses, composed of 128,811 data points. More details regarding this expanded data set are available in Table S 56, presented in the supporting information (section 7.4.2). Fit statistics associated with measure estimates in the model determined to be the best explanation of the data were recorded to further assess construct validity.

4.1.3 Best explanation of ASLE data

Modelling student dependent factors to be question specific appears to be the worst explanation of the data, as demonstrated by the fact such models have the highest AICc values of all (Figure 32). A significant improvement is gained by assuming the bias of each student to be non-specific; constant regardless of which question is asked or which experiment is being conducted. However, the best, most parsimonious explanations of the ASLE survey data appear to be provided by models which allow the bias of each student to differ between different experiments. The model explaining the highest portion of observed data (66.64%) is the $\beta_E+\beta_Q-\delta$ model, in which student dependent factors are modelled as having both an experiment specific component and a separate question specific component, whilst student independent factors are modelled as being jointly specific to both question and experiment. The high portion of data explained is not worth the substantial lack of parsimony imparted by

the many modelled parameters needed to do so, however. The extra variance explained here appears not to be due to the model accounting for consistent and significantly evident trends in the data, but rather the model's extra parameters allowing predictions to fit more closely to random 'noise'.

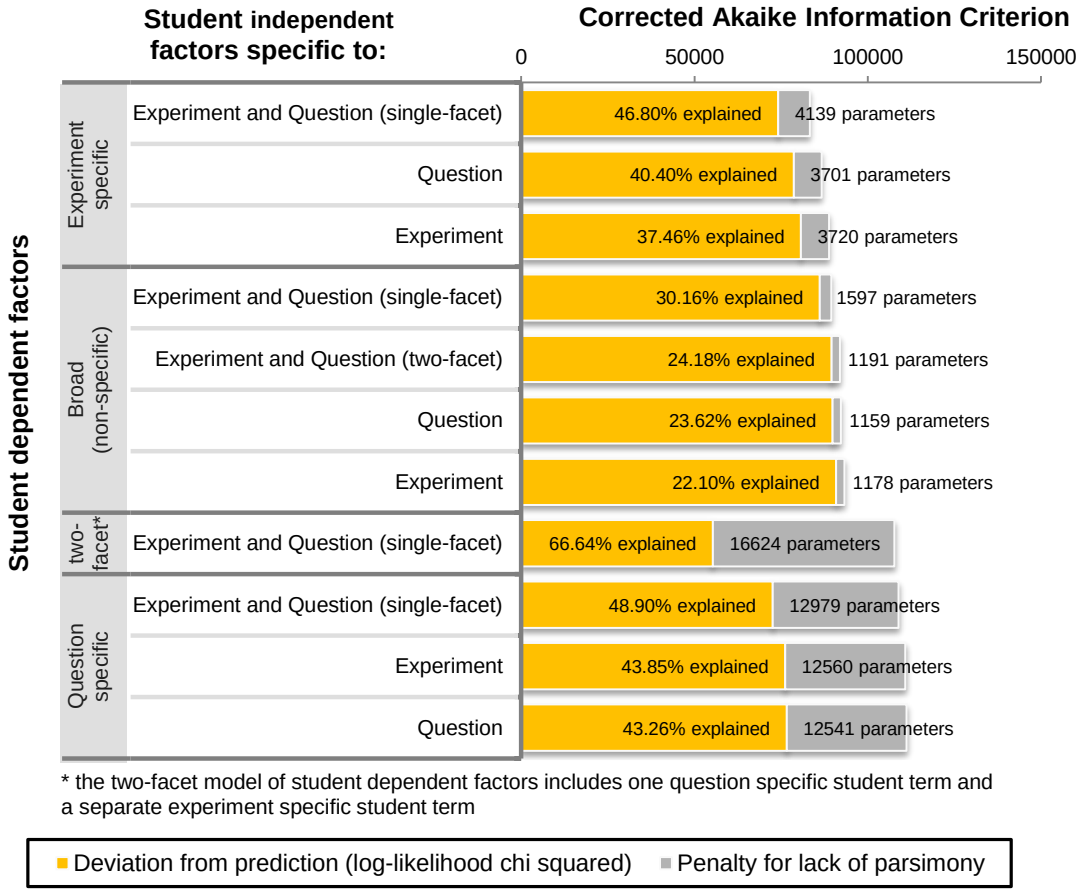


Figure 32: Explanatory Rasch models for the ASLE survey data

Models are arranged from best explanation of the observed data to worst, grouped based on the way in which students dependent factors appear to vary. Model descriptions correspond to those presented in Table 18. For each model, student dependent factors are modelled as either "broad (non-specific)", "question specific", "experiment specific" or "two-facet", using β, β_Q, β_E or $(\beta_Q+\beta_E)$ respectively (see Table 17). Student independent factors are modelled as specific to "experiment", "question", "experiment and question (single-facet)" or "experiment and question (two-facet)" using E, Q, δ or (E+Q) respectively (see Table 16).

Both the question specific and experiment specific interpretations of student factors drastically increase the proportion of observed data explained by the models, as compared to modelling each student's bias as broad and non-specific. However, only the experiment-specific student factor models achieve this in a sufficiently parsimonious manner. Assuming each student to have fourteen different bias parameters, one for each different question they may be asked, is extremely costly in terms of parsimony. Thousands of extra parameters need to be modelled under this interpretation compared to simply assuming a single bias parameter per student, and the extra portion of the data explained does not make up for this substantial cost. Assuming each student has a different bias value for each experiment they conduct, on the other hand, achieves a similar amount of extra observed data explained by the model whilst modelling far, far fewer extra parameters to do so. Crucially, the fact that the AICc values for

these models (those where student factors are experiment specific) are lower than the alternatives where student measures are broad and non-specific demonstrates conclusively that the additional parameters are explaining real, consistent variations in the data. Modelling student biases as if they differ for each experiment the student conducts, rather than remaining constant, provides a better explanation of ASLE survey results.

Based on this assessment, student dependent contributions to the ASLE survey responses appear to vary from experiment to experiment, and do so in a manner not equivalent between different students. One student's tendency to respond positively may increase from one experiment to the next, whilst another student's may decrease between the exact same two experiments. This appears to be the case regardless of how student independent factors are conceptualised, and is present to such a significant degree that ignoring these effects sharply decreases the proportion of observed data able to be accounted for.

Student independent factors contributing to ASLE survey responses appear to be best modelled as being both experiment and question specific. Removing the experiment specific nature of these factors consistently provides a worse explanation of the data, as does the removal of the question specific nature of the data. This trend is true regardless of how student dependent factors are conceptualised. The null-hypothesis that these student independent factors are non-existent was firmly rejected in the best explanatory model of the data (χ^2= 7324.9, df=451, p<<0.001), and modelling the student effects on their own (β_E) yielded a worse explanatory model by comparison (log-likelihood χ^2=80729.9, 3688 free parameters, 37.47% of variance explained, ΔAICc=+5476.4 compared to the β_E-δ model). This data collectively provides strong support for the hypothesis that ASLE survey data reflects student independent factors which are specific to both the experiment evaluated and the question asked.

Given this aligns with what would be expected if the usual assumptions about the way the ASLE survey works were true (see section 3.3.2.1), this is a promising result. However, the matter is complicated by the way the student dependent effects appear to function. The previous study discussed the usual assumptions of utilising the ASLE surveys, and the necessary implications these assumptions had for the construct of Rasch models of the data. One assumption discussed was the comparability of data across various different occasions, and it was described that this assumption requires that student measures remain invariant between different survey occasions (see section 3.3.2.1). As seen in Figure 32, the best model of all possible models is the model in which student dependent effects vary from experiment to experiment, whilst student independent factors not only vary from question to question but also vary differently in different experiments. Utilising the notation introduced in this investigation to specify the different facets, the relevant Rasch model is given in Equation 35, where the τ parameter is specific to the survey question asked as well as the relevant category threshold. This is highly problematic, as it results in a model where nothing remains constant between different experiments evaluated.

$$ln\left[\frac{\mathrm{P}(X = x_k)}{\mathrm{P}(X = x_{k-1})}\right] = \beta_E - \delta - \tau_k \qquad \textbf{35}$$

When modelled in this way, the data are split up into 33 isolated subsets, each of which contains the data gathered from one specific student group sampled. Different sample groups correspond either to different years, different student cohorts (Foundations of Chemistry IA/IB or Chemistry IA/IB), or different practical exercises conducted. Measures are not comparable

across different subsets when data sets lack connectivity in this way (see section 2.3.3), meaning that if this genuinely is the way the data are best modelled, data gathered from one experiment cannot easily be validly compared to data gathered from another. Practically speaking, this corresponds to an interpretation of the data where something a student thinks is "good" on one occasion could be the same as what they consider "poor" on another occasion. Even if the identical students are used to evaluate two experiments on two occasions, there is no reason to assume their responses can be considered as meaning the same thing each time; in fact it appears that, based on these results, they most likely will not. This could potentially ruin any ability to use student evaluations as a tool of comparing the qualities of each experiment, conflicting with the entire purpose of the ASLE instrument as a tool of evaluation and comparison.

4.1.4 Investigating comparability between different sample scores

Rasch models of the data may be made comparable using 'equating' techniques, however such techniques are not available to researchers using the usual ASELL scoring procedure. For scores to be comparable between different occasions under this explanatory model, the same overall distribution of student measures needs to be assumed. The hypothesis that the overall student measure distribution remains constant despite variation in individual students was tested here by equating the Rasch measurement scales of the different experiment specific subsets, and comparing the distribution of student measures observed for each.

The different experiment-specific subsets of data were equated using two techniques. First, experiments which were structured identically despite being presented at different times and to different students were assumed to have the identical student independent (δ) measures. This reduced the number of isolated subsets from 76 down to only 29, and resulted in a better explanatory model of the data (log-likelihood χ^2=217851.8, 9734 free parameters, 49.03% of variance explained, ΔAICc=-76.6 compared to the unequated β_E-δ model). Secondly, the remaining measurement subsets were equated by identifying 82 students who appear to have invariant bias (both infit and outfit mean square values <0.8 for the β-δ model of the same data, with responses for at least five experiments) and modelling them to have the same (β_E) measure for every experiment in which they responded. All other students were still modelled to have a measure different for each experiment, as per the best explanation of the ASLE data. This new model proved to be a better explanation of the data than the previous equated model, due to the fact it accounted for the consistency in these students' responses and avoided the need to model multiple bias parameters for each (log-likelihood χ^2=218693.7, 9286 free parameters, 48.68% of variance explained, ΔAICc = -202.6 compared to the previous partially equated model). This validates the selection of students to consider as having invariant βE measures for the purposes of equating.

Following application of these equating techniques, the data set contained only 7 isolated subsets. The largest of these subsets (subset 1) was used to assess the comparability of the student bias (experiment-specific student measure) distributions gathered from each separate sampling occasion. The distributions of measures from each of the separate groups of students are shown in Figure 33, where sample groups are labelled by student cohort (C for Chemistry IA/B, F for Foundations of Chemistry IA/B or u for an unknown or mixed cohort), then labelled by year, then labelled the title of the experiment conducted. A more positive bias measure (β_E) reflects a greater propensity to provide a positive response to all fourteen Likert-type items posed.

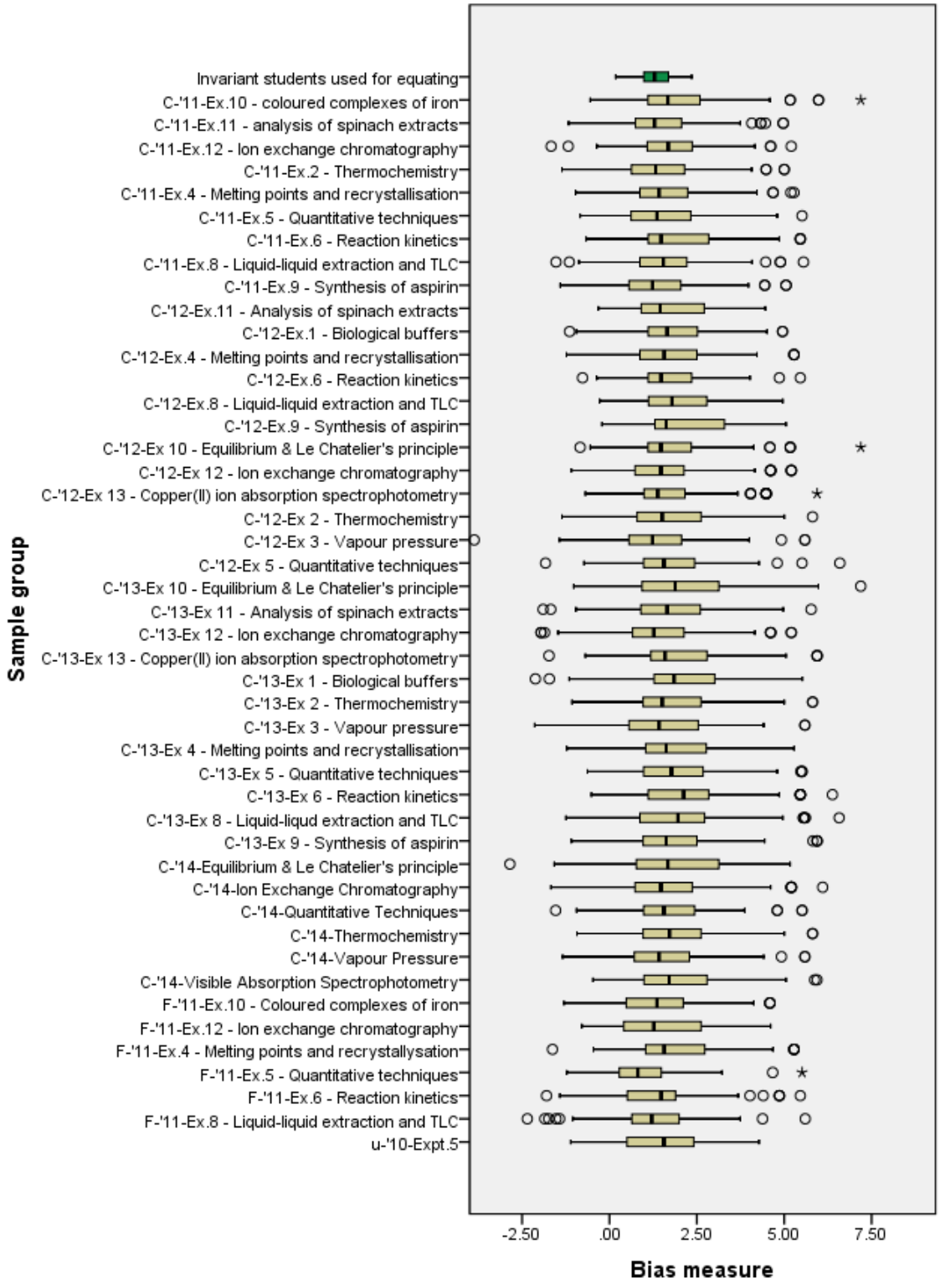


Figure 33: Comparison of the distribution of student biases samples on different occasions

The distributions appear quite variable, and almost all appear to significantly deviate from the normal distribution (see Table S 57 in the supporting information, section 7.4.2). Both the shape and centre of the distributions appear to significantly differ as judged by the Kruskal-Wallis test (test statistic = 164.855, df = 45, $p<0.001$) and Mood's median test (grand median = 1.510, test statistic = 135.545, df: 45, $p<0.001$). The students used for the purposes

of equating appear to have a much smaller range of bias values, reflecting the fact that those students who appear to respond in a consistent manner tend not to do so in an extremely positive or negative way. They were not included in the comparative statistical tests.

4.1.5 Other notable features of the equated model

Final results were split into seven isolated measurement subsets. Subset 1 contained most experiments conducted by the Chemistry IA/B cohort, whilst subset 6 contained most experiments conducted by the Foundations of chemistry cohort. The other subsets include experiments which could not be connected to the rest of the data. These experiments included the original form of the three experiments which made use of handheld data loggers[256] (as opposed to using laptop computers in their revised forms), the initial form of the Foundations of chemistry "introductory experiment" and the original form of the Foundations of chemistry version of the "Reaction kinetics" experiment before revising the order and phrasing of the questions presented. Experiment quality measures for the connected data are presented in the supporting information (Table S 58, section 7.4.2).

The reliability of the experiment quality measures obtained in this equated model is very high (separation = 5.87; reliability = 0.97), whilst student bias measure reliability could be improved, yet is still acceptably high (separation = 2.70; reliability = 0.88). These values are promising, particularly given that misfitting students were not removed from consideration as is often common practice in Rasch measurement. Fit statistics for the question-specific experiment measures reveal that in the majority of cases, questions 11 and 13 (concerning teamwork and time availability respectively) fit the Rasch model poorly. This is true for both inlying data points and for outlying data points (judging by infit and outfit respectively). Question 12, concerning students' opportunity to take responsibility for their own learning, also appears to misfit, but to a lesser degree. This may mean that perceptions for questions 11, 12 and 13 are poorly modelled under this interpretation of the data. The general overfit observed in other survey items may indicate the presence of confounding factors forcing student responses to align more than expected. A description of the variety of statistics described in Table 19 is available in sections 2.5.1 and 2.5.2.1.

Table 19: Fit statistics associated with ASLE survey items in the equated model

#	Survey item topic	Data points	Infit		Outfit		Estim. Discrm	Correlations	
			MnSq	ZStd	MnSq	ZStd		PtMea	PtExp
1	data interpretation	9250	0.93	-3.8	0.94	-3.6	1.06	0.64	0.61
2	laboratory skills	9276	0.95	-3.0	0.92	-4.7	1.08	0.68	0.65
3	interest	9264	0.95	-3.4	0.94	-3.5	1.08	0.68	0.66
4	clear assessment	9263	0.91	-5.7	0.89	-6.4	1.10	0.66	0.62
5	expected learning	9260	0.83	-9.0	0.81	-9.0	1.17	0.69	0.61
6	increased understanding	9258	0.79	-9.0	0.77	-9.0	1.22	0.71	0.62
7	background in introduction	9244	0.95	-2.8	0.94	-3.4	1.06	0.64	0.62
8	demonstrators	9253	1.07	3.9	1.00	0.0	0.96	0.51	0.53
9	procedure in manual	9247	1.03	1.7	1.01	0.6	0.99	0.62	0.63
10	relevance to chemistry studies	9254	0.90	-6.5	0.87	-7.8	1.13	0.67	0.62
11	teamwork	8628	1.53	9.0	1.64	9.0	0.51	0.55	0.68
12	responsibility for own learning	9207	1.11	7.1	1.18	9.0	0.86	0.54	0.60
13	time availability	9196	1.64	9.0	1.77	9.0	0.61	0.38	0.61
14	overall learning experience	9211	0.82	-9.0	0.81	-9.0	1.14	0.68	0.61

4.1.6 Discussion

4.1.6.1 Responses do reflect qualities of the experiment

Broadly speaking, it appears clearly evident from these analyses that student independent factors play a significant role in the generation of ASLE survey question responses. A clear question-specific element to the survey responses is present, reflecting qualities that both appear generally true for most students and also appear to be specific to the design of the experiment being evaluated. This is evidenced by the fact that removing the experiment-specific or question-specific nature of student independent factors (or removing student independent factors entirely) invariably results in a worse explanatory model of the observed data, regardless of how student dependent factors are conceptualised. The ability to simplify the Rasch model by presuming all experiments with the same design have the identical student independent, question-specific measures (as seen during the equating procedure) lends considerable validity to the usual interpretation and practical use of ASLE survey results. Based on these results, the ASLE surveys do indeed appear to target the 'objective' quality of the experiment designs being evaluated, with respect to the questions being asked.

The misfit of survey items 11 and 13, and to an extent item 12, suggests that the manner in which the data were modelled here does not align with the way in which the data appear to behave for these specific items. This suggests survey item 11, pertaining to the benefit of teamwork, item 12, pertaining to the opportunity to take responsibility for one's own learning and survey item 13, concerning time availability, exhibit poor construct validity under this interpretation of the data. A model in which student bias remains constant for different survey items yet differs between occasions, whilst experiment quality measures are specific to each experiment and survey item posed, appears not to explain responses to these specific items well. Factor analysis previously conducted on an entirely separate set of ASLE data[54] has previously revealed that items 11 and 13 appeared independent to other questions of the instrument, potentially suggesting that student dependent factors for these two items may not be equal to those applying to other survey items, as was modelled here. A future refinement to the current best explanation of the ASLE data would therefore be to model student bias parameters differently for these two items specifically, maintaining the current model for the others. The validity of measurements obtained from these survey items remains unknown until these further analyses are conducted. The other survey items, however, appear to be well explained by the current best model determined.

4.1.6.2 ASELL scores obtained from few different cohorts contain inherent error

Despite the fact that ASLE responses do reflect the qualities of the experiment being evaluated, these factors are clouded by the influences of student biases on the responses given. Student predisposition towards positive response appears to vary widely between individuals, and individuals mostly appear not to have the same predisposition from one occasion to the next. This interpretation of student factors appears to be the best way of modelling responses, irrespective of the way student independent factors are conceptualised.

Student bias towards positive response appears best modelled as if it is constant for all questions on a given occasion, but changes from one occasion to the next. This means that even using the exact same student group, results from one experiment may not be comparable to another given that the same student's responses may mean fundamentally different things. These effects, being the major contribution to individual responses, are liable to perturb overall scored data received unless the sample size gathered is sufficient to

'average out' this variability. Achieving this, however, is problematic. There are two sources of student bias variation here: within group variation, which may be minimised by gathering a large number of samples from each single student group, and between group variation, which may only be minimised by gathering a representative sample of many different groups of students. Gathering a vast array of data from a single student group will naturally minimise any student sampling effects caused by variation in student biases within that sampled group, but because variation in the average bias occurs between different student groups (eg. from different years of sampling in the case of this study), error introduced by sampling effects can never be minimised unless a large number of *groups* are sampled, not just a large number of students from each one. This is the reason why results of this study still exhibit errors introduced by student bias variation, despite the fact some ASELL scores have been calculated from over 700 responses. These responses still emerged from a small number of different groups, and so the between group variation has not been minimised.

Because of the presence of between cohort bias variation, scored responses gathered from a single cohort or few cohorts contain an inherent degree of unreliability beyond that imparted by low sample sizes or imprecision in the response scale. The inability to separate student bias effects from measures of experiment quality in scored data serves as a constant limitation to the use and interpretation of the ASLE survey data. Advanced techniques such as Rasch analysis are capable of separating these effects, though such methods are not nearly so simple and available to educators as the usual scoring methods of survey analysis. A practical implication for researchers utilising scoring methods would be to only infer a genuine difference in experiment quality once score differences appear to be large, even if calculated standard errors are small. Statistical tests conducted on scored data cannot differentiate between genuine change in the experiment quality as opposed a change in the average bias of the broad student cohort, and this should be acknowledged in all ASELL survey research. Due to the fact student bias distributions may unavoidably differ between sampling occasions, small variations in scores between two ASLE evaluations should be dismissed as expected variability, *even if statistical tests reveal a significant difference*.

4.1.6.3 Experiment quality correlations cannot be revealed by correlating scores

The fact that student bias values remain constant between different survey items, and the fact they also contribute to the vast majority of variation between responses provided means that correlations between individual response scores for different questions are most likely due to person biases staying constant between questions, not relationships between the actual experiment qualities. Correlations based on individual response scores are therefore mostly unrelated to anything able to be altered by restructuring the experiment and therefore not of any practical use whatsoever in designing appealing laboratory activities. The correlations are most likely revealing factors "beyond our control". In the case of correlating mean scores rather than scored individual responses, this problem would become less prevalent as more responses were used to generate each data point. Increasing the number of students whose scores are averaged to compute the mean allows the student bias effects to be further 'averaged out', meaning the scores more closely reflect the student independent experiment quality measures.

Further research needs to be conducted using sample independent means of measurement, such as Rasch analysis, in order to determine which experiment qualities correlate to the overall broadly appealing nature of the experiment, in a sense true for most students. A wider variety of different experiments needs to be sampled than was the case in this study.

4.1.6.4 Future model refinements and deeper understandings

During the course of this investigation, the best explanatory model of the ASLE survey data studied here was determined. As a consequence of the nature of this model, namely the fact that student bias terms appear to vary between occasions, yet remain constant for all items on each occasion, the generation of Rasch models of the ASLE data including students who have not provided their identification, or who have only completed a small number of experiments has become possible. This is a substantial improvement upon previous models, which presumed the necessity of tracking student identifiers, consequently severely limiting data availability.

Additionally, the model determined here to be the current best explanation can be used as a starting point in subsequent studies to test further refinements. As discussed in the introductory material, it is not currently possible to confidently predict the theoretical expectation of the way perception measurements should change, given a specific change in experiment design. Given a sufficiently wide array of different experimental designs and accompanying ASLE response data, this goal now appears obtainable. Further refinements to the way experiment quality measures are modelled, connecting their value to the design features of the experiment, could be tested similarly to the manner implemented here, using AICc values. Using the current best explanatory model as a starting point, a refinement could be hypothesised, and AICc values could be used to contrast the current best explanation with the newly proposed explanation. The best model of the two could then be taken to be the new best explanatory model, iterating this process continually as progressively more data becomes available and more hypotheses are able to be tested. With the knowledge that experiment quality measures for the identical procedure do remain constant for different student cohorts, this process could feasibly utilise any and all ASLE survey data that has ever been collected, as it is now known that student bias is entirely occasion specific and hence tracking student identification between occasions is unnecessary for bias measure estimation.

Ways to probe the reasons experiment quality measures take the values they do could potentially involve the development of a specification equation;[132] an equation deriving the value of δ as a function of other components, based on some theoretical framework. This could include expressing δ as the sum of a number of facets each related to some aspect of experiment design, or using other similar methods such as using the linear logistic test model[142, 143] or multidimensional Rasch models.[141, 152] Establishing a specification equation using techniques such as these would not only serve to further complete the process of validating the ASLE survey measurements, but more importantly would establish a quantitative, predictive and testable model of student perception outcomes as a direct function of experiment design. Such a model would be invaluable knowledge for any educator implementing, designing or researching laboratory learning exercises.

4.1.7 Conclusion

Throughout the course of this investigation, the ASLE survey responses have been clearly demonstrated to contain a student independent component, specific to both the experiment being evaluated and the survey item posed. This establishes that these surveys can validly be used to compare the quality of different laboratory learning exercises, in a sense that is generally true for most students. It is however, necessary to conduct further investigation to establish this for both the time availability question and the teamwork benefit question of the survey. Student bias effects appear to be inconsistent between different sampling occasions,

meaning that scored ASLE survey data should be assumed to contain an inherent and expected error, unless a representative sample of different student cohorts is used. Gathering many samples from the same student cohort does not alleviate this effect, and differences in experiment quality should only be inferred from scored data if differences are large, even when small differences appear statistically significant. Correlations between scores obtained for different survey items are more likely to reflect similarity in student biases than factors which may be exploited by educators to develop more generally appealing laboratory sessions. This is certain for correlated individual responses, and progressively less of an issue as more responses are used for each data point if mean scores are correlated. The current model serving as the best explanation of the ASLE survey data determined in this research could feasibly be used as a starting point to develop other models, potentially connecting experiment quality measures directly to facets of the experiment design.

4.2 Gender differences in the perception of laboratory learning experiences in chemistry

4.2.1 Outline

Given the determined best explanatory model of the ASLE data, the opportunity arises to investigate finer level trends in both the person measures and the experiment quality measures. Given Rasch analysis' ability to estimate these measures separately, analyses are possible here which would not be achievable if a scoring method were implemented. This section utilises these advanced techniques to contrast the perceptions of male and female students during their chemistry laboratory sessions. This is investigated from two different perspectives: identification of any difference in the average general tendency to provide positive response to any given laboratory exercise, as well as investigation of experiment and question specific differences consistent across all students.

4.2.2 Specific methods

All data utilised for this analysis includes the data used to estimate the previously determined best explanatory model for the ASLE data (see section 4.1). For those students who provided their identification numbers on the survey, the gender of the student was recorded and used for comparative purposes. Responses which were unable to be identified were excluded from the comparative tests, but still contributed to formulation of the Rasch model.

A difference between genders in the tendency to provide positive response as broadly applicable to any experiment in general was tested by taking the average of all person measures estimated for a specific individual, for each individual in turn, then contrasting the distribution of these measures between genders. Because of the connectivity issues of the determined best explanatory model, this procedure was repeated for two distinct subsets of the data: "subset 1" containing experiments conducted by Chemistry IA/B students (or in some cases both by Chemistry IA/B and Foundations of Chemistry IA/B students), and "subset 6" containing experiments conducted by the Foundations of chemistry IA/B cohort. More specific details regarding which experiments are contained in these subsets and how many survey responses contributed to estimating their measures is available in the supporting information relevant to determination of the best explanatory model of the ASLE data (Table S 56 and Table S 58 in section 7.4.2). Student measures specific to subset 1 experiments were disregarded when finding average measures for the subset 6 comparison and vice versa, as measures are not comparable across subsets (see section 2.3.3).

Experiment quality measures were tested for differential item functioning (see section 2.5.3) between the two genders using the *Facets* software, aiming to reveal any experiment or question specific differences between genders. Specific experiment quality measures for which no students who listed their ID number provided response were not compared, meaning that of the 406 experiment quality measures estimated in the equated model (one measure for each of the 14 survey items, for 29 equated experiments), only 350 were able to be tested for significant differential item functioning (DIF). Because this analysis therefore involves 350 distinct hypothesis tests, multiple comparisons are an issue in this study. Under the null hypothesis of no evident DIF in any case, it would be expected that 5% of the 350 tests performed would reject the null hypothesis at $p<0.05$ (by definition) as a simple consequence of natural random variation. Therefore, the proportion of DIF tests resulting in $p<0.05$ was tallied, and was tested using the normal approximation to the standard error of a proportion

(see section 2.4.3) to determine whether the observed proportion of rejecting hypothesis tests was significantly greater than the expected 5%. The Bonferroni correction to the significance level was also considered in order to correct for family-wise error, when interpreting results of isolated hypotheses within the full set.

4.2.3 Results

4.2.3.1 General predisposition toward positive response

Non-parametric statistical tests were required for the comparison of student measures between genders, as the distribution of values obtained appeared to significantly deviate from normality as judged by the Shapiro-Wilk and Kolmogorov-Smirnov tests (Table 20).

Table 20: Normality tests for distribution of students' average measures

	Data set	Kolmogorov-Smirnov test			Shapiro-Wilk test		
		Statistic	df	p	Statistic	df	p
Subset 1	Female students	.056	471	.001	.980	471	<0.001
	Male students	.055	502	.001	.982	502	<0.001
Subset 6	Female students	.047	325	.075	.990	325	.023
	Male students	.110	317	<0.001	.963	317	<0.001

The general predisposition toward positive response for any given experiment was not found to differ between genders for either the subset 1 data or the subset 6 data. As is evident in both Figure 34 and Figure 35, the centre and breadth of distributions of student measures appears invariant between genders. Tests revealed that neither the distribution nor the median differed significantly between genders for either subset, as judged by Mood's median test (grand median = 1.667, $\chi^2 = 3.116$, $df = 1$, p = 0.078 for subset 1; grand median = 1.586, $\chi^2 = 0.399$, $df = 1$, p = 0.528 for subset 6) and the Mann-Whitney U test (standardised test statistic = 1.619, p = 0.105 for subset 1; standardised test statistic = 0.794, p = 0.427 for subset 6) respectively.

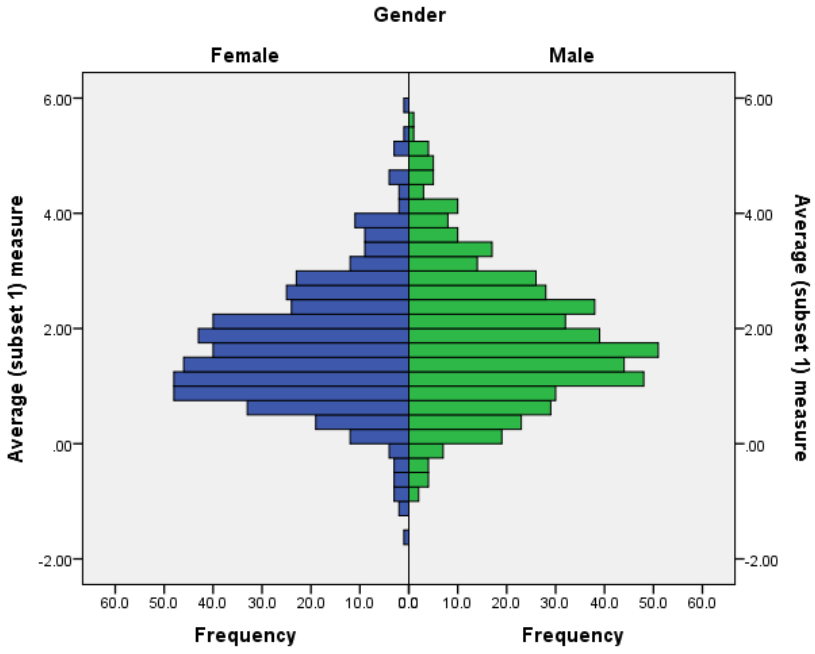


Figure 34: Distribution of student predispositions toward positive response in subset 1

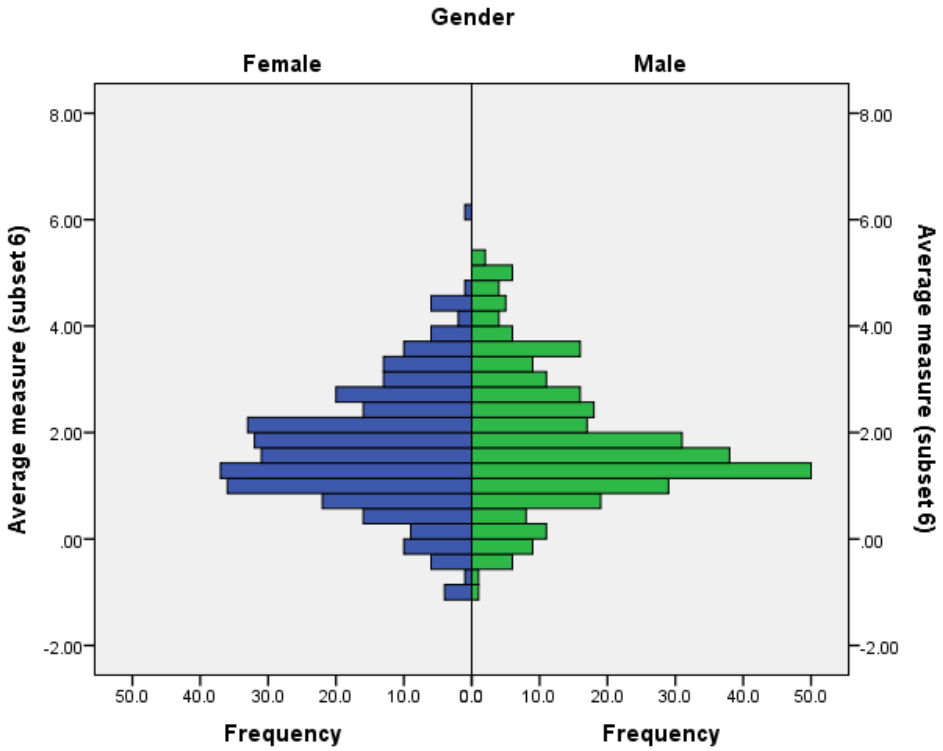


Figure 35: Distribution of student predispositions toward positive response in subset 6

From this data, male and female students appear not to differ in their general tendency to provide positive responses to ASLE survey items for the experiments evaluated.

4.2.3.2 Student independent measures of experiment quality

Following the performance of DIF analysis on the student independent measures of experiment quality, a number of tests reported significant difference between genders at $p < 0.05$. However, no single hypothesis test was so significant as to avoid attribution to family-wise error ($p < 0.05/350$). This means that no individual test of DIF may be used to definitively claim a difference between genders for that specific experiment and survey item in this case. Regardless, the use of a Z test revealed that the proportion of tests suggesting a difference between genders at $p < 0.05$ was significantly higher than the 5% which would be expected under a scenario of total equality. A total of 31 out of the 350 tests conducted (8.86%) reported p values less than 0.05, meaning that the statistical tests conducted indicated gender differences in significantly more cases than would be expected if no true DIF existed in any case ($Z = 2.54$, $p = 0.011$). This result was affirmed by use of an improved approximation to the confidence interval of a proportion, with similar results: the Wilson score interval reports a 95% confidence interval of 6.3% to 12.3% in the observed data, which does not overlap with the expected 5% proportion. The evidence therefore suggests that some degree of difference does exist in the quality of some experiments between genders, however identifying what the specific points of difference are is highly problematic. A full table of all DIF tests conducted is provided in the supporting information (Table S 59, section 7.5).

In order to investigate which specific facets of the laboratory experience genuinely differ between genders, and for which laboratory experiments, the number of DIF tests suggesting inequality between genders at $p < 0.05$ was tallied for each item of the survey, and additionally for each experiment conducted in order to observe where differences are detected most

frequently. Whether the test revealed significantly greater difficulty of providing a positive response for male students or for female students was also recorded.

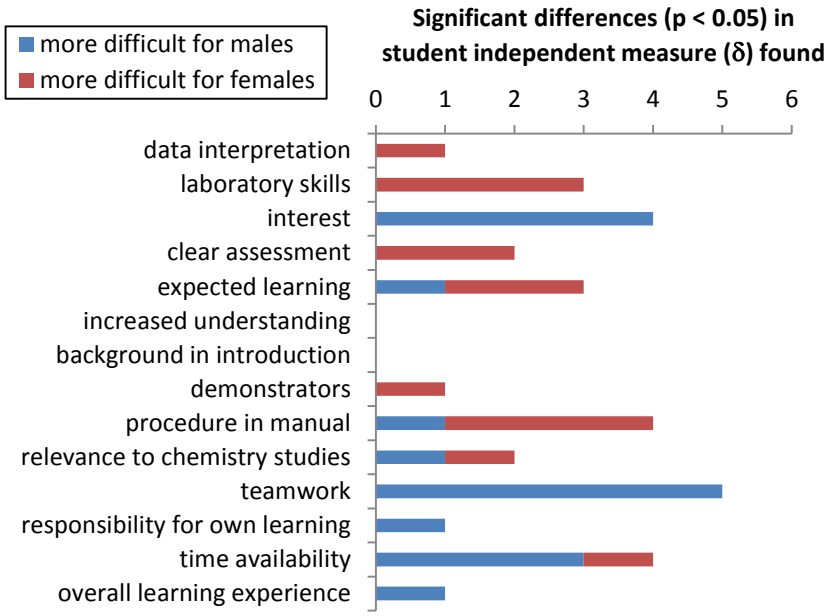


Figure 36: Possible gender differences grouped by survey item

As can be seen in Figure 36, neither the perception of increased understanding nor the perceived sufficiency of the background information provided ever consistently differed between the male and female student groups. All other items exhibit at least one difference significant to $p < 0.05$ detected in one of the 25 equated experiments able to be compared, with some survey items exhibiting differences exclusively in favour of a single gender. Providing a positive response to survey item 11; "working in a team to complete this experiment was beneficial", appears significantly more difficult for males in the case of five different experiments, and evidently equivalent between genders for the other experiments. This result should, however, be taken with extreme caution: it has been shown previously that this survey item significantly misfits this Rasch model of the data (see Table 19, section 4.1.5) and therefore measures may not even be validly attributable to this survey item, let alone any gender differences in those measures. Many other survey items are not subject to this problem, however, and do exhibit multiple occasions of evident DIF. Survey item 3 for example, concerning interest, consistently appears either more difficult to provide positive response for males (in 4 cases) or exhibits no gender difference (the other 21 cases). It should again be emphasised, however, that the issue of multiple comparisons implies that no single test result here was significant to the degree it could not be attributed to family-wise error. The differences at $p < 0.05$ enumerated and displayed in Figure 36 are easily attributable to random chance. These results therefore represent grounds for further investigation more than they reflect a conclusive characterisation of precise differences between genders.

Differences at $p < 0.05$ were similarly enumerated and grouped based on the experiment in which they occur (Figure 37). No specific experiment appears to exhibit possible gender difference in any more than three of the fourteen survey items, and again any differences reported here are attributable to family-wise error regardless.

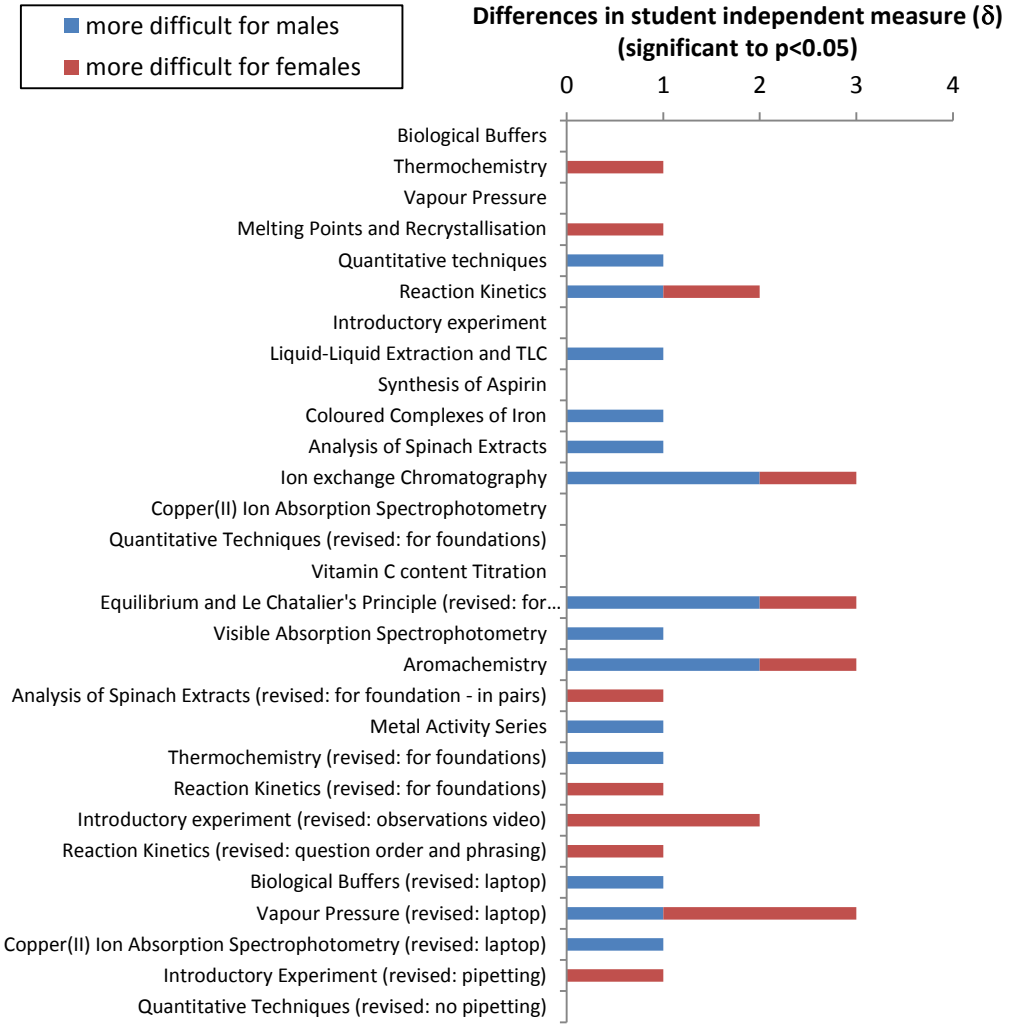


Figure 37: Possible gender differences grouped by experiment

Given that it is known that at least some degree of DIF is likely real as judged by the results of the group level Z-test described earlier, an effort was made to identify any particularly large differences amongst those detected as significant to $p < 0.05$. Here a 'large' difference between genders was deemed at a difference in gender-specific δ measure greater than or equal to $ln(2)$; that is, cases in which the odds of one gender responding in the next highest category is twice the odds of the other gender doing the same.

Large gender differences where males were substantially less likely to provide positive response were detected twice in the case of item 13 and once in the case of item 11, both of which fit poorly to the Rasch model and may therefore not be meaningful comparisons due to construct invalidity (see Table 19, section 4.1.5). Four large differences were found where female students were substantially less likely to provide positive response; one case for item 13, again possibly not a meaningful comparison, and the other three all concerning different iterations of the "Reaction Kinetics" experiment. These differences, however, were not in the same survey item in every case. Figure 38 shows the observed difference between estimated gender-specific experiment quality measures (δ), with error bars representing standard error values.

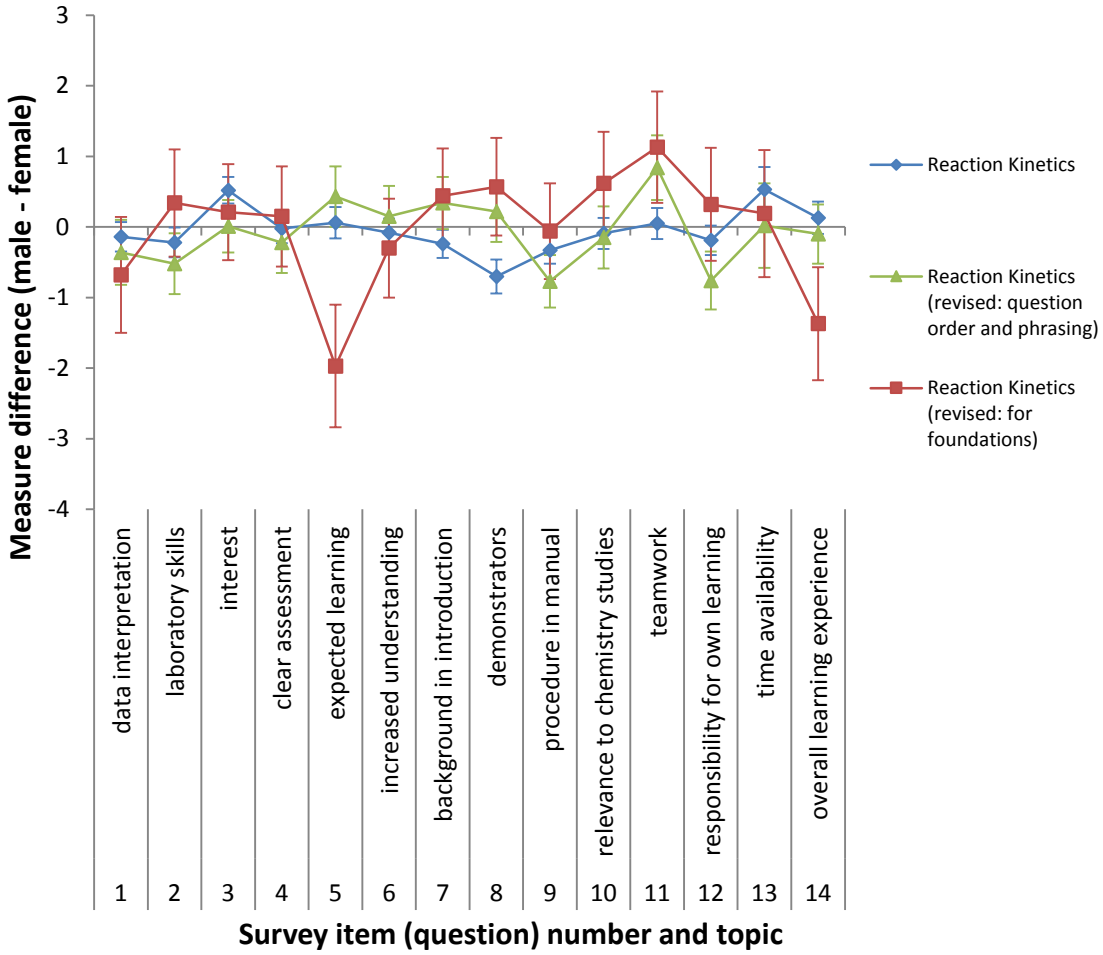


Figure 38: Gender specific DIF analysis for different forms of the "Reaction Kinetics" experiment

Whilst cases of DIF in the "Reaction Kinetics" experiment occasionally appear large and significant, there appears to be little consistency in these occurrences across occasions which could not be equated due to subtle differences in the experiment structure. The largest difference appears to be that once the experiment was revised for the Foundations of Chemistry cohort, the odds of male students responding in the next highest category became seven times that of the female students when posed with item 5: "It was clear to me what I was expected to learn from completing this experiment" (odds *exp(1.97)* times as high, as determined by a difference of 1.97 logits in the gender-specific δ values). This test was, however, one of the few which were based on a small sample of responses, and so has a wider margin of error than in other cases. Consequently this difference only appears significant at $p = 0.0415$.

This experiment was not the only one to be revised to suit the Foundations of Chemistry IA/B cohort, however. Numerous experiments, all of which appear in subset 6 of the equated data set, were revised in similar ways. These revisions included aligning the timing of experiments with the lecture material as much as possible rather than having different student groups conducting different experiments at different times in the semester, as well as slightly modifying the laboratory manual and question booklet accompanying each experiment. Modifications include some small procedural simplifications or amendments (such as working in pairs rather than individually) and rephrasing or amendment to the background information

provided, often in the form of “chemistry connections”: small paragraphs giving extra assistance or connecting theory presented to other theory previously encountered. Hypothesising that these changes made to the “Reaction Kinetics” experiment were the cause of the evident DIF of survey item 5, it would be expected that similar changes to the experiment and manual in other cases would result in a similar observation of DIF. However, this expectation is not evident in the data.

As can be seen in Figure 39, gender difference in survey item 5 is not consistently observed in all cases where the experiment was revised for the Foundations of Chemistry cohort, meaning these revisions are unlikely to be the cause of the large difference observed in the revised form of the “Reaction Kinetics” experiment discussed previously. The fact that the large DIF observed is absent without these revisions, however, means the DIF of item 5 in the revised “Reaction kinetics” experiment is not attributable to any difference in the structure of the experiment itself, and is likely an artefact of random error. This illustrates the difficulty of making multiple comparisons simultaneously: it is expected that some large significant differences will be reported falsely (type 1 error) as an artefact of random error, and more comparisons inevitably means more chances for these errors to occur. Regardless, the number of null hypotheses rejected here remains significantly higher than the expectation presuming no DIF, as previously discussed. It simply remains problematic to identify which DIF is genuine and which is a result of random error.

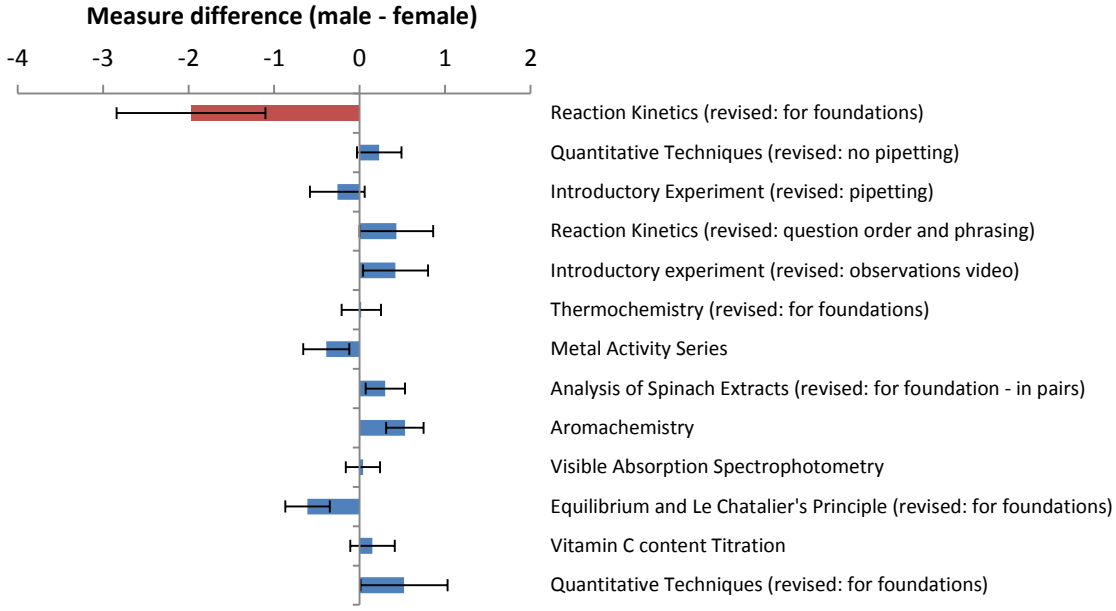


Figure 39: Gender DIF observed in item 5 for Foundations of Chemistry revised experiments

4.2.4 Discussion

Whilst this investigation was unable to pinpoint any specific cases of gender differences in the perception of the laboratory exercises evaluated, it would be incorrect to claim no significant gender difference was detected. The fact that more significant differences were detected than expected under the null hypothesis of perfect equality suggests at least some small cases of genuine difference are likely to exist in this data set. The issue is that the problem of multiple comparisons necessitates very strict criteria for identifying a confident conclusion of genuine difference, with many false positives at usual significance criteria expected for so many hypothesis tests conducted on the same data set. Vast sample sizes are needed to draw

conclusions of difference not attributable to family wise error in this way. Regardless of this difficulty, it can still be concluded that some differences between genders exist in the student independent measures of experiment quality, though those differences which are present appear not to be particularly prominent. It could feasibly be the case that multiple small yet ultimately inconsequential differences exist, summating to produce what is a detectable effect in this study when observing the data overall. Given the small proportion of cases which showed difference even at $p<0.05$, it appears that presuming genders to behave differently as a general rule would be a poor explanation of the data. Rather, what differences exist are likely small and/or infrequent. In terms of more general, broad scale predisposition towards laboratory experiences, rather than specific cases, there also appears to be no detectable difference between genders.

4.2.5 Conclusion

Whilst this analysis was unable to pinpoint any specific, conclusive differences between male and female students' perception of the laboratory experience, it has been shown that some small differences are likely to exist. Any differences present appear not to be due to a general tendency of one gender to be less positive toward chemistry laboratory sessions than the other. Instead, any differences present appear to occur in a manner specific to the experiment conducted and survey item posed. The vast majority of cases appear not to be detectably different between the two genders.

4.3 Empirical estimation of a Linear Logistic Test Model Q-matrix

4.3.1 Outline

Previous research presented in this thesis has yielded the best general explanatory model for the ASLE data, based on observed response patterns. Whilst this model reveals several notable features of the way students interact with experiments to give rise to ASLE survey responses, the model is not informative as to why some experiments are associated with more preferable measures whilst others are not. As such, a deeper understanding of which features of experiment design elicit positive responses remains to be determined.

The Linear Logistic Test Model (LLTM, see section 2.2.2) allows student independent measures to be expressed as a linear combination of more basic components. That is, the model "explains" the observed item parameters as the sum of several underlying factors. In the context of the ASLE surveys, a LLTM formulation may explain the fourteen question-specific quality measures associated with an experiment to be directly resultant of a small number of elementary features of the laboratory experience. Most notably, it may express the measure for survey item 14: the "overall learning experience", as a direct function of other facets of the laboratory experience.

$$\begin{matrix} \\ q1\text{: } interest \\ q2\text{: } clarity \\ q3\text{: } overall \end{matrix} \begin{matrix} Ex1 & Ex2 & Ex3 \end{matrix} \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \delta_{1,3} \\ \delta_{2,1} & \delta_{2,2} & \delta_{2,3} \\ \delta_{3,1} & \delta_{3,2} & \delta_{3,3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \times \begin{matrix} Ex1 & Ex2 & Ex3 \end{matrix} \begin{bmatrix} \eta_{1,1} & \eta_{1,2} & \eta_{1,3} \\ \eta_{2,1} & \eta_{2,2} & \eta_{2,3} \end{bmatrix} \begin{matrix} \\ interest \\ clarity \end{matrix} \quad \textbf{36}$$

Equation 36 above is an example of a simple Linear Logistic Test Model. Three survey questions (q1, q2 and q3) have been asked of three different experiments (Ex1, Ex2 and Ex3) and student independent measures for each have been obtained (δ). The example above explains the nine observed δ measures as linear combinations of only six underlying η measures. For example, the δ measure for question 3 ("overall") is always twice the η measure associated with "interest" plus the η measure associated with "clarity". As a result, only six parameters (η) would need to be estimated to explain the observed data rather than the original nine (δ). In this way the model is made more parsimonious, whilst providing an explanation for all δ measures in terms of more fundamental factors.

The matrix of weighting coefficients in the above example is known as a "Q-matrix". Typically, the Q-matrix is stipulated *a priori* by the analyst, since the LLTM is often used when an explanatory model is already established. Unfortunately, an underlying model of the ASLE survey items similar to the above example is currently unknown. Attempts have been made to evaluate the extent to which other survey items are relevant to the "overall learning experience" rating associated with question 14, but these past attempts have been based on integer scoring methodologies. Addressing this question from a Rasch modelling perspective requires formulation of a Linear Logistic Test Model (or similar), either based on theory or based on data. Given little theory exists concerning student perceptions and their interrelationships, *a priori* stipulation of a Q-matrix for the ASLE surveys would require a substantial degree of 'trial and error' before a successful matrix was found. As such, a method of estimating a satisfactory Q-matrix directly from observed data is needed. Such a method is described in this section. A method for deriving a satisfactory Q-matrix for the ASLE survey data is presented, then applied to the existing data set of survey responses. Adequacy of the Q-matrix estimated as an improved explanation of the observed data is also demonstrated.

Features of the model obtained and their pedagogical implications are to be discussed separately in a subsequent section of this thesis.

4.3.2 Specific methods

4.3.2.1 Estimation of a Q-matrix using factor analysis

Factor analysis (see section 2.4.6) expresses Z-standardisations (see section 2.4.3) of the observed variables as a linear combination of F latent factors. In the case of the ASLE surveys, an example of this would be the expression of the 14 different question-specific quality measures for each experiment as estimated using a partial credit model (δ_{PCM}), as being linear combinations of underlying, basic experiment qualities (factors).

$$z_{i,m} = \frac{\delta_{PCM\,i,m} - \overline{\delta_{PCM\,i}}}{\sigma_{PCM\,i}} \cong \sum_{f=1}^{F} l_{i,f}\varepsilon_{f,m} \qquad \textbf{37}$$

In Equation 37 above, subscripts i and m indicate specificity to the ith survey item and mth experiment respectively. The $\overline{\delta_{PCM\,i}}$ term denotes the mean value of the set of all $\delta_{PCM\,i,m}$ values for the ith survey item, whilst $\sigma_{PCM\,i}$ represents the standard deviation in the set of $\delta_{PCM\,i,m}$ values for the ith survey item. There are F many factors (indexed by f) underpinning the responses to the 14 survey items for any given experiment evaluated. The factor loading of the fth factor onto the ith survey item is given by $l_{i,f}$, with $\varepsilon_{f,m}$ acting as the measure of the fth factor for the mth experiment.

This equation may be rearranged to express the $\delta_{PCM\,i,m}$ value in terms of the factor model loadings and measures:

$$\delta_{PCM\,i,m} \cong \sigma_{PCM\,i} \sum_{f=1}^{F} l_{i,f}\varepsilon_{f,m} + \overline{\delta_{PCM\,i}} \qquad \textbf{38}$$

Which in turn may be rewritten to incorporate the $\overline{\delta_{PCM\,i}}$ and $\sigma_{PCM\,i}$ values into the summation term as follows:

$$\delta_{PCM\,i,m} \cong \sum_{j=1}^{F+14} l'_{i,j}\varepsilon_{j,m} \quad ; \quad l'_{i,j} = \begin{cases} \sigma_{PCM\,i} l_{i,f}, & j \le F \\ 1, & j = F+i \\ 0, & otherwise \end{cases}, \forall m,\ \varepsilon_{(F+i),m} = \overline{\delta_{PCM\,i}} \qquad \textbf{39}$$

This equation now resembles a linear logistic test model (LLTM). The l' values are analogous to the weighting values in the LLTM Q-matrix, whilst the ε values are analogous to measures of the basic underlying variables, of which the observed variables are a linear combination. Subtle differences do, however, exist between this formulation and an estimated LLTM.

A key issue is that the $\delta_{PCM\,i,m}$ values above are estimated such that they sum to zero, but there is no guarantee this will eventuate if the data are modelled using an LLTM structure. Rather, it is the basic parameters of the LLTM, analogous to the ε values in Equation 39, which are instead defined to sum to zero. This has the implication that any estimated LLTM parameters analogous to the $\varepsilon_{(F+i),m}$ term will not necessarily be equivalent to the $\overline{\delta_{PCM\,i}}$ values (as Equation 39 would otherwise appear to suggest). Instead, the estimated values serve to define the location of one item's set of experiment specific measures relative to those of another item, after accounting for the different linear combinations of the underlying

factors. Their sum would equate to the negative sum of the set of all other factor measures estimated. For these reasons, a LLTM analogous to this formulation is therefore best expressed using different variables, as shown in Equation 40.

$$\delta_{LLTM\,i,m} = \sum_{j=1}^{F+14} q_{i,j}\eta_{j,m} \quad ; \quad q_{i,j} = \begin{cases} k\sigma_{PCM\,i} l_{i,f}, & j \leq F \\ 1, & j = F+i \\ 0, & otherwise \end{cases} \quad ,\forall m,\ \eta_{(F+i),m} = \mu_i \qquad \textbf{40}$$

The above model may more simply be stated in matrix form, to illustrate the fact it is directly analogous to the LLTM example presented in the introductory material (Equation 36). In Equation 41 below, Q is the matrix of $q_{i,j}$ values (which serve as the LLTM weighting factors) and H is the matrix of $\eta_{j,m}$ values. Note that values within matrix Q may be calculated from the results of a factor analysis, whilst values within matrix H require estimation using Rasch modelling software.

$$\left[\delta_{LLTM\,i,m}\right] = \boldsymbol{Q} \times \boldsymbol{H} \qquad \textbf{41}$$

In Equation 40 above, a scalar value k is also introduced in order to allow the matrix of weighting factors to be scaled up or down such that all Q-matrix values ($q_{i,j}$) approximate integers. This is necessary for running the LLTM within the Facets software, which is only capable of using integer values for the $q_{i,j}$ weightings, but unnecessary within other more capable Rasch measurement programs. The value k may be selected arbitrarily such that all q values satisfy this constraint approximately, and q values may then be rounded to the nearest whole number to generate a matrix Q' used in the analysis in the Facets software. Larger values of k would allow Q' to more closely approximate the matrix of unrounded values Q, however this would also be more taxing on the Facets software.

Constructing the LLTM Q-matrix in this way, it can be seen that the $\eta_{j,m}$ values serve the same purpose as the factor measures $\varepsilon_{f,m}$, though are not equivalent in value to them. The value of $\eta_{j,m}$ is a measure of the jth basic factor for the mth experiment, and can be estimated based on observed data. The value $\sigma_{PCM\,i}$ is the standard deviation in the $\delta_{PCM\,i,m}$ measures for the ith survey item, estimated from the previous, non-LLTM Rasch model. Working this value into the $q_{i,j}$ parameters allows one survey item to have more variable measures than another survey item without the need of working this item specific variation into the $\eta_{j,m}$ measures, which are desired not to be survey item specific. For easier interpretation of the estimated measures assigned to the basic underlying factors ($\eta_{j,m}$), it is also convenient to define the LLTM measures to have the opposite orientation to the PCM measures thusly:

$$\delta_{LLTM\,i,m} \cong -\delta_{PCM\,i,m} + \gamma_s \qquad \textbf{42}$$

The $\delta_{LLTM\,i,m}$ measures are redefined in this way such that a more positive measure implies increased likelihood of more positive response. That is, the Rasch model is reformulated such that:

$$\varphi_{n,i,m} = \beta'_{E\,n,m} + \delta_{LLTM\,i,m} \qquad \textbf{43}$$

where φ is the latent trait measure input into Equation 1, which is modelled as giving rise to the observed responses for the nth student, ith survey item and mth experiment. The β'_E term here is the experiment specific student bias facet discussed in previous investigations (see section 4.1.2.2, Table 17), shifted in value because of the differences between δ_{LLTM} and δ_{PCM}.

The equating procedures previously detailed for both the student dependent and student independent measures are also maintained: some students are defined to have the same β'_E measure regardless of the experiment conducted, whilst the δ_{LLTM} measures are defined to be equivalent for the identical experiment (see section 4.1.4). The value γ in Equation 42 translates measures by an amount specific to the measurement subset (indexed by s), owing to differences between the partial credit model and LLTM formulations described previously.

4.3.2.2 Resolving disconnected subset issues

The ASLE survey data, as analysed thus far, contains a number of subset disconnects. This implies that the absolute location of measures estimated from one subset relative to the absolute location of measures in another is unknown. This is problematic, as the factor analysis necessary for the procedure above requires all measures for each given ASLE survey question to be correlated against all measures for each other ASLE survey question. This is impossible unless the absolute location of each measure relative to the others associated with the same survey question is known. As such, all subsets of the data must be equated prior to application of this LLTM estimation technique.

One way to achieve this equating of subsets is to artificially force δ_{PCM} measures in one subset to be equal to δ_{PCM} measures in another subset. As such, analysis was carried out to identify cases where all δ_{PCM} measures associated with a specific experiment in *subset one* of the current model were likely equivalent to all δ_{PCM} measures associated with another experiment in *subset six* of the current model (see section 4.1.5). This was achieved though correlating the fourteen estimated δ_{PCM} measures for each experiment with those for every other.

Under the assumption that two experiments have identical measures, two key expectations exist. If the set of measures associated with each of the two experiments were estimated separately, and a linear relationship were drawn between the two, then:

(1) A strong correlation would be observed between the two sets of measure estimates
(2) The slope of the line would be approximately one

Prediction (1) is trivially the case, given that the separate sets of measures, if truly equivalent, would observably yield a set of estimates in the same order and the same relative difference from one another. Prediction (2) is justified by the fact that if the two sets of measures were equal, the magnitude of the differences between any given pair of measures within each set should be the same as the differences observed in the analogous measures of the other set. For example, if measures for items 1 and 2 differ by 0.5 logits in one experiment, the measures for items 1 and 2 should also differ by 0.5 logits in the other experiment, if the two experiments are equal. Experiments with sufficiently similar measures for equating purposes were identified in this way and stipulated to have equal δ_{PCM} measures prior to factor analysis and subsequent LLTM formulation.

4.3.2.3 Features of the factor analysis

Following the forced equating of two experiments described above, the remaining disconnected data subsets were removed from consideration. The δ measures from the remaining 23 connected experiments were then used for the purposes of conducting factor analysis. Image factoring was chosen as the factor extraction method for two reasons. Firstly, image factoring operates via linear regression techniques, which are appropriate for deriving linear models as desired. Secondly, only image factoring was capable of yielding sensible

results for larger numbers of underlying factors. Other factor extraction techniques resulted in so called "Haywood cases" in multiple instances (data not reported), whereas image factoring is not susceptible to this issue (see section 2.4.6). Varimax rotation was used to obtain more easily interpretable factors, maintaining their orthogonal nature.

The method of deriving a Q-matrix described above relies on a factor model having been performed on the δ_{PCM} measures. It is at this stage of the analysis that the number of basic underlying factors explaining the observed responses is defined. A different number of factors stipulated at the outset of the analysis would result in a different Q-matrix estimated, and therefore a procedure is needed to select the most appropriate number of factors modelled.

A range of techniques are typically applied in factor analysis to select the appropriate number of factors. In this study, the appropriate number of factors was selected based on the adequacy of the LLTM model generated from that factor model. A new Q-matrix was generated for each possible number of factors stipulated, then the most appropriate Q-matrix was identified as the Q-matrix which yielded a LLTM with the minimum corrected Akaike Information Criterion value (see section 2.5.4.2).

4.3.3 Results

4.3.3.1 Model estimation

A small number of disconnected experiments with very strong correlations between δ measures were observed (see supporting information: section 7.6.1). The strongest correlation, between "Coloured complexes of iron" and "Equilibrium and LeChatelier's principle (revised: for foundations)" (r=0.975), may be somewhat expected, as these two experiments are equivalent except for minor changes to the instruction manual. The verification that their measures align is useful, however, in that it may imply the two forms of the experiment can reasonably be assumed equivalent, thereby allowing equating between the Chemistry IA/B and Foundations of Chemistry IA/B cohorts, at present divided into distinct measurement subsets in the current best model. Considering these experiments equivalent would not only be a more parsimonious model, but would also therefore allow direct comparisons between experiments which could otherwise not be contrasted. However, treating these two experiments as equivalent in the Rasch model resulted in a poorer explanation of the observed data (ΔAICc = +3.78) due to the proportion of observed data explained by the model lost in making this simplification (χ^2 = 36.2969, df = 14, p = 0.0009). Regardless, these two experiments were the most similar of any pair contrasted, and therefore equated for the purposes of the following factor analysis. Data which remained disconnected from the bulk of the data after this equating procedure were removed from consideration, and a partial credit Rasch model (formulated in the same manner as the current best explanatory model, section 4.1.3) was estimated using the remaining equated data set of 120 701 individual data points, gathered from 23 experiments.

The 322 δ_{PCM} measures obtained from the initial partial credit model were organised by experiment (row) and survey item (column) in preparation to conduct the factor analyses. The KMO measure of sampling adequacy and Bartlett's test of sphericity were used to confirm the data were adequate for factor analysis, revealing the sample size was marginally adequate at best (KMO = 0.511) but contained a significant degree of correlation (χ^2 = 195.186, d.f. = 91, p < 0.001). As will be seen, the poor sample adequacy appears not to compromise the final results obtained. Factor models were generated for 9, 10, 11, 12 and 13 factors underpinning

the fourteen observed survey item measures for each experiment. Modelling lower numbers of factors became unnecessary based on results (discussed in conjunction with Figure 40 below).

Following the computation of Q-matrices from factor models as described in the specific methods section, one for every different number of latent factors modelled, the *Facets* software was used to estimate a corresponding LLTM for each (see section 7.6.1 for the structure of the specification files). The corrected Akaike Information Criterion (see section 2.5.4.2) was used to identify which of the different LLTM formulations generated provided the best explanation of the observed data.

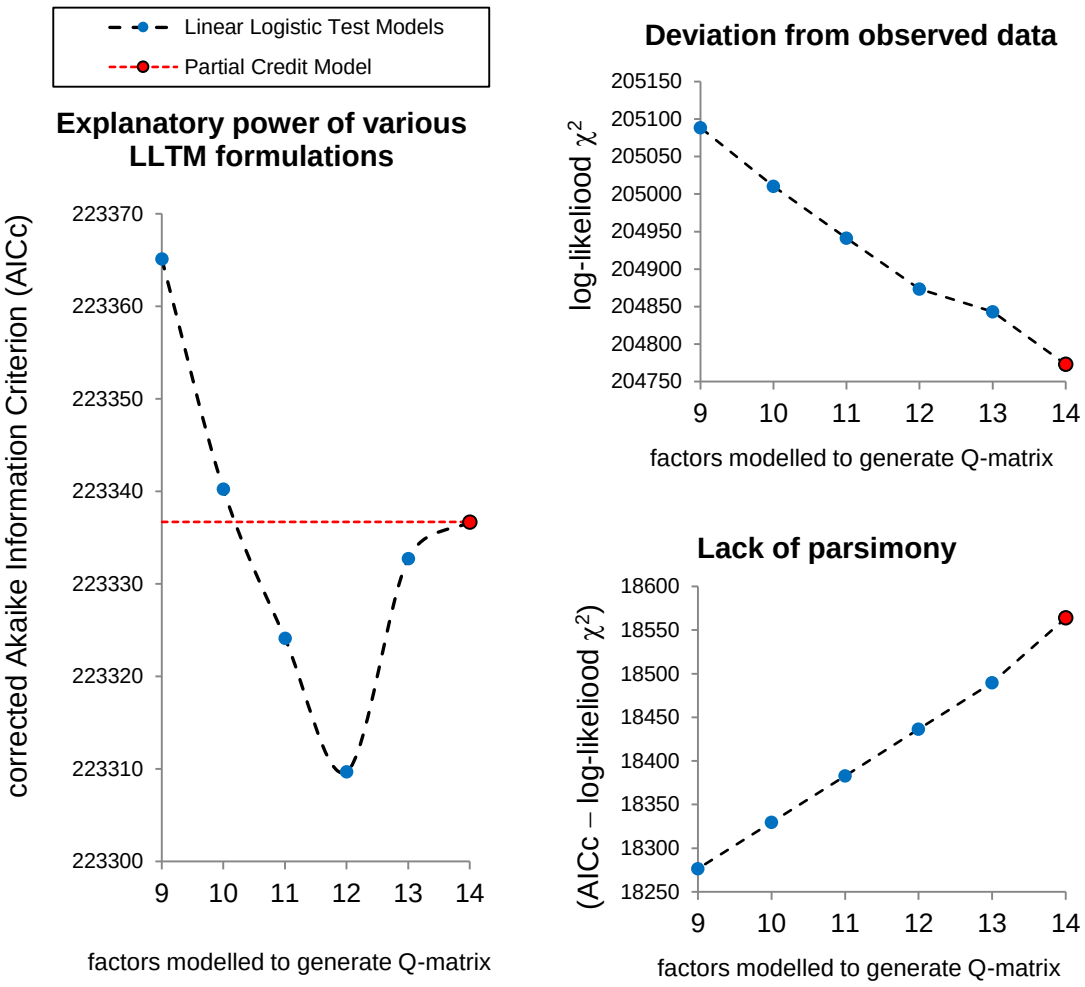


Figure 40: Efficacy of Linear Logistic Test Models formulated using factor analysis results

Modelling a greater number of factors underpinning ASLE survey responses is less parsimonious (lower right) but explains a greater proportion of observed data variance (upper right). An optimum balance between these two competing considerations is found when 12 underlying factors are modelled, corresponding to a minimum on the "surface" of AICc values (left). This model explains the observed data better than all other models, including the previous best explanatory model of the ASLE data discussed in section 4.1.3. As can be seen, modelling fewer than 9 factors would likely produce progressively less desirable models, extrapolating from the data presented.

Figure 40 (left) displays the AICc values corresponding to different numbers of latent factors modelled as explaining the survey item specific measures for each experiment. The plots

displayed to the right of the figure break down these AICc values into their deviation from observed data and lack of parsimony components. Shown in red are the values obtained for the unsimplified (partial credit) model: the model up until this point deemed to be the best explanation of the observed data (see section 4.1.3). The best explanatory model of the data was observed to be a LLTM with 12 latent factors explaining response to the 14 ASLE survey items for each experiment. Following the identification of the 12-factor model as optimal, the Q-matrix for the 12 factor model was used to estimate a LLTM for the full data set (128,881 data points) without artificially equating "Coloured complexes of iron" and "Equilibrium and Le Chatelier's principle (revised: for foundations)", as these were only equated previously to enable the factor analysis to be performed. This LLTM proved to be the best proposed explanatory model of the full data set thus far (log-likelihood χ^2=218807.7, 9213 free parameters, 48.62% of variance explained, ΔAICc = -55.5 compared to the previous best model).

The parsimony of the LLTM appears not to have resulted in a loss of explained data variance. The raw estimated δ_{LLTM} values approximate non-LLTM estimates for the same data closely, as seen in Figure 41. As can be seen, there is a very high level of agreement between the LLTM measures and the analogous non-LLTM measures. The more parsimonious LLTM accounts for 99.56% of the variance in the non-LLTM measures. The "displacement" value in Figure 41 is a simple error term reflecting the difference between the LLTM model prediction and the "optimal" value.

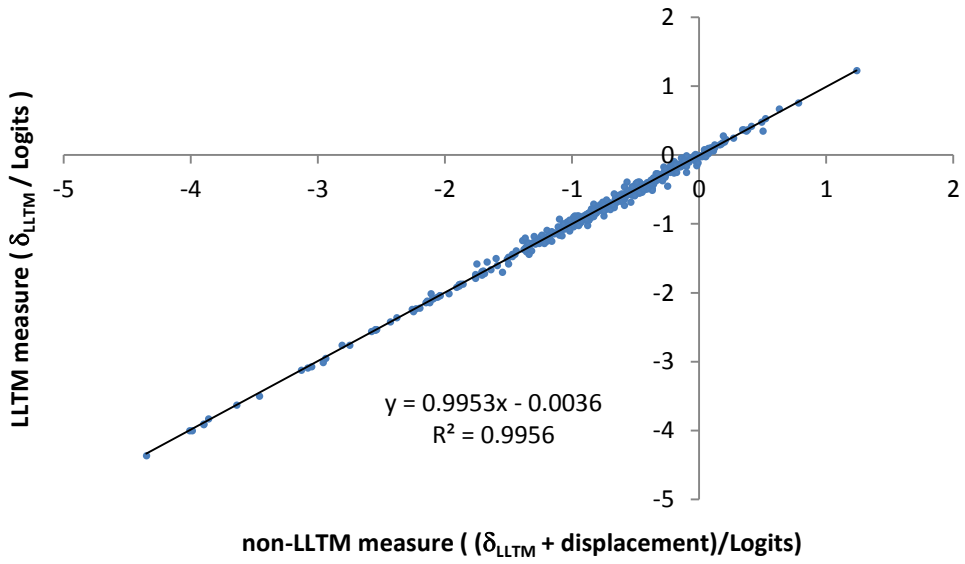


Figure 41: Accuracy of Linear Logistic Test Model approximatons

The LLTM's linear combinations of the estimated factor measures estimated do not necessarily sum to values which would explain the data optimally. The difference between the LLTM estimate (δ_{LLTM}) and this optimal value is expressed as the "displacement". Thus, adding the displacement value back to the LLTM model's estimates yields the optimal measure which would otherwise be reported in a non-LLTM model. The strong correlation between δ_{LLTM} and δ_{LLTM} + displacement reflects the fact that very little variance in the observed measure estimates is lost in the LLTM approximation.

Performing a likelihood ratio test (see section 2.5.4.1) reveals that some degree of explained variance is lost when this LLTM is applied to the full data set ($\chi^2(73) = 113.98$, $p = 0.002$), but when applying this LLTM to the somewhat restricted data set used to generate the factor

analysis results, no significant loss of explained data is observed ($\chi^2(55) = 27.01$, $p = 0.999$). The LLTM obtained appears to be a definitively superior model to the partial credit model, when applied to the exact data set used for estimation. The model also appears somewhat generalizable beyond the estimation data set, given the AICc value still affirms it as the superior model in the wider data set. An even better LLTM may therefore have been possible for the broader data set, were it better connected and able to be used for Q-matrix estimation.

4.3.3.2 Data connectivity and errors in typical scored data

A useful feature of the LLTM applied to the full data set is that the data now appear fully connected, where disconnects otherwise existed. Experiments which were previously in isolated subsets of the data and unable to be contrasted now appear in the same subset of measurement, because the δ_{LLTM} values are necessarily composed of the same underlying factor dimensions. As can be seen in Figure 42, each previously isolated subset within the original set of δ_{PCM} measures appears offset from the other subsets, rendering comparison between different subsets of data invalid. Connectivity is achieved with the LLTM, however, and the artificial offsets in measure values reported due to subset disconnects (γ_s, see Equation 42) are now known.

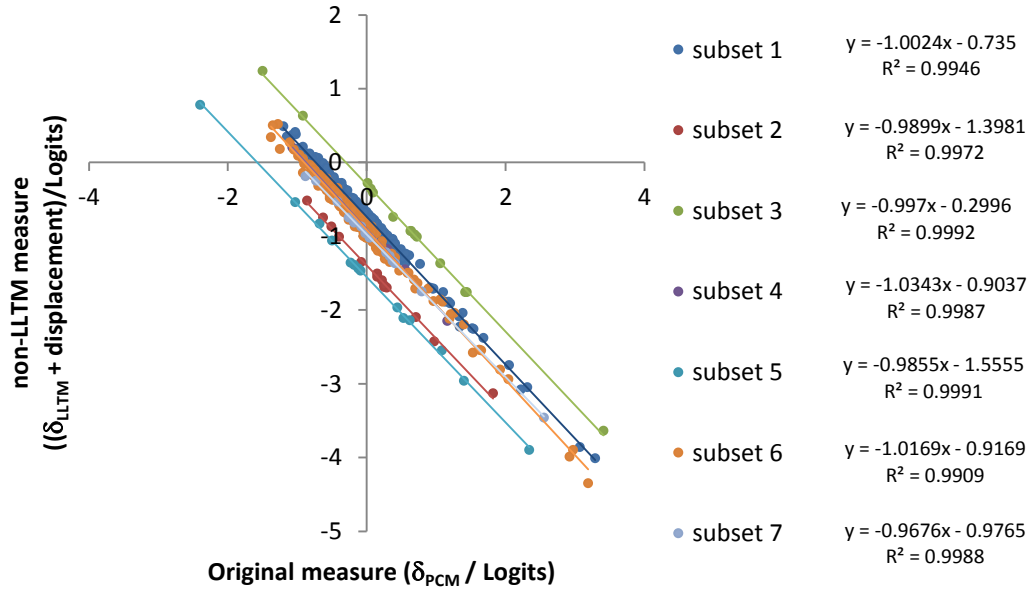


Figure 42: Previous measure offsets in disconnected subsets of data revealed by the LLTM

Linear relationships observed are the linear relationship between δ_{LLTM} and δ_{PCM} described previously by Equation 42, presented in the introductory material for this section. The original δ_{PCM} estimates appear in seven disconnected subsets of the full data set, each visualised as a separate trendline. Vertical translation of the different trendlines results from a lack of ability to assign δ_{PCM} measure values relative to those in other subsets previously. The y = mx + c equations shown directly emulate Equation 42, where the slope value m is always approximately -1 and the intercept value c is the subset specific offset γ_s. The δ_{LLTM} term featuring in Equation 42 is replaced here by δ_{LLTM} + displacement, where the displacement is simply an "error" term resulting from imperfect approximation of the LLTM summations to their optimal (non-LLTM) values. These error corrected values were used here to better estimate the subset offset parameters. Lack of perfect correlation observed within each subset likely results from subtly different Rasch model optimisation when the Facets software specification file is structured to accommodate an underlying LLTM as opposed to a simple Partial credit Model (PCM).

The comparability of measures afforded through this connectivity allows for broad scale statistics to be drawn, quantifying the average error introduced in typical ASELL integer score values. Mean scores calculated from observed responses can be contrasted with the "fair" scores which would be expected in absence of any sampling errors (computed from LLTM predicted category frequencies). As can be seen in Figure 43, the bulk of the data closely fit to a singular linear relationship, with a high correlation. However, data from the original (data logger) variant of "Vapour Pressure" deviate from this trend, exhibiting far lower calculated scores than would be expected without sampling bias.

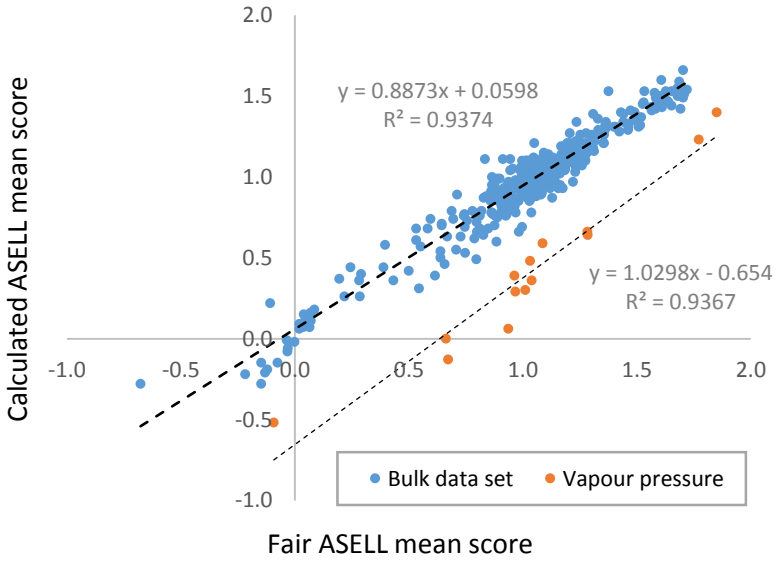


Figure 43: Effects of student biases on calculated ASELL mean scores

Scores here refer to values calculated using the typical technique of assigning successive integer values to the successive rating scale categories, associated with the final 29 equated experiments of the full data set (see Table S 56 in section 7.4.2). Statistics shown associated with the linear relationship drawn are relevant to the bulk of the data only (excluding "Vapour Pressure"). Mean scores for the "Vapour pressure" experiment are substantially offset, indicating unfair evaluation compared to the other experiments.

It is known that this offset for the "Vapour pressure" experiment values is not due to data connectivity issues. Because all basic parameters in the LLTM were included in a single facet (see section 7.6.2 in the supporting information), η parameters for each LLTM factor associated with "Vapour pressure" have known position relative to the item locations μ, which in turn have known location relative to the η measures for all other experiments. This observation therefore suggests that students evaluating the "Vapour pressure" experiment had a broad scale negative bias against the experiment as a whole, to a degree which never occurred for other experiments evaluated. Given this experiment was received poorly, it therefore appears that this variant of the "Vapour pressure" experiment was received so poorly that students began to judge it unfairly. The trendlines shown in Figure 43 illustrate that for the data logger variation of "Vapour Pressure", students provided responses approximately 0.7 score units lower than would be fair, for all items of the survey. The fact that such an effect can exist is important both for the interpretation of survey data and for the design of laboratory activities.

Discounting “unfair” evaluations like this, the rest of the data can be used to quantify expected sampling errors present in calculated score values. Based on the central limit theorem, it may seem intuitive that increasing the number of observations used to calculate a mean score would reduce the error observed. However, this does not appear to be the case.

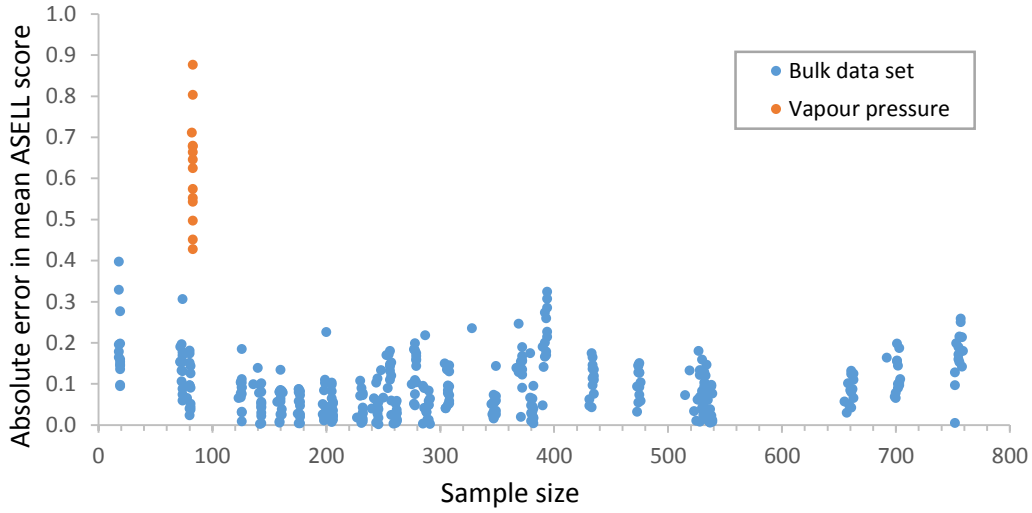


Figure 44: Sample size independence of ASELL mean score errors

A “baseline” level of systematic error remains for samples of any size, due to between sample variance. Average disposition towards positive response for entire sample groups changes from occasion to occasion and may differ between different groups, and this is not eliminated by simply increasing the number of observed cases sampled from a single group, at a singular time. The “absolute error” is the magnitude of difference between the observed mean score and the fair mean score, calculated from LLTM measures.

As shown in Figure 44, the size of the errors in the bulk data set appears largely independent of the sample size. This can be explained using an observation made previously in sections 4.1.4 and 4.1.6.2: errors in ASELL score arise not only from variance in student biases within individual sample groups, but also variance in the average bias between different sample groups. Here the term “sample group” is used to refer to a subset of the total observed data set, defined by the specific set of circumstances in which surveys were gathered. For example, all surveys gathered in the morning may form one sample group, whilst all surveys gathered in the afternoon may form another. Alternately, surveys gathered from students enrolled in Foundations of Chemistry may form one group, whilst students enrolled in Chemistry IA/B may form another. Different student circumstances and contexts (for example time of day, course enrolled) influence broad scale student disposition, as was noted previously as an explanation for the fact that student specific measures (β_E) change from occasion to occasion (see section 4.1). This affects the location of mean score value large sample sizes converge to, meaning different sample groups converge to different population means. It is differences between these population level means particular to the different groups which remain present, independent of the sample sizes used for each group.

A very narrow diversity of sample groups was combined in these analyses, therefore meaning that the between sample variance has been reduced very little. Sample sizes used are very large, however, meaning the within sample variance has little to no effect in most cases. As such, variance in ASELL mean score error appears constant, due almost entirely to the inherent differences between the small variety of sample groups combined in these data. Reducing the

impact of these constant errors would not only require large numbers of observations from each group, but also a diverse range of different sample groups.

Drawing a histogram of the errors observed (which reflect between sample variance), it can be seen that the distribution of errors for the bulk of the data appears roughly normal, with a standard deviation of approximately 0.1 (see Figure 45). It is this value which may be used as a quantification of the average size of any "baseline" level of error in these ASELL mean scores. Observations made here suggest that this constant margin of inherent error should be presumed to exist in ASELL mean score values, and should be included in any statistical analyses of ASELL mean score data.

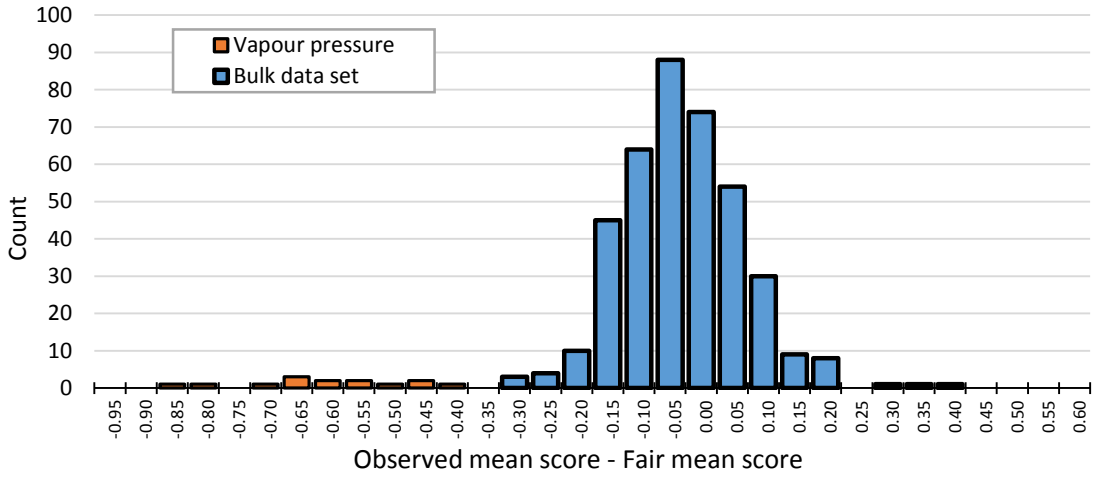


Figure 45: Distribution of errors in ASELL mean score

The unfairly judged "Vapour pressure" experiment appears as an outlier to the rest of the data gathered. The bulk of the calculated error margins in mean ASELL scores follow an approximately normal distribution with $\sigma \cong 0.1$. Because these errors are due almost entirely to between sample variance, this error margin should be considered as an expected level of variation in ASELL survey data, regardless of sample size.

4.3.3.3 Major identifiable factors

The now comparable values of δ_{LLTM} are merely combinations of the more fundamental underlying factors, and it is measures associated with these factors which become the focus of comparative studies when using the LLTM. The relative contribution of each of these estimated factors to the variance in ASLE survey responses can be roughly gauged by referring to the results of the factor analysis used to generate the final LLTM, shown in Table 21. As can be seen the vast majority of variance in δ_{PCM} values is explained by the first seven extracted factors only, whereas factors 8 through 12 all explain less than 1% of the variance in these measures (the other factors all explaining at least greater than 6% each). Were the number of factors to be retained decided using a scree plot, only these first seven factors would be retained judging by these values. The often used, but problematic factor extraction technique of retaining only those factors with eigenvalues of 1 or above would advise retaining only the first five factors. Here, in contrast, 12 factors have been retained based on an optimal balance between parsimony and proportion of observed data explained by the model (Figure 40 previously). Factors beyond factor 7 are thus considered necessary for a full explanation of the observed data, but show little substantial contribution in comparison to the other factors of the model. Factor numbers have been assigned based on relative proportion of variance in the initial δ_{PCM} measures explained.

Table 21: Variance in student independent measures explained by LLTM factors

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.583	25.590	25.590	3.065	21.895	21.895	2.540	18.144	18.144
2	3.133	22.377	47.967	2.811	20.077	41.972	2.313	16.519	34.663
3	2.020	14.431	62.399	1.726	12.331	54.303	1.249	8.921	43.584
4	1.599	11.423	73.821	1.195	8.532	62.836	1.172	8.373	51.957
5	1.213	8.663	82.485	.917	6.550	69.385	1.158	8.269	60.225
6	.784	5.600	88.084	.402	2.869	72.254	.927	6.618	66.844
7	.618	4.411	92.496	.221	1.578	73.832	.869	6.205	73.049
8	.356	2.543	95.039	.056	.403	74.235	.119	.848	73.897
9	.217	1.550	96.589	.004	.026	74.261	.098	.698	74.595
10	.155	1.110	97.699	.019	.132	74.393	.071	.510	75.105
11	.124	.884	98.583	.036	.260	74.653	.036	.260	75.364
12	.096	.687	99.270	.104	.741	75.394	.004	.029	75.394
13	.066	.471	99.740						
14	.036	.260	100.000						

"Initial eigenvalues" reflect the relative variance explained by each factor as estimated by preliminary principal component analysis, prior to factor extraction. Totals sum to the number of initial variables (14). "Extraction sums of squared loadings" refer only to the shared variance among the 12 factors retained following extraction by image factoring. "Rotation sums of squared loadings" are similar values computed following factor rotation. It can be seen that a sharp drop in the % variance explained occurs for factors 8 to 12 as compared with the seven major factors (1-7), particularly in the final rotated solution.

It should be noted that proportions of variance explained in Table 21 above were not calculated using all 128,881 individual data points following LLTM estimation, but instead refer to calculations performed only on the limited number of artificially equated δ_{PCM} values used to initially estimate the factor loading matrix. As such, these values serve as rough indicators rather than reflections of the properties of the final LLTM estimated. As was seen in Figure 41 previously, the final estimated LLTM was able to explain 99.56% of variance in non-LLTM values, rather than the 75.394% explained in the factor analysis results. The discrepancy is likely due to the re-estimation of non-LLTM measures (δ_{LLTM} + displacement, analogous to δ_{PCM}) during the Facets software Rasch model optimisation process and the increase in number of data points included.

The loadings of each factor onto each of the ASLE survey items further support the conclusion that first seven factors are responsible for the majority of variance in observed responses, whilst the remaining five factors contribute little. Table 22 displays the factor loadings associated with each factor, reflecting the correlation between factor measures and δ_{PCM} measures associated with the original fourteen survey items. It is these loading values which were utilised to generate the final Q matrix for the 12 factor LLTM, via Equation 40.

Table 22: Annotated factor loading matrix ($L = [l_{i,f}]$) for the 12 factor model

		Factor number (f)											
i	**Survey item**	1	2	3	4	5	6	7	8	9	10	11	12
1	This experiment helped me to develop my data interpretation skills	0.02	0.01	0.08	0.81	-0.10	0.19	0.02	0.01	0.01	-0.01	-0.01	0.00
2	This experiment helped me to develop my laboratory skills	-0.70	0.05	-0.07	0.46	0.33	-0.05	0.07	-0.01	-0.03	0.16	0.01	0.02
3	I found this to be an interesting experiment	-0.72	0.36	0.33	0.07	0.07	-0.18	0.24	0.05	0.15	0.04	-0.07	0.01
4	It was clear to me how this laboratory exercise would be assessed	0.13	0.84	0.13	0.01	-0.03	0.10	0.23	-0.05	0.18	-0.05	0.02	0.01
5	It was clear to me what I was expected to learn from completing this experiment	0.69	0.50	0.07	0.14	0.09	0.18	0.19	0.08	0.14	-0.06	0.03	-0.03
6	Completing this experiment has increased my understanding of chemistry	0.06	0.06	0.81	0.03	-0.12	-0.08	0.05	-0.05	0.01	0.00	0.00	0.00
7	Sufficient background information, of an appropriate standard, is provided in the introduction	-0.27	0.70	-0.09	0.30	0.16	0.14	0.12	-0.10	-0.01	0.15	-0.03	-0.03
8	The demonstrators offered effective supervision and guidance	0.12	0.10	-0.04	0.23	-0.02	0.69	0.13	0.02	0.00	0.00	-0.01	0.00
9	The experimental procedure was clearly explained in the lab manual or notes	0.07	0.76	-0.06	-0.14	0.21	-0.02	-0.04	0.06	-0.15	-0.02	0.00	0.01
10	I can see the relevance of this experiment to my chemistry studies	0.78	0.10	0.27	0.15	-0.13	-0.01	-0.17	0.02	0.00	0.11	-0.05	0.04
11	Working in a team to complete this experiment was beneficial	0.07	-0.18	0.56	0.09	-0.48	0.10	-0.21	0.29	-0.01	-0.01	-0.01	0.00
12	The experiment provided me with the opportunity to take responsibility for my own learning	-0.16	0.18	-0.19	-0.06	0.77	-0.03	0.06	0.02	0.00	0.00	-0.01	0.00
13	I found the time available to complete this experiment was	0.50	0.04	-0.07	-0.26	-0.25	0.46	0.40	-0.08	0.01	-0.02	0.16	0.00
14	Overall, as a learning experience, I would rate this experiment as	-0.27	0.29	0.01	0.09	0.20	0.31	0.67	-0.01	0.01	0.01	-0.02	0.00

Factor loading values in the table above reflect correlations between factor measures and survey item measures. Negative correlations are shown in blue, whilst positive correlations are shown in red, with darker colours reflecting correlations of greater magnitude. These values therefore reflect the "character" of each factor, described in terms of the original survey items: what each factor resembles.

These loading values may be used to identify what features of the laboratory experience each factor generally appears to reflect. That is, the loading values reflect the identity or "character" of each factor. A summary of the strongest observed loadings for each of the first seven factors and hence the assigned character of each is provided in Table 23. Factors 8 to 12 show

minimal loadings only for any given survey item, and their character is therefore unknown. This is not problematic, however, as these factors contribute very little to variance in observed data as discussed.

Table 23: Character of major factors contributing to ASLE survey responses

Factor	Strong negative loadings Characterise more negative values	Strong positive loadings Characterise more positive values	Assigned character
1	Laboratory skills development Interest	Relevance to chemistry studies Clear expected learning Time availability	*Theory focus (vs practical/ lab focus)*
2		Clear assessment criteria Clear procedure in manual Sufficient background information Clear expected learning outcomes	*Instructions*
3		Increased understanding of chemistry Teamwork beneficial	*Collaborative understanding*
4		Data interpretation skills development	*Data interpretation*
5	Teamwork beneficial	Responsibility for own learning	*Independent learning*
6		Demonstrator supervision and guidance	*Demonstrators*
7		Positive overall learning experience	*Unexplained overall*

A number of factors appear to load strongly on singular survey items and have therefore been assigned a character reflective of the content of those items (factors 4, 6 and 7 loading on survey items 2, 8 and 14 respectively). This does not necessarily imply that it is only these factors which contribute to each of these respective survey items, merely that these factors are primarily characterised by singular aspects of the laboratory experience and not others targeted by the survey. Factor 7, for example, is not the only factor to contribute to the overall learning experience: it is simply a factor that has no other clear defining characteristic, and appears unexplained by characteristics targeted by other survey items.

Other factors appear to have more complex character. The loadings of factor 1 appear to suggest a dimension whereby a stimulation of interest and development of laboratory skills comes at the cost of a perceived lack of relevance and clarity of learning objectives, as well as a lack of time to complete the experiment. Conversely, for learning objectives to be clear and relevant with ample time to address them, the task evidently typically lacks a development of laboratory skills and is less interesting. These characteristics appear to resemble the differences between hands-on “skills-based” experiments and experiments intended to reinforce lecture content rather than technical skills. Factor 1 has therefore been labelled as the spectrum from lecture theory focus to practical/ laboratory skills focus to reflect this. The quality of the instructional material provided appears to be a singular factor, with factor 2 loading on four separate survey items all concerned with the information provided to students. The character of factor 2 has therefore been assigned accordingly as the

"instructional material" factor. Factor 3 primarily loads on the perceived increase in understanding, but also has a prominent loading on the perceived benefit of teamwork. Factor 3 has been assigned as "collaborative understanding" to reflect this. This factor does not negatively load on responsibility for own learning, however, meaning it may therefore be more accurately characterised primarily by the increase in understanding. Another factor, factor 5, does appear to represent the contrast between independence and collaboration, loading positively on responsibility for own learning and negatively on the benefit of teamwork. Factor 5 has therefore been assigned as relating to "independent learning".

4.3.3.4 Identity of remaining low contribution factors

Despite the fact that factors 8 – 12 have little contribution to the variance in observed responses overall, their identity may still be useful information if it can be determined. However, their very low correlation with (and therefore resemblance to) aspects of the laboratory experience targeted by the ASLE survey makes characterisation of these factors problematic.

A clue as to the role of one of these factors, factor 8, can be gained from examining the effects of excluding factor 8 from the model. Factor 8 has no clearly discernible identity based on its factor loadings, but does appear to have some role in determining response to item 11 of the survey: "working in a team to complete this experiment was beneficial". The final LLTM computes the measure for item 11 as a sum of various contributions from the first 8 factors, with no contribution from factors 9 and onwards (see Table 24 later discussed). Computing the measures for this item by only considering the seven primary identified factors discussed previously therefore gives insight into the role of factor 8. Such a comparison is displayed in Figure 46.

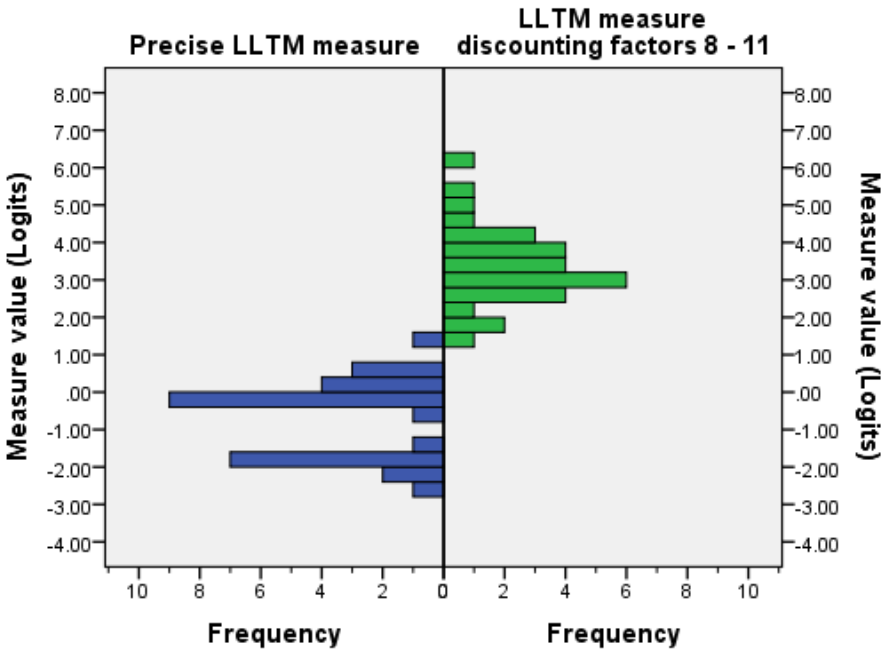


Figure 46: Role of factor 8 in determining the "benefit of teamwork" measure

The left of the figure shows student independent measures for item 11 of the ASLE survey, as computed using all factors involved. The right of the figure shows measures computed for the same item, this time only using the first seven identifiable factors in the model. Factors 9 onwards had no contribution in either case, meaning any change in the distributions observed is solely due to the exclusion of factor 8.

As can be seen, student independent measures for item 11 appear in two clusters when all factors of the model are included. When observing which experiments' measures appear in which cluster, it is quickly discovered that all values in the upper cluster ($\delta > -1$) are, without exception, the experiments conducted in pairs. Conversely, again without exception, all values in the lower cluster ($\delta < -1$) are from experiments conducted individually. This demonstrates quite clearly that item 11 of the ASLE survey yields an effectively binary response: when asked if teamwork was beneficial, students respond positively in all cases they worked in pairs, whereas they respond negatively if they worked individually.

Curiously, however, this binary response is entirely absent unless factor 8 is included in the calculation. Computing measures for item 11 using only the first 7 factors, a single cluster of measurements is observed. Reasons for this are speculative, but the conclusion must be drawn that factor 8 "corrects" the measure for item 11 back to a binary response once the effects of the first 7 factors have been accounted for. As has been seen in the identity of the first seven factors, teamwork or lack thereof is an inherent defining feature of at least two factors underpinning ASLE survey responses (factors 3 and 5), and therefore a full range of (non-binary) perceptions regarding the role of teamwork is accounted for by the students. However, it appears that despite a tacit acknowledgement of this full spectrum when answering other items of the ASLE survey, students are still compelled to respond to item 11: "working in a team to complete this experiment was beneficial", as if it simply asked "did you work in a team?" The correcting of the full spectrum of possible perceived levels of teamwork back to a binary response is the observed role of factor 8, judging by this analysis.

Behaviour such as this could never be expected to correlate strongly to any item asked on the ASLE survey, and therefore never manifest as a substantial factor loading value which could be used to "characterise" the factor's identity. Factors 9 through to 12 could have similarly obscure roles in survey response, however this remains unknown. It is effects such as this, which could never have manifested in the results of any simple factor analysis, which justify the inclusion of factors beyond those which are clearly identifiable (the first seven here). The identity of factors contributing to ASLE survey responses which cannot resemble any question asked on any survey, such as factor 8, could be a problem unable to be rectified easily.

4.3.3.5 Factor impacts

The true advantage of the LLTM generated is not the identification of the key factors involved in ASLE survey response, as this could largely be achieved from a simple factor analysis alone. Rather, the greater advantage is the quantification of each of these factors' contribution to each individual ASLE survey item. The final Q-matrix estimated from this factor loading matrix is shown in Table 24. These values are the "weightings" of each factor's contribution to the original ASLE survey items' measures. Of note, the pattern in weightings does not necessarily match the patterns observed in the factor loadings. This is because of different observed variances in item measures estimated in the Partial Credit Model for each question. These weightings reflect the "impact" of each factor, rather than reflecting the "character" of each factor like the factor loadings. Negative impacts are coloured blue, whilst positive impacts are coloured red, with darker colours indicating higher magnitude of impact. Of note, factor 12 has zero impact on any survey item, and can therefore be discounted from consideration entirely. It's inclusion in the model seems to be an advantage only in that it allows more accurate estimation of the other 11 factor weightings. It should be noted that Rasch measurement software other than *Facets* could accommodate non-integer values in the Q-matrix, giving factor 12 some small degree of impact.

A wealth of information regarding how to improve all fourteen specific aspects of the laboratory learning experience targeted by the ASLE surveys is contained within the Q-matrix. The Rasch measure reflecting "objective" quality of the experiment with respect to any given survey item can be known by making use of the Q-matrix coefficients and measures associated with each of the twelve basic factors identified above, for the experiment concerned (Equation 40). Impacts of altering any of these twelve factors of the laboratory learning experience can also be known and quantified using the Q-matrix coefficients. The full breadth of conclusions regarding best practice in structuring laboratory learning exercises gained from this investigation is therefore highly extensive and to be discussed separately (see section 4.4).

Table 24: Annotated weighting matrix ($Q = [q_{i,j}]$) for the 12 factor model

		Factor number ($j = f$)											
i	**Survey item**	1	2	3	4	5	6	7	8	9	10	11	12
1	This experiment helped me to develop my data interpretation skills	0	0	1	7	-1	2	0	0	0	0	0	0
2	This experiment helped me to develop my laboratory skills	-15	1	-1	10	7	-1	2	0	-1	3	0	0
3	I found this to be an interesting experiment	-10	5	5	1	1	-3	3	1	2	1	-1	0
4	It was clear to me how this laboratory exercise would be assessed	1	6	1	0	0	1	2	0	1	0	0	0
5	It was clear to me what I was expected to learn from completing this experiment	5	3	1	1	1	1	1	1	1	0	0	0
6	Completing this experiment has increased my understanding of chemistry	1	1	7	0	-1	-1	0	0	0	0	0	0
7	Sufficient background information, of an appropriate standard, is provided in the introduction	-2	5	-1	2	1	1	1	-1	0	1	0	0
8	The demonstrators offered effective supervision and guidance	1	1	0	1	0	3	1	0	0	0	0	0
9	The experimental procedure was clearly explained in the lab manual or notes	1	8	-1	-1	2	0	0	1	-2	0	0	0
10	I can see the relevance of this experiment to my chemistry studies	7	1	2	1	-1	0	-1	0	0	1	0	0
11	Working in a team to complete this experiment was beneficial	2	-4	14	2	-11	2	-5	7	0	0	0	0
12	The experiment provided me with the opportunity to take responsibility for my own learning	-1	1	-1	0	5	0	0	0	0	0	0	0
13	I found the time available to complete this experiment was	14	1	-2	-8	-7	13	12	-2	0	-1	5	0
14	Overall, as a learning experience, I would rate this experiment as	-2	2	0	1	1	2	5	0	0	0	0	0

Note that the matrix shown above is not the full Q-matrix. The full matrix has matrix elements for columns j=13 to j=26 as described in Equation 40 and can be seen in full in the supporting information (section 7.6.3). The matrix elements shown are those relevant to the twelve basic experiment specific factors (indexed by f) underpinning survey question responses.

4.3.4 Discussion

4.3.4.1 Successful model estimation

The fact that the 12-factor LLTM model yields an AICc value lower than the initial partial credit model definitively establishes that this technique of Q-matrix estimation achieves the desired outcome. For the data set used to estimate the matrix, the explanatory model of the data has

been successfully reduced to a smaller number of factors for each experiment (7 major factors, 12 factors total) than the initial fourteen, with no evident loss in the proportion of observed data explained. Generalising the model to a slightly wider data set, the estimated LLTM still appears superior to the initial partial credit model.

The outstanding success of this technique was achieved despite two key sources of error in the Q-matrix estimation:

1) Artificial equating of two experiments to resolve data connectivity issues
2) Rounding errors in the conversion of factor loadings to Q-matrix weightings

Utilising Rasch measurement software capable of using non-integer values for Q-matrix weights and estimating the Q-matrix from a fully connected data set would resolve these sources of error, generating a model which fits the data even more closely than the one presented here. A model such as this would be capable of modelling even more than the observed 99.56% of variance in δ_{PCM} values obtained here. Further, this study's success was achieved with a near inadequate number of data points for the factor analysis. Evidently this method is capable of estimating a superior model for a given data set despite this. The use of data points obtained from a small number of experiments limits only the generalizability of the model obtained, not the capability of estimating a superior model for the given data set.

The fact that the final model estimated here was generated from the measures associated with only 23 experiments is a substantial limit on the generalizability of any conclusions drawn from the model features. However, the LLTM estimated still represents a significant improvement in understanding: the previous partial credit model contained no inherent information about why each experiment is associated with the set of quality measures observed (δ). The final LLTM, however, explicitly reveals patterns in the quality measures observed, explaining them as combinations of more basic factors which are identifiable through their factor loading values on the initial survey items. The LLTM therefore represents an advance in knowledge.

4.3.4.2 Corrected standard error in a fair mean ASELL score

Due to the data connectivity afforded by the LLTM, observed ASELL mean score results could be easily contrasted with fair predictions of the LLTM, which eliminate errors introduced through broad scale biases in individual samples.

Data presented here can be used to derive a corrected, more accurate formulation for the standard error in any given ASELL score. The fair mean ASELL score (A_{fair}) can be considered as the observed mean score ($A_{observed}$) minus the error introduced due to broad scale bias in the sampled group (E_{bias}):

$$A_{fair} = A_{observed} - E_{bias} \quad \textbf{44}$$

Through the variance sum law, this therefore implies that the expected variance in a fair ASELL mean score can be described by the following, presuming the observed mean score and the bias present are independent:

$$var(A_{fair}) = var(A_{observed}) + var(E_{bias}) \quad \textbf{45}$$

Given that the standard error is simply the square root of the sample variance, Equation 45 is effectively an equation for the most appropriate error margin in any sampled ASELL score. The $var(A_{observed})$ term is the variance in the observed mean score calculated: the within sample variance. The $var(E_{bias})$ term is the variance introduced through differences between the biases of separate samples: the between sample variance. These two terms sum to give the total variance in a fair ASELL mean score $var(A_{fair})$. Using the central limit theorem to obtain $var(A_{observed})$ and Figure 45 to obtain $var(E_{bias})$, the standard error in a fair ASELL mean score may be given as follows:

$$SE(A_{fair}) = \sqrt{\frac{\sigma_A^2}{n} + 0.01} \qquad \textbf{46}$$

Where SE is the standard error, σ_A^2 is the observed variance in scored ASELL survey responses and n is the sample size. This corrected standard error value could be used in T-tests conducted on ASELL mean score data, which are far more accessible to most ASELL survey researchers. It can also be inferred that an error margin of approximately 0.1 score units is to be expected for any ASELL score obtained, regardless of sample size. This can be used to suggest a simple rule of thumb that no significant difference in two ASELL mean scores can be claimed unless those two means differ by at least 0.1 score units.

This specific minimum error value of 0.1 score units may well be particular to this data set: the value is a reflection of the degree of difference between the variety of sampling occasions and contexts combined in these analyses. Notably, the two student cohorts: Foundations of Chemistry IA/B and Chemistry IA/B, are expected to have different perspectives and biases towards the experiments, due to differences in their backgrounds. Similarly, students conducting experiments at different times of day or at different times during the week may have consistent differences in their average predispositions. It would be unexpected, however, to observe differences in average bias greater than these in most cases: the two separate cohorts particularly are quite dissimilar, and most cases in which data sets are compared are liable to use cohorts more similar than these in the comparison by design. If cohorts of a greater degree of dissimilarity are used, however, the value of $var(E_{bias})$ used above (0.01) may need to be increased.

The corrected standard error formula presented above may prove useful for more rigorous statistical testing, but use of ASELL surveys in practice does not often need to be this precise. Surveys are often used for exploratory purposes, and rough guidelines for significant differences are usually sufficient. By considering the maximum possible value of σ_A^2 in the formula above, a guideline for the expected maximum standard error value for an ASELL mean score gathered from n samples can be estimated. Assuming the distribution of scored responses gathered has a singular peak, the maximum possible variance in the observed scored responses (σ_A^2) arises from a distribution whereby each response category has the same proportion of responses: a uniform distribution. Calculating the population variance in this case (as per Equation 10) yields a value of 2. This value can be substituted for σ_A^2 in either the corrected (Equation 46) or uncorrected (Equation 11) formula of the standard error value to yield an expected maximum error margin (SE_{max}) in the mean ASELL score. As can be seen in Figure 47, the correction for between group variance has little practical significance. Standard error values at sufficient sample sizes appear not to exceed approximately 0.25 score units, regardless of the inclusion of any correction.

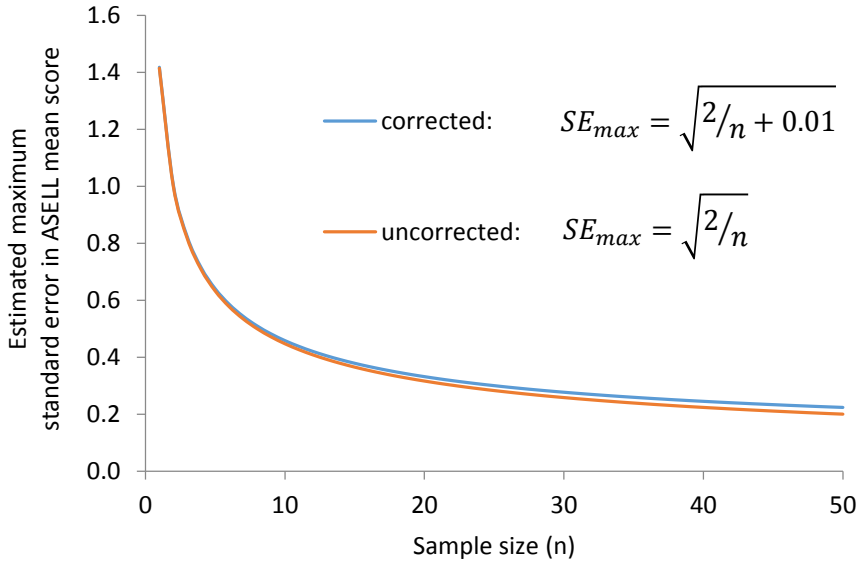


Figure 47: Maximum expected standard error values in calculated ASELL mean scores

The corrected standard error value differs little from the uncorrected value. At appreciable sample sizes ($n \geq 30$), the standard error in a mean ASELL score is not expected to exceed approximately 0.25 score units. This can be used as a rough guideline for exploratory comparisons.

4.3.4.3 Student independence of results

Recently, Barrie *et al.*[61] also conducted a factor analysis on ASLE survey response data, discussing the role of various factors in generating a positive overall learning experience. Factors reported in this previous study differ substantially from those reported in the new analysis presented here, and this is to be expected for a number of reasons. The differences may partly be due to the fact this study exclusively made use of data from a single year level at a single university (as opposed to the broader scope of the Barrie *et al.* study), but the differences are far more prominently due to different estimation methodology.

Because Barrie *et al.* base their factor analysis on scored individual responses, they conflate student dependent and student independent effects (see section 3.2). As the majority of variance in individual ASLE survey data points is due to student dependent effects (as identified in section 3.3), correlations underpinning the factor analysis in the Barrie *et al.* paper will reflect correlations between student dependent factors (biases) far more than they reflect correlations between student independent factors (experiment quality) (see section 4.1.6.3). In contrast, the study presented here exclusively analyses student independent measures, meaning the factors estimated reflect properties of the experiments evaluated, not the students doing the evaluation.

As an example, in the Barrie *et al.* study, scored individual responses to survey item 14 (overall learning experience) correlate with scored responses to survey item 7 (concerning background material provided). In this study, Rasch measures for these same items also correlate. However, the conclusions which can be drawn are very different. From the score based study, an appropriate conclusion would be:

> ***Students** who rate their overall experience highly also commonly rate the background information to be sufficient.*

Whereas from the Rasch based study, an appropriate conclusion would be:

Experiments *for which the overall learning experience measure is high also commonly have a high background information sufficiency measure.*

The former conclusion, most appropriate for the Barrie *et al*. study, could feasibly be the case even if background information quality were irrelevant to genuine overall experience: it could be that students in a "good mood" typically answer both questions more positively regardless. The latter conclusion, able to be drawn from this LLTM formulation, has controlled for student dependent factors such as this.

What becomes apparent is that the conclusion able to be drawn from the score based study is not necessarily informative of good experiment design: the correlation could appear simply because students who have a high positive bias to one question also happen to have a high positive bias to other questions, regardless of the experiment conducted at the time. In fact, based on results presented in section 4.1, this is precisely the way student responses operate. The use of the Barrie *et al.* study in informing best practice for design of experiments is, for this reason, highly questionable. In contrast, the Rasch technique used here exclusively makes use of measures that have been shown (see section 4.1) to reflect 'objective' measures of quality for the experiments themselves, independent of the biases of students responding.

Factors extracted in the Barrie *et al.* paper are therefore expected to reflect factors underpinning student disposition towards positive response, not factors underpinning experiment quality. Had the factor analysis been conducted on mean scores – which "average out" student bias effects to a degree – this would be less of an issue. However, doing so would have reduced the study to having only 784 data points for the factor analysis as opposed to the reported 3153. Additionally, even in the case of using mean scores, only 56 responses per experiment on average may not have been sufficient to overcome all student bias variations within each group. Even if it were sufficient to do so, this would still not remedy the remaining between-group variation in student biases (see section 4.1.4).

The use of a larger number of data points in the Barrie *et al.* study highlights one substantial difficulty with the methodology used here: data from a large number of experiments are needed for this type of analysis to be generalizable. Though generation of the LLTM here made use of nearly three times as many individual responses, the factor analysis was based on only 322 Rasch measures, which was seen to be barely adequate at best. As has been discussed, however, the model generated still represents an improvement in understanding.

4.3.4.4 Factor extraction: objectivity and quantification

Another advantage of this technique over the previous score-based factor analysis is the identification of how many factors to extract via statistical means rather than judgement by the researcher. In the previous study by Barrie *et al.*, the appropriate number of factors in model was decided by retaining the smallest number of factors whilst meeting somewhat arbitrary (though common) researcher-chosen criteria. The method of selecting for eigenvalues ≥ 1 was used, which has been shown to be less preferable than other alternate methods.[209] The technique used in the study discussed here requires no input by the researcher as to the number of factors extracted: the number of factors is determined objectively by the algorithms involved, which are based on well-established statistical considerations of parsimony and fit of data to the model. This also means that factors with unconventional characteristics necessary for the model's accuracy are retained where they would otherwise be dismissed inappropriately. The retention of factor 8 in this model is a

prime example. This is possible here because the factor extraction is unaffected by the expectations or predispositions of the researcher.

A drawback of this methodology, however, is the substantial amount of time needed for the estimation: each data point corresponding to a different LLTM in the plot at the right of Figure 40 required approximately 10 to 13 hours of computing time. This is partly due to the complexity of the model, but more likely due to the fact that the "landscape" of possible solutions to each LLTM is a very "flat" surface, meaning the optimal solution differs very little from many other solutions, thereby increasing the time needed to find the global minimum during the optimisation.

The final LLTM obtained in this study also has the advantage of explicitly quantifying the effects of each factor on the experiment's quality with respect to each item of the survey. In this way, measures such as those for the "overall learning experience" item are directly explained in terms of the experiment's basic properties. Moreover, they are explained in a quantitative manner: it is known which factors contribute the most, which contribute the least and the proportion yet to be explained. As an example, the LLTM formulation for the "overall learning experience" measure (associated with survey item 14) is expressed below in Equation 47 (discounting the item location parameter μ_{14}). The column vector containing integer value coefficients is lifted directly from the appropriate row of the Q matrix (Table 24), whilst the vector of η values contains measures of the seven major factors identified, each labelled using their assigned character (see Table 23). The integer weightings are constant for all experiments, whereas the factor measures (η) are all experiment specific. The final measure value δ can be input into Equation 43 (as δ_{LLTM}), which in turn can be substituted into Equation 1 to give the probability of observing a student to respond in each of the five response categories on the ASLE survey.

$$\delta_{14\,(overall\ learning\ experience)} = \begin{bmatrix} -2 \\ 2 \\ 0 \\ 1 \\ 1 \\ 2 \\ 5 \end{bmatrix} \cdot \begin{bmatrix} \eta_{theory\ focus} \\ \eta_{instructions} \\ \eta_{collaborative\ \ understanding} \\ \eta_{data\ interpretation} \\ \eta_{independent\ learning} \\ \eta_{demonstrators} \\ \eta_{unexplained\ overall} \end{bmatrix} \quad \textbf{47}$$

Here the Q-matrix reveals the relative weighting of each factor's contribution to the overall learning experience. Included is a contribution of 5 × factor 7, which itself is characterised as mostly resembling overall learning experience. This shows and explicitly quantifies a substantial contribution to overall learning experience which is not explained by the topics of other items within the ASLE survey. The model estimated here thereby identifies a gap in knowledge, revealing a goal for future research. A substantial portion of the overall learning experience measures are explained in this model, however. As previously stated, the extent of these contributions are explicitly quantified in the model, advancing understanding further than merely a "yes or no" answer to the question of whether various considerations are of importance. A full spectrum of partial contributions is recognised in this model, as opposed to only identifying full, partial or absent factoring as was the case in the study by Barrie *et al*. Equations similar to Equation 47 are obtained for all fourteen items of the ASLE survey through this model.

Having obtained this model, the results can now be analysed in depth to determine how experiments should be structured to be received as positively as possible by students, which positive features of the laboratory experience are more important than others, and the likely effects of changing various features of the laboratory experience both qualitatively and quantitatively.

4.3.5 Conclusion

A technique has been devised and implemented here to yield a Linear Logistic Test Model for the ASLE survey data. The method yielded the best explanatory model of the data to date, resulting in a more parsimonious solution without sacrificing the proportion of data explained by the model. The objective, sample independent nature of the model means conclusions can now be drawn regarding best practice for the design of laboratory exercises, independent of the specific students conducting the activity. This topic is to be discussed in the next section of the results presented in this thesis. Formulation of the model has also allowed for the derivation of a correction to the standard error in a given mean score value calculated using more typical integer scoring methods. This correction is readily usable for the majority of ASELL survey practitioners.

4.4 Recipes for a positive laboratory experience: pedagogical implications of the ASLE data LLTM

4.4.1 Outline

The previous study described a method for using observed data to derive a Linear Logistic Test Model (LLTM) capable of adequately explaining ASLE survey response patterns. Application of this procedure resulted in the generation of a factor loading matrix, detailing the character of 12 factors: seven major interpretable factors (Table 25) and 5 factors with little clear contribution or identity. As part of the LLTM, a Q matrix was also estimated, detailing how each factor contributes to responses for each ASLE survey question. This section will now discuss the key features of the estimated model, revealing more practical interpretations of the results able to inform future teaching practice. A summary of the seven interpretable factors and the symbols used to refer to them in the following discussion is presented in Table 25 below. The low-contribution factors, factors 8 through 11 (not presented in Table 25), will be referred to as η_f where f is the factor number.

Table 25: interpretable factors contributing to laboratory perceptions

#	Symbol	Factor description
1	$\eta_{theory\ focus}$	Positive values for this factor reflect a focus on (lecture) theory, whereas negative values reflect a focus instead on practical activity.
2	$\eta_{instructions}$	Positive values for this factor reflect high quality of the instructional notes provided
3	$\eta_{collaborative\ \ understanding}$	Positive values for this factor reflect an increase in perceived understanding of chemistry, associated with the benefit of teamwork
4	$\eta_{data\ interpretation}$	Positive values for this factor are associated with perceived development of data interpretation skills
5	$\eta_{independent\ learning}$	Positive values for this factor are associated with perceived opportunity to take responsibility for own learning, whereas negative values are instead associated with a sense that teamwork was beneficial
6	$\eta_{demonstrators}$	Positive values for this factor are associated with students frequently reporting effective supervision and guidance by their demonstrator.
7	$\eta_{unexplained\ overall}$	Positive values for this factor are associated with a more positive reported overall learning experience, unexplained by other factors

Refer to Table 22 for more precise characterisation of each factor's defining features and Table 24 for quantifications of their impacts on each ASLE survey item.

4.4.2 Skills-based versus theory-based laboratory activities

The fact that $\eta_{theory\ focus}$ appears as the first factor (explaining 18% of the variance in δ_{PCM} estimates, see Table 21 previously) appears to suggest that an experiment's focus on (lecture)

theory or on practical activity is the single most important consideration in designing an experiment which will rate positively with students. Student interest and "overall laboratory experience" as reported in items 3 and 14 of the ASLE survey respectively appear to improve as focus is directed away from theory and instead towards practical activity. However, this appears to come at the cost of reduced clarity of the learning objectives and a lack of time availability. Measurements for this factor associated with each surveyed experiment are displayed in Figure 48.

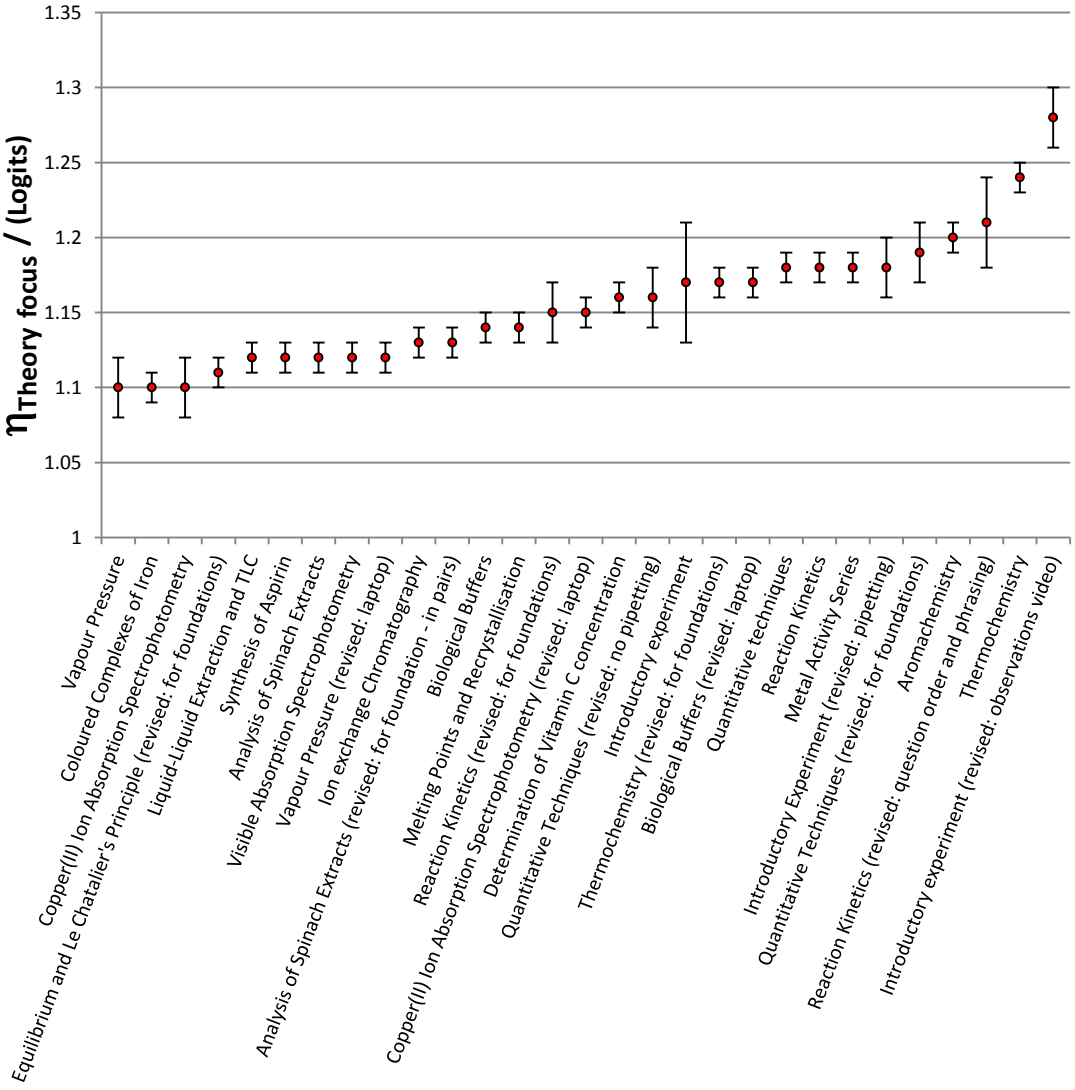


Figure 48: LLTM basic parameter measures for factor 1 (theory focus)

More positive values correlate with perceived relevance, clarity of expected learning outcomes and time availability whilst negative values correlate to increased interest and perceived development of laboratory skills. Error bars represent the standard error value of the measure. Experiment titles are sequenced from lowest measure to highest measure. A summary of experiment descriptions has previously been presented in section 2.1.3.

Given pre-existing chemistry education literature detailing the importance of practical activity for engaging with chemistry at the "macroscopic" and concrete level,[13, 257, 259, 260, 304-306] as well as the substantial importance of bridging the gap between concrete observations and formal

theory in science education,[261, 307, 308] it is tempting to interpret this factor as a measure of whether students perceive the experiment (or the concepts involved) to be more concrete and hands on or more abstract and theoretical in nature. However, interpreting $\eta_{theory\ focus}$ as a measure of concrete versus formal interaction with the relevant concepts would lead to some very complex rationalisations of the observed factor values.

For example, the "Metal activity series" experiment has one of the higher $\eta_{theory\ focus}$ measures, despite the fact that students work almost exclusively with direct experimental observations in this practical. Interpreting $\eta_{theory\ focus}$ as a concrete vs abstract measure, the concrete nature of "Metal activity series" would suggest a low $\eta_{theory\ focus}$ measure, not a high one. A clearer understanding of the $\eta_{theory\ focus}$ factor values comes from recognition of the factor's defining characteristics: clear, relevant learning objectives as opposed to high development of laboratory skills, *from the perspective of the students*. The explanation for the "Metal activity series" experiment's high $\eta_{theory\ focus}$ measure comes from the fact that "Metal activity series" is aligned with lecture content, as part of the Foundations of Chemistry course. The experiment therefore has a high focus on theory that students recognise from lectures, leading to perceptions of relevance and clarity of the intended learning outcomes: primary components of the $\eta_{theory\ focus}$ factor. It is also reasonable to expect that students therefore recognise the purpose of the practical to be about exploring the lecture content *as opposed to* development of any laboratory skills. Under this interpretation, a high value of $\eta_{theory\ focus}$ is logical.

Other experiments of high $\eta_{theory\ focus}$ measure are also predominantly experiments conducted as part of the Foundations of Chemistry course, which are similarly aligned well with lecture content. In contrast, many experiments of low $\eta_{theory\ focus}$ measure are those without lecture content alignment, often conducted by the Chemistry IA/IB cohort, whose practical activities were randomly sequenced. One apparent exception to this broad explanation of the observed values appears to be "Equilibrium and Chatelier's principle" as first revised for the Foundations of Chemistry cohort. From the teacher's perspective this experiment aligns well with the lecture content, and therefore should have a high $\eta_{theory\ focus}$ measure. However, the experiment is observed to have a low $\eta_{theory\ focus}$ value despite this. The measure assigned also appears to explain the observed responses well given the fit statistics (see Table 26), and cannot be dismissed as a statistical anomaly given its narrow error margin.

Recalling that ASLE data reflects *student perceptions* rather than teacher intentions may resolve this apparent problem. The low $\eta_{theory\ focus}$ measure of "Equilibrium and Chatelier's principle" as first revised for the Foundations of Chemistry cohort may be interpreted to suggest students typically do not see the connection to the lecture content, despite teacher intent. This interpretation is sensible in light of the conceptual complexity and abstract reasoning required to draw conclusions from the observations made in the experiment; the topic of equilibrium is known to be problematic with respect to connecting concrete experiences with abstract theory,[260, 305, 309-312] and this may explain why students might "miss the point" of the experiment. Observations made in "Equilibrium and Chatelier's principle" largely consist of changes in pH and colour, with conclusions about equilibria only accessible through subsequent deductions and correct interpretation in light of more abstract theory. Disconnects between macroscopic phenomena and the unseen "sub-microscopic" level of chemistry such as this have often been cited as cause for students being unclear about what

intended learning outcomes they are supposed to gain from their observations,[257, 259, 260, 306, 313] and such a reduced clarity of intended learning outcomes is a defining characteristic of low $\eta_{theory\ focus}$ measures (see Table 22).

Generally speaking, the existence of this factor and the trends in its observed values appear to suggest that a clear connection to lecture content, from the perspective of the students, is the primary (known) factor in laboratory experience perceptions. Sequencing of the experiment with lecture content appears not to be sufficient in achieving this, however: the connection must be readily apparent to the students, not just the teacher. This connection may be achieved through careful design of experiments such that the relevant theory is clearly associated with the experimental observations and procedures; the macroscopic observations need to be *illustrative* of the abstract concept to achieve a high measure of $\eta_{theory\ focus}$, in conjunction with the experiment's sequencing with the appropriate lecture material. Achieving this effectively lends itself to promoting clarity of the learning objectives of the practical and a perception of relevance (as seen in the Q-matrix, Table 24).

Unfortunately, this also has the consequences of decreasing student interest and reducing perceived development of laboratory skills. The explanation for the impacts on student interest could simply be that students find lecture theory boring, or alternately that they find novel material interesting; reasons are not clear from the model alone. In either case the trend is somewhat unsurprising, as is the fact that students perceive the experiment to be less about developing laboratory skills if they perceive a stronger connection to lecture content. Similarly, increased time availability for practicals more concerned with lecture theory may be intuitively explained by the fact that practicals more directed towards technical skills typically require more time consuming manual "work" to complete.

Additionally, benefits of having a clear connection to lecture content seem only to involve increased *clarity* of the intended learning outcomes, with only a small impact on actually attaining them. The Q-matrix weighting coefficient of this factor on the survey item "completing this experiment has increased my understanding of chemistry" is only +1, whereas other factors appear to have far more substantial contribution. Again, however, the perceptions of students may not necessarily align with teacher perspectives or indeed with the actual objective gains in student learning resulting from the practical. The implication for teachers here simply appears to be that a stronger and easily recognised connection to lecture content clarifies learning goals for students, but at the cost of students liking the exercise (both in terms of their interest and perceived "overall learning experience"). In this way, student preference appears to lean towards "skills-based" laboratory tasks rather than "lecture theory-based" laboratory tasks.

4.4.3 Collaborative and independent learning

The notion that learning is a social process is a very familiar concept in education, often viewed as being at the heart of constructivist ideologies. The views of Vygotsky[314, 315] notably emphasise the significance of social interactions during the learning process, whilst past and current trends in pedagogy of science (and other disciplines) have emphasised the benefits of small group discovery activities,[316-318] guided inquiry activities involving group work[319-321] and the benefits of involving peers in problem based learning activities.[322-326]

The appearance of the third factor in the ASLE data LLTM: $\eta_{collaborative\ understanding}$, seems to suggest that interaction with peers is relevant to student enjoyment of their activities as well as in their learning, as suggested in the literature. This factor appears most strongly characterised as relating to perceived understanding of chemistry (item 6) and also has a slightly less strong association with perceived benefits of teamwork (item 11), as seen in Table 22. However, the impacts of high values for this factor appear reversed to this: item 11 (concerning teamwork) displays the most improvement as $\eta_{collaborative\ understanding}$ is increased (Q-matrix weight of +14), with perceived understanding slightly less so, but still substantially improved (Q-matrix weigh of +7). A possible explanation for this could be that increasing understanding (through groupwork) is the heart of the factor's definition (hence the factor loadings observed), whereas in practice, students notice the fact they are now working together more than they notice the gains in understanding achieved (hence the Q-matrix coefficients observed). Regardless of which of these features can most accurately be described as the primary description of this factor, it seems clear that a perception of increased understanding of chemistry and a perception of teamwork being beneficial are so strongly associated that they manifest as one singular indistinguishable factor.

Perceived increase in the understanding of chemistry as reported by students on ASLE surveys seems almost entirely due to this single factor, with little contribution from any other factors identified in the LLTM (see Equation 48 below, discounting item location parameter μ_6). The only other ways to promote understanding appear to be clear focus on lecture theory, well written instructional material, less responsibility for the student's own learning and reduced guidance from demonstrators, though these factors each appear far less important by comparison.

$$\delta_{6\,(understanding\ of\ chemistry)} = \begin{bmatrix} 1 \\ 1 \\ 7 \\ 0 \\ -1 \\ -1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} \eta_{theory\ focus} \\ \eta_{instructions} \\ \eta_{collaborative\ \ understanding} \\ \eta_{data\ interpretation} \\ \eta_{independent\ learning} \\ \eta_{demonstrators} \\ \eta_{unexplained\ overall} \end{bmatrix} \quad \textbf{48}$$

The reasons behind the strong association between perceived increase in understanding and perceived benefits of teamwork are not made clear by the model itself. It could be that an experiment which improves student understanding often tends to elicit (or even require) conversation with peers, or that peers are more likely to be of assistance when needed. It could also be that experiments which require teamwork naturally assist with student understanding. The "direction" of causality here is unknown, but it seems reasonable to suggest that the existence of the $\eta_{collaborative\ understanding}$ factor was not unexpected based on existing education literature. Not only does the value of this factor influence the two survey items reflecting its primary characteristics (items 6 and 11 concerning understanding and teamwork respectively), but also positively influences student interest in the activity (Q-matrix weight of +5).

The role of teamwork (or lack thereof) on student perceptions of learning is further revealed by another factor; $\eta_{independent\ learning}$. This factor appears much more directly characterised as the spectrum from independent work (for positive values) to collaborative work (for

negative values). Somewhat confirming the conclusions drawn from the $\eta_{collaborative\ understanding}$ factor's behaviour, $\eta_{independent\ learning}$ reveals a slightly negative impact on understanding as students report they are more able to take responsibility for their own learning (see Equation 48 above). Positive values of $\eta_{independent\ learning}$ also show a small positive impact on perceived overall learning experience (Q-matrix weighting of +1), which was not observed for the $\eta_{collaborative\ understanding}$ factor. This suggests students prefer individual work to group work, aligning with the previous conclusion that "skills-based" practicals (which benefit from independent learning) are preferred to "theory-based" practicals (which benefit from collaborative understanding), seen when analysing the $\eta_{theory\ focus}$ factor previously. More independent learning also appears to lead to a decrease in perceived availability of time (Q-matrix weighting of -7), again in keeping with what was seen for "skills-based" practicals generally when analysing $\eta_{theory\ focus}$.

One of the largest impacts of the $\eta_{independent\ learning}$ factor, however, is the substantial increase in reported development of laboratory skills as $\eta_{independent\ learning}$ is increased (Q-matrix weighting of +7, see Equation 49 in later discussion). Evidently, students perceive their laboratory skills to be better developed in the experiments where they report teamwork to be of little benefit and instead have responsibility for their own learning. The argument could be made, therefore, that skills-based practicals are more beneficial when conducted independently.

However, this initially seems to conflict with the pedagogical implications of promoting "increased understanding of chemistry", which was seen to be improved when group work is seen to be beneficial (seen analysing the $\eta_{collaborative\ understanding}$ factor). The apparent conflict may be reconciled if "understanding of chemistry" is primarily interpreted by students to mean understanding of chemistry *theory*, not laboratory skills. Rather than the model presented here containing any contradiction, this interpretation would then simply imply that the most appropriate pedagogy depends on the primary learning objectives of the practical: skills development is promoted by independent learning, whereas understanding of theory is promoted by teamwork.

4.4.4 Different factors may apply for different student groups

Given the hypothesis that individual work is beneficial for "skills-based" practicals as suggested in the preceding discussion, it is a natural next stage of research to examine the changes observed when the identical "skills-based" experiment is conducted in pairs as opposed to individually. The "Analysis of spinach extracts" experiment is one such experiment, and is in fact the only case in this study where whether students worked individually or in pairs was ever modified. The experiment appears to be perceived as one of the more "skills-based" experiments, with the relatively low $\eta_{theory\ focus}$ measures of 1.12 ± 0.1 Logits and 1.13 ± 0.01 Logits for the Chemistry IB (individual work) and Foundations of Chemistry IB (working in pairs) iterations respectively (see Figure S 59 in the supporting information). Given the preceding discussion, it would therefore be expected that perceived development of laboratory skills and overall learning experience would receive more positive responses in the case of working individually. The empirical Rasch measures associated with each survey item (δ), reflecting "objective" quality of the learning experience, were compared to test this prediction. Measures estimated when students conduct the experiment in pairs as opposed to the measures estimated when students conduct the experiment individually are contrasted in Figure 49.

As would be expected, the measure for the item pertaining to the perceived benefit of teamwork appears greatly decreased when working individually. The next greatest affected facet of the learning experience is the perceived time availability, with group work appearing to be associated with less available time. This is in direct contrast to the implications of the behaviour of the $\eta_{independent\ learning}$ factor described previously. Skills development also behaves contrary to prediction: the student group working in pairs appeared to broadly report a greater perceived development of laboratory skills, in contrast to the prediction made based on the effects of the $\eta_{independent\ learning}$ factor. This specific "skills-based" practical appears not to have benefited from individual work as hypothesised. Patterns in the measures for this experiment appear to differ from the behaviour more generally observed in the full data set.

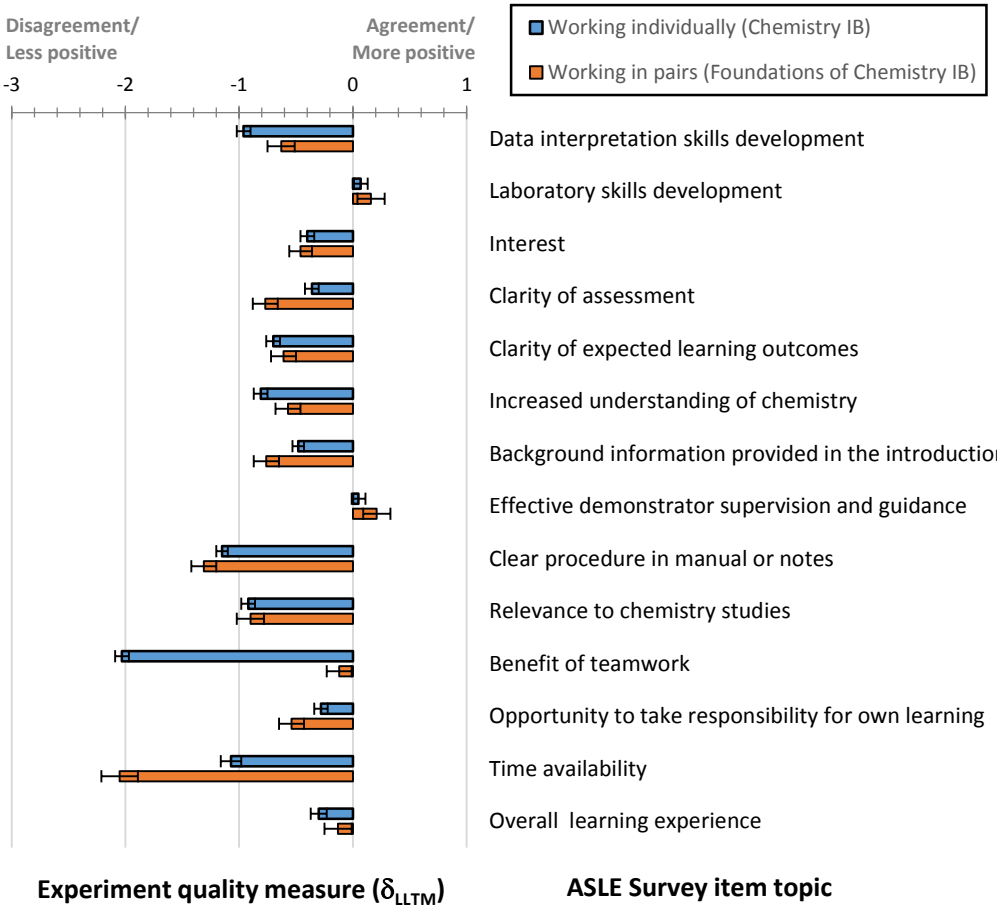


Figure 49: Measures of laboratory learning experience quality associated with different forms of the Analysis of Spinach extracts experiment

Here the experiment quality measures presented are the empirical values as would be estimated from a partial credit Rasch model, not the LLTM approximation (corrected by adding the "displacement" value). This is to ensure that observed values are contrasted rather than values simply predicted from the theoretical model. Error bars represent standard error values. Note that the position of the "zero" value is arbitrary. The student group "working individually" were Chemistry IB students, whilst the student group "working in pairs" were Foundations of Chemistry IB students (see section 2.1.1).

It is important to note that although the experiment conducted here is equivalent, different cohorts of students were involved for the two groups. Despite the fact that student bias measures have been controlled for, the possibility of differential item functioning between the two student cohorts still exists. That is, all other factors being equal, objective measures of

learning experience quality (δ) may be the same for all students within a single student cohort, but differ between the two cohorts. Possibilities such as this confound any conclusions drawn from the comparison displayed in Figure 49. This could explain why time availability appears to be less of an issue for the students working individually in this case: that group is also the group of students with past experience in chemistry, who may objectively execute procedures more quickly as a general rule, because of their past experience. Likewise, the experiment's learning objectives appear slightly less discernible for the "in pairs / Foundations of chemistry" group, possibly for similar reasons. The "individual workers / Chemistry IB" group may also have had lower perceived development of laboratory skills simply because they already had the required skills through previous chemistry experience, unlike the Foundations cohort.

A valid comparison of working in pairs as opposed to working individually cannot yet be made here. This observation raises a critical issue: if the quality of an experiment can objectively differ depending on the capability of the student cohort, then so too must the Q-matrix. In real world terms, this means that the relative importance factors determining the quality of an experiment differ based on the student group to which the experiment is presented. As currently presented, the LLTM does not incorporate potential differences in the factors contributing to a positive student learning experience for the two different student cohorts: "Foundations of Chemistry IA/B" and "Chemistry IA/B". Inherent differences in the appropriate pedagogy applicable to each student group therefore confound the Q-matrix weighting coefficients estimated.

Differences in the design of the experiments conducted which, by chance, happen to correlate with differences in the student cohort conducting those experiments may alter the correlations observed in the factor analysis, hence also the factors extracted and the associated Q-matrix. For example, it was previously suggested that the (often) more experienced "Chemistry IA/B" cohort may be more likely to finish their experiments quickly than the (often) inexperienced "Foundations of Chemistry IA/B" cohort. Experiments where students work individually are in many cases the experiments exclusively conducted by the Chemistry IA/IB cohort, suggesting that this could be one reason why the perceived benefit of teamwork appears to be associated with increased time availability more generally in the final LLTM; it may be the cohort difference which impacts time availability rather than teamwork, but the two just happen to correlate. Likewise, experiments with greater theory focus were noted to often be the experiments conducted by the Foundations of Chemistry cohort, possibly helping to explain why $\eta_{theory\ focus}$ appears to control a large proportion of the variance observed in experiment quality measures estimated for the ASLE survey items. The fact that experiments conducted individually were almost exclusively conducted by the more experienced Chemistry IA/B cohort could also explain why teamwork showed such a strong association with the increase in understanding. The actual extent to which to cohort difference confounds the results of this investigation is unknown. Rectifying this issue would require estimating a separate Q-matrix for each student group; something not possible with this data set due to the low number of experiments investigated.

The fit statistics associated with some LLTM measure estimates appear to support the possibility of different factor models applying to different student groups. Many $\eta_{theory\ focus}$ measures which appear contrary to intuition can often be attributed to misfitting (specifically underfitting) the model generated. In cases such as these, reported student perceptions appear not to be modelled well by the Q matrix and Rasch model proposed here, suggesting that factors contributing to the observed perceptions are either weighted differently or are

different in nature to those modelled. Table 26 provides a summary of fit statistics associated with each measure for the $\eta_{theory\ focus}$ factor.

Table 26: Fit statistics for $\eta_{theory\ focus}$ measures

#	Experiment	Data points	Infit		Outfit		Estim. Discrm	Correlations	
			MnSq	ZStd	MnSq	ZStd		PtMea	PtExp
3	Vapour Pressure	5141	1.17	7.5	1.54	9.0	0.68	0.72	0.74
10	Coloured Complexes of Iron	32971	1.04	3.8	1.12	9.0	0.94	0.74	0.75
13	Copper(II) Ion Absorption Spectrophotometry	12750	0.98	-1.2	1.07	4.1	1.00	0.62	0.63
16	Equilibrium and Le Chatalier's Principle (revised: for foundations)	17902	1.00	-0.3	1.07	4.4	0.96	0.75	0.76
8	Liquid-Liquid Extraction and TLC	22860	0.92	-7.6	0.92	-6.0	1.08	0.72	0.71
9	Synthesis of Aspirin	15637	0.96	-3.1	1.02	1.5	0.99	0.75	0.76
11	Analysis of Spinach Extracts	26703	1.02	2.4	1.11	9.0	0.97	0.59	0.61
17	Visible Absorption Spectrophotometry	23596	1.07	6.3	1.16	9.0	0.94	0.56	0.60
26	Vapour Pressure (revised: laptop)	24331	0.98	-1.5	1.07	5.3	1.00	0.70	0.71
12	Ion exchange Chromatography	46824	1.07	9.0	1.11	9.0	0.94	0.63	0.66
19	Analysis of Spinach Extracts (revised: for foundation - in pairs)	19020	0.83	-9.0	0.89	-7.9	1.14	0.71	0.69
1	Biological Buffers	12322	0.89	-7.3	0.91	-4.9	1.08	0.71	0.72
4	Melting Points and Recrystallisation	29207	0.96	-4.1	0.99	-1.0	1.03	0.67	0.67
22	Reaction Kinetics (revised: for foundations)	4561	0.81	-8.3	0.85	-5.3	1.06	0.74	0.74
27	Copper(II) Ion Absorption Spectrophotometry (revised: laptop)	33347	0.97	-2.8	1.06	5.2	0.98	0.62	0.65
15	Determination of Vitamin C concentration	16109	0.98	-1.4	1.04	2.6	1.01	0.76	0.76
29	Quantitative Techniques (revised: no pipetting)	10914	0.98	-1.5	1.04	2.3	1.02	0.69	0.69
7	Introductory experiment	1149	1.14	3.1	1.20	4.4	0.81	0.66	0.69
21	Thermochemistry (revised: for foundations)	17253	1.11	8.3	1.32	9.0	0.85	0.64	0.68
25	Biological Buffers (revised: laptop)	15165	0.99	-0.5	1.03	1.8	1.02	0.68	0.69
5	Quantitative techniques	40666	1.12	9.0	1.16	9.0	0.90	0.71	0.72
6	Reaction Kinetics	32704	1.03	3.5	1.07	6.5	0.96	0.64	0.67
20	Metal Activity Series	14348	0.89	-8.2	0.92	-5.0	1.07	0.70	0.70
28	Introductory Experiment (revised: pipetting)	8822	1.19	9.0	1.33	9.0	0.81	0.49	0.54
14	Quantitative Techniques (revised: for foundations)	9903	1.05	3.0	1.12	6.1	0.92	0.73	0.75
18	Aromachemistry	21530	1.23	9.0	1.41	9.0	0.71	0.51	0.57
24	Reaction Kinetics (revised: question order and phrasing)	4984	0.99	-0.5	1.08	2.5	0.99	0.70	0.70
2	Thermochemistry	43480	1.23	9.0	1.44	9.0	0.70	0.56	0.62
23	Introductory experiment (revised: observations video)	7797	1.22	9.0	1.45	9.0	0.68	0.55	0.63

Numbers (#) associated with experiments are equivalent to those listed in Table S 56 (section 7.4.2) following previously described equating procedures. Cases shaded in orange exhibit statistically significant underfit of a large magnitude. Experiments have been ordered from lowest measure to highest measure, as displayed in Figure 48. Elaborations on the statistics quoted and their interpretations are presented in section 2.5.2).

Many of the 29 experiments appear to underfit the model significantly (ZStd > 2, see section 2.5.2.1). It is important to note, however, that statistical significance of the misfit does not reflect the magnitude of the misfit; many experiments have several thousand data points associated, leading to significance of even very small misfit to the model. Only seven experiments show underfit of any substantial magnitude (mean square values above approximately 1.2, see section 2.5.2.1). Of the experiments which do show significant misfit, many are different variants of the same experiments. "Thermochemistry" and "Introductory experiment" each have multiple different variants, yet consistently show some misfit to the model in each case. This lends some assurance to the idea that the observed misfit is characteristic of the experiment design itself rather than being chance variation within the data set.

Using the mean square values as an indicator of the magnitude of the misfit, outlying data points appear to deviate from the model more than inlying data points in all cases, suggesting that most student perceptions for survey items relevant to $\eta_{theory\ focus}$ are well explained by the model, but extreme perceptions are explained poorly. This is additionally observable in the fact that misfitting experiments (shaded in Table 26) are at the edges of the distribution of measures rather than being closer to the average. Differential item function may exist here; the objective quality of experiments may differ depending on the specific student group performing the evaluation, for reasons independent of the students' broad scale tendency to provide positive response on the survey. The "Thermochemistry" experiment, for example, may have an objectively less clear connection to the lecture content for students of a specific learning style (or of other particular characteristics). In such a case, those specific students would then respond in a manner substantially different to the bulk of the student population, as the factors determining learning experience quality for them differ from the factors determining learning experience quality for other students. These students' response patterns would then misfit the LLTM, manifesting as the high outfit values observed (since their perceptions would be outlying rather than inlying).

This issue highlights a limitation of analysing large scale datasets of ASLE survey responses: the student population is presumed to be "homogeneous". Correlating responses, either using Rasch measures such as in this study or using scores as previous research has done, reveal conclusions only about the average behaviour of the bulk student sample. The nuances which arise when considering how to appeal to a specific cohort of students are lost when observing broad scale correlations or average responses. Conclusions drawn from large scale studies of ASLE survey data about *the* definitive factors contributing to a positive laboratory experience may therefore encourage "teaching to the centre", whilst neglecting students with atypical learning requirements or capabilities. This raises consequential validity issues for the use of ASLE survey data; something to be explored further in later discussion.

4.4.5 Supporting laboratory skills development through data interpretation

A means of improving the vast majority of aspects of the laboratory learning experience targeted by the ASLE survey appears to be the effective development of data interpretation skills within practicals. The only survey items not impacted by $\eta_{data\ interpretation}$ are items 4, 6 and 12, concerning clarity of assessment criteria, increased understanding and responsibility for own learning respectively. Increasing $\eta_{data\ interpretation}$ does also appear to have a large negative impact on perceived time availability (Q-matrix weighting of -8), likely because extra data interpretation requires more time for the analysis. However, all other items see at least a

small improvement as $\eta_{data\ interpretation}$ is increased, with substantial gains seen in the perceived development of laboratory skills (Q-matrix weighting of +10, see Equation 49 below). This suggests that the ability to interpret the data gained from executing practical procedures may improve the ability to execute the procedure itself, at least from the perception of the learner. Intuitively this is sensible, given that understanding the data to be gained from a procedure (and hence how to interpret them), helps in executing the procedure with an understanding in mind as opposed to 'blind' recipe following. Equation 49 shows the composition of the data interpretation skills item measure, as modelled in the LLTM (discounting item 2 relative location μ_2).

$$\delta_{2\ (laboratory\ skills\ development)} = \begin{bmatrix} -15 \\ 1 \\ -1 \\ 10 \\ 7 \\ -2 \\ 2 \\ 0 \\ -1 \\ -3 \end{bmatrix} \cdot \begin{bmatrix} \eta_{theory\ focus} \\ \eta_{instructions} \\ \eta_{collaborative\ \ understanding} \\ \eta_{data\ interpretation} \\ \eta_{independent\ learning} \\ \eta_{demonstrators} \\ \eta_{unexplained\ overall} \\ \eta_{8} \\ \eta_{9} \\ \eta_{10} \end{bmatrix} \quad \textbf{49}$$

As can be seen, $\eta_{data\ interpretation}$ appears as a prominent contributor to the perception of laboratory skills development. The two other most prominent contributors are $\eta_{theory\ focus}$ and $\eta_{independent\ learning}$, both of which could theoretically be manifestations of cohort difference issues already noted. Chemistry IA/B experiments lacked lecture synchronisation (low $\eta_{theory\ focus}$) and were often the only experiments to be conducted individually (high $\eta_{independent\ learning}$), meaning a cohort difference in the perceived development of laboratory skills would be expected to manifest as heavy weightings in these two factors. However, the contribution of these two factors appears to be opposite to what would be expected, were this the case: Chemistry IA/B students have increased prior experience and therefore would develop fewer new laboratory skills, meaning the factor weightings observed would be positive for $\eta_{theory\ focus}$ and negative for $\eta_{independent\ learning}$respectively under this interpretation. If the observed heavy weightings are indeed a result of cohort differences, it appears the more experienced cohort generally perceives more skill development, not less. This is in direct contrast with reasonable expectation, suggesting that the large weightings of these factors cannot be dismissed merely as cohort differences confounding the data. Independent work and a low focus on lecture content can therefore reasonably be concluded to support the development of laboratory skills generally.

A cohort difference may, however, contribute in other ways. The measures for $\eta_{data\ interpretation}$, which also strongly contributes to laboratory skills development, appear to be dependent on students' ability to interpret the data as required. Measures also appear to be dependent on the *improvement* of data interpretation skills, which could feasibly differ depending on the skills students have to begin with. The reaction kinetics experiment exemplifies this fact well: as initially presented, "Reaction kinetics" shows a $\eta_{data\ interpretation}$ measure of -0.32 ± 0.01 Logits. When revised and presented to the Foundations of chemistry cohort, $\eta_{data\ interpretation}$ improved slightly to -0.27 ± 0.03 Logits, suggesting the Foundations cohort's data interpretation skills were developed more than for the Chemistry IA cohort, using this same practical exercise. This could sensibly be attributed to differences in past

chemistry experience. Anecdotally, some students from Chemistry IA noted that they had conducted the essentially identical experiment in high school; something which would decrease any data interpretation skills gained from the experiment. This was typically not an issue for the Foundations cohort.

Further, when initially modified to suit the Foundations of chemistry cohort, the questions asked of the students in "Reaction kinetics" were mistakenly re-ordered, changing the way students were guided through the data analysis process. This change is observable as a decrease in quality of the material provided ($\eta_{instructions}$ shifted from 0.50 ± 0.01 down to 0.44 ± 0.03 Logits). When this was noticed and amended such that the questions were asked in their original, more intuitive order, the increased quality of the provided material ($\eta_{instructions}$ = 0.56 ± 0.03) allowed a more easily followed interpretation of the data, seen in an improved $\eta_{data\ interpretation}$ measure from the original -0.27 ± 0.03 Logits up to -0.19 ± 0.03 logits. Again, these observations illustrate that measures for $\eta_{data\ interpretation}$ appear to be related to student ability to interpret the data as required, not just the extent to which it is required of them. Quality of data interpretation activities appears to outweigh quantity here.

The role of the $\eta_{data\ interpretation}$ factor as a measure of students' ability to interpret their data, rather than merely be required to do so, is further corroborated using measures for experiments previously studied in depth. Section 3.1: "Typical score-based analysis of ASLE survey data: an example" discussed the effects of replacing a handheld data logger with a more likely intuitive laptop interface for the purpose of data analysis. Improving the means by which data was gathered and viewed by students would intuitively alter the students' development of data interpretation skills as a result of the practical, meaning this study presents an opportunity to examine the behaviour and validity of the $\eta_{data\ interpretation}$ factor. All three of the experiments discussed in section 3.1, which were amended to be conducted using a laptop rather than a handheld data logger, show similarly improved $\eta_{data\ interpretation}$ values with the laptop interface (Table 27). This further validates conclusions presented earlier about the laptop interface being superior to the data logger interface: all three experiments show a greater perceived development of data interpretation skills with the laptop interface, presumably because the laptop interface (where data is gathered and analysed) is more often intuitive and more easily understood.

Table 27: Data interpretation skills development measures for different technological interfaces

Experiment	$\eta_{data\ interpretation}$ /(Logits)	
	Data logger interface	Laptop interface
Vapour pressure	-0.45 ± 0.03	-0.28 ± 0.02
Biological buffers	-0.36 ± 0.02	-0.26 ± 0.02
Copper (II) ion absorption spectrophotometry	-0.32 ± 0.02	-0.30 ± 0.01
Visible absorption spectrophotometry		-0.26 ± 0.02

The first three experiments presented in the table above were previously studied in depth, described in section 3.1. "Visible absorption spectrophotometry" was not discussed at length previously, but is a version of "Copper (II) ion absorption spectrophotometry" presented exclusively to the Foundations of Chemistry cohort.

Perhaps even more encouraging, however, is the high degree of similarity in the measures obtained for each experiment. Values for $\eta_{data\ interpretation}$ using the laptop interface are highly consistent between all relevant experiments, each with a measure of approximately

0.28 Logits. This is a substantial validation of the qualitative meaning of the $\eta_{data\ interpretation}$ factor: factor measures take on the same value when students interact with their data in the same way, suggesting a direct connection between the $\eta_{data\ interpretation}$ measure value and real world concepts pertaining to data interpretation. Values associated with the handheld data logger vary to some extent, though again appear comparable for two of the three experiments listed. This variation may be due to the different purposes for which the technology was used in each case.

4.4.6 High quality written material is broadly beneficial

During the initial discussion of experiments in which the technological interface was changed, it was suggested that a change in the instructional material for the three experiments may have been a contributor to the reason for the observed improvements, rather than the change in technological interface (see section 3.1.4.5). This concern was dismissed, given little evidence existed to suggest the new instructions were any better than the originals, and if anything may have been of lesser quality given the qualitative comments received. The LLTM offers the possibility of quantifying the quality of the provided material, using $\eta_{instructions}$ measures. This single factor encompasses all ASLE survey items concerning materials provided to the students, including background information (item 7), experimental procedure (item 9) assessment criteria (item 4) and to some extent the intended learning outcomes (item 5).

Table 28: Changes in quality of the provided material when using different technological interfaces

Experiment	$\eta_{instructions}$ /(Logits)	
	Data logger interface	Laptop interface
Vapour pressure	0.48 ± 0.02	0.47 ± 0.01
Biological buffers	0.44 ± 0.02	0.42 ± 0.02
Copper (II) ion absorption spectrophotometry	0.52 ± 0.02	0.55 ± 0.01

As can be seen, improvement for any the three experiments is minimal at best (Table 28). Error margins in $\eta_{instructions}$ measures for the two forms of each experiment overlap except in the case of the Copper (II) ion absorption spectrophotometry experiment, which shows some small improvement in the instructional material once the laptop was introduced. Again, the fact that any improvement observed is minimal matches the prediction which would be made from the fact the instructional notes were only changed minimally: the only section altered in each case was a small appendix containing instructions for the technology. The suggestion that a change in quality of the instructions was unlikely to be the cause of the observed improvements appears sound.

Had the $\eta_{instructions}$ measures been substantially different, however, this could have had large scale impacts on the responses observed. The $\eta_{instructions}$ factor is the second most prominent factor in the LLTM, and appears as one of the equal largest contributors to positive responses to the question “I found this to be an interesting experiment”, with a Q-matrix weighting of +5 (the other equally strong contributor being $\eta_{collaborative\ understanding}$). The $\eta_{instructions}$ factor also appears as one of the strongest identifiable contributors to the “overall learning experience” item, with a Q-matrix weighting of +2. At least small improvements to every item on the ASLE survey appear likely with greater quality of the provided materials (reflected in a high $\eta_{instructions}$ measure), only with the exceptions of item 1 (development of data interpretation skills) which appears unaffected by the $\eta_{instructions}$

factor, as well as item 11 (concerning the benefits of teamwork). Item 11 "working in a team to complete this experiment was beneficial" appears as the only item whose responses are negatively impacted by $\eta_{instructions}$, with a Q-matrix weighting of -4. This seems to suggest that a higher quality of instructional notes for an experiment leads students to see less benefit in teamwork, perhaps explained by a lowered need for assistance by their peers. This may not necessarily be a bad thing, given that it may indicate students can grasp the intended messages within the notes more easily without assistance.

Again, however, the effect of student cohort differences can be seen in this factor. The identical instructional notes can evidently be objectively better for one student group than another. The Quantitative techniques experiment is a revealing example: when presented to the Chemistry IA cohort, "Quantitative Techniques" has a $\eta_{instructions}$ measure of 0.55 ± 0.01 Logits, whereas when the same experiment was presented to the Foundations of chemistry cohort, the perceived quality of the provided material dropped to $\eta_{instructions}$ = 0.44 ± 0.02 Logits. A similar value was seen even after the experiment was further amended for the Foundations of chemistry cohort, removing the pipetting section of the activity ($\eta_{instructions}$ = 0.46 ± 0.02). Once again, however, the fact that the same notes can be more useful to one student cohort than for another is intuitively sensible. Different student cohorts likely have different learning styles, cognitive abilities, background knowledge and an array of other differing predispositions, meaning the same set of notes may not be read, interpreted or understood equivalently. Pitching the written material at a level appropriate to the abilities of the reader is naturally advisable, and this is seen in the behaviours of $\eta_{instructions}$ as described.

4.4.7 Engaging the students: interest and positive overall experience

One of the major goals of the ASELL project and its predecessors has been to improve the student experience of their laboratory activities, promoting interest and a positive overall learning experience. The LLTM now reveals how to achieve this, based on patterns present in the data set analysed. A wide array of factors appear to contribute to student interest in a laboratory activity, as seen in Equation 50 (discounting item 3 relative location μ_3).

$$\delta_{3\,(interest)} = \begin{bmatrix} -10 \\ 5 \\ 5 \\ 1 \\ 1 \\ -3 \\ 3 \\ 1 \\ 2 \\ 1 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} \eta_{theory\ focus} \\ \eta_{instructions} \\ \eta_{collaborative\ \ understanding} \\ \eta_{data\ interpretation} \\ \eta_{independent\ learning} \\ \eta_{demonstrators} \\ \eta_{unexplained\ overall} \\ \eta_{8} \\ \eta_{9} \\ \eta_{10} \\ \eta_{11} \end{bmatrix} \quad \textbf{50}$$

By far the strongest contributor to student interest in the activity is a decreased connection to lecture content. Naturally this is undesirable for many laboratory activities from the perspective of the teacher, since laboratory activities are often intended to strengthen content knowledge through concrete interactions. Student preference naturally appears to lie with "skills-based" rather than "theory-based" experiments.

As it was previously discussed, however, low $\eta_{theory\ focus}$ measure does not necessarily imply more concrete interaction with the material. Heavily concrete practicals can occasionally be observed to have high $\eta_{theory\ focus}$ measures, and this was previously explained by suggesting that a clear connection to current theory from the perspective of the students was the key consideration (see section 4.4.2). This raises an unusual situation: it is not necessarily the inclusion of theory students are opposed to, but the fact that they can recognise the theory as relevant to the course. Theory recognisably connected to the lecture content appears to drastically decrease student interest, where it apparently may not if it were absent from lectures. Reasons for this effect are speculative at best, but this may simply be a "knee-jerk" reaction of boredom to any theory which has been laboured upon as part of the course.

The only possible remedy to this problem appears to be to "disguise" the theory in any practical which also appears in the lecture content. Lecture content to be reinforced by the practical activity should, by this logic, be addressed differently to the way it is in other components of the course. In this way it may not be quite as recognisable to students, ideally reducing this "boredom" response. Alternately, practicals could be used as a means of introducing the lecture theory rather than reinforcing it. This would have the effect that students still gain the educational benefits of the practical relevant to the theory, but without the risk of students losing interest because of the lecture content connection. Students would, ideally, not be able to be "bored" by the material if they had not been extensively exposed to it yet. Using practical activity as introductory rather than to reinforce content already presented also follows recommendations in the education literature, keeping the laboratory as the place to explore the "macro" aspects of chemistry,[4, 13] then shifting from the concrete to the abstract as recommended by Johnstone[257-260, 304, 306] and in line with cognitive development described by Piaget.[261, 307, 308, 327]

Other strong contributors to interest in the activity appear to be well constructed written material (high $\eta_{instructions}$) and understanding through teamwork (high $\eta_{collaborative\ understanding}$). Interestingly, this creates another problem for educators: understanding through teamwork was previously seen to be beneficial for highly theory focussed practicals, whilst more skills-based practicals, conducive to student interest, were seen to benefit from individual work. The decision whether students work in teams or not therefore presents a dilemma to the teacher when running a skills-based practical: working in teams makes the experiment more interesting, but working individually promotes laboratory skills development. Strengthening the learning objectives appears to be at odds with maintaining student interest.

A more moderate contribution to student interest is to reduce perceptions of effective demonstrator guidance. This seems unintuitive if the demonstrator guidance item of the ASLE survey (item 8) is taken to reflect the teaching ability of the demonstrator, however, correct interpretation of the $\eta_{demonstrators}$ factor is problematic. The factor measures can not possibly reflect the quality of specific demonstrators, since the specific demonstrator was not constant for each practical. Rather, the demonstrator was specific to the practical group students were assigned to: each experiment was conducted by a range of student groups in each semester, each group with a potentially different demonstrator. At best, values for this factor could be attributed to the "average" quality of demonstrators for each given practical. However, there is reason to suggest that values for this factor may also depend on students' *reliance* on their demonstrators, and hence greater appreciation of their assistance.

Before "Vapour Pressure" was amended to utilise the laptop as opposed to the handheld data logger, the value of $\eta_{demonstrators}$ was far higher than for any other experiment at a value of 0.92 ± 0.03 Logits. For the entire suite of 28 other experiments studied, $\eta_{demonstrators}$ only ranges between values of 0.29 ± 0.02 and 0.67 ± 0.02 Logits (for the revised "Vapour pressure" using the laptop and "Determination of Vitamin C concentration" respectively). A conclusion that quality of the demonstrators' teaching abilities drastically rose for one specific experiment seems unlikely, particularly since their teaching abilities would then have to be interpreted as returning to normal again once the identical experiment used a different technological interface. The demonstrator measure seems to reflect something about the design of the activity, not the demonstrators themselves. A likely candidate for correct interpretation of $\eta_{demonstrators}$ is that it measures students' broad scale appreciation for the fact they have a demonstrator there to help and guide them: the extent to which students require and appreciate help, not the quality of the demonstrator irrespective of how needed they were. Under this interpretation, it makes sense that student interest would be improved under low $\eta_{demonstrators}$ measures: it simply means that experiments are more interesting if students do not need to rely on their demonstrators to complete or understand the exercise.

The role of $\eta_{demonstrators}$ seems to be opposite for the overall learning experience, however. Increasing reliance on demonstrators may decrease student interest in the activity, but it promotes a better overall perception of the activity as a whole. Equation 47, which has been previously presented, is replicated below to show the LLTM model for the overall learning experience item of the ASLE survey.

$$\delta_{14\,(overall\ learning\ experience)} = \begin{bmatrix} -2 \\ 2 \\ 0 \\ 1 \\ 1 \\ 2 \\ 5 \end{bmatrix} \cdot \begin{bmatrix} \eta_{theory\ focus} \\ \eta_{instructions} \\ \eta_{collaborative\ \ understanding} \\ \eta_{data\ interpretation} \\ \eta_{independent\ learning} \\ \eta_{demonstrators} \\ \eta_{unexplained\ overall} \end{bmatrix} \quad \textbf{47}$$

Again it is seen that students show a preference for less lecture theory connection in their practical activities, indicated by the -2 weighting coefficient for $\eta_{theory\ focus}$. This factor, as well as assistance by the demonstrators and high quality instructional material appear to carry equal weight in generating a positive overall learning experience. Independent learning also appears to play a small role, leaving understanding of the theory through groupwork with no discernible contribution. This does not mean group work need not be included in practical activities – it was seen to be highly beneficial for understanding of chemistry. However, it does suggest students have a slight preference towards individual work. Given student preferences for "skills-based" practicals discussed at length previously, this is perhaps unsurprising. It appears that constructive alignment[328, 329] plays a role here: the laboratory activity needs to be designed to suit the key learning objectives, either increased theoretical understanding or increased technical skills. Skills oriented practicals are likely to be perceived well overall regardless, but attainment of the learning objectives is likely to be improved using individual work and the inclusion of some data interpretation (see section 4.4.5). However, theory oriented practicals are unlikely to be perceived well unless connection to lecture content is well hidden (section 4.4.3) and understanding is most liable to be increased through group work, in contrast to the recommendation for skills development.

More concerning, however, is the fact that $\eta_{collaborative\ understanding}$ has no discernible weighting on the overall learning experience. Given that this factor's primary defining characteristic is the perception of increased understanding of chemistry, the question arises as to whether students are generally ignoring the word "learning" in the phrase "overall learning experience". Certainly ASLE survey researchers would typically interpret responses to this survey item as incorporating, at least in part, some learning gained from the exercise. The observations made here suggest this may be in error, however. A serious issue is now apparent for the consequential validity of the ASLE survey: if experiments are structured such that the "overall learning experience" item receives the most positive response possible, this has evidently no connection to increased student understanding. The pursuit of optimal perceptions by the students (pursued using ASLE surveys) needs to be appropriately balanced with the educational goals of the course. If students were to receive their preferences, judging by Equation 47, experiments would be entirely disconnected from the lecture content, students would not have to develop group work skills, demonstrators would guide students through any thinking processes required and whether any understanding was gained would be considered irrelevant. This difference between student preference and teacher intentions may explain the evident disconnects between staff expectations of what would promote a positive overall learning experience and the results observed in student feedback.[51, 57]

This is by no means a full explanation of students' "overall learning experience" rating, however. The largest contribution of all to this survey item's response is from the seventh factor, interpreted only as "overall learning experience" unable to be explained as connected with the other 13 items of the ASLE survey. The vast majority of variance in "overall learning experience" appears to occur for reasons unknown. It may be that these reasons are connected to attainment of learning objectives, but it also may not. It could also be that this seventh factor has unusual "corrective" behaviour as factor 8 was previously seen to exhibit (see section 4.3.3.4). The identity of this factor, and hence the identity of the current unidentified sources of variance in perceived overall learning experience, remains as a goal for future research.

4.4.8 Conclusion

The ASLE data LLTM reveals a wide range of insights into the laboratory experience from the student perspective. Trends observed in student preferences appear to suggest that individual work and more skills-based practical activities are preferred by students. Activities like these appear to benefit from the inclusion of data interpretation and a lack of clear connection to lecture theory. Conversely, more theory oriented experiments appear to be best structured as group work activities to foster the understanding of chemistry, in spite of student preferences. There is also evidence to suggest that activities should be structured appropriate to the ability level of the student cohort and that different subsets of the student population may benefit from different approaches. The engagement of students as judged by ASLE surveys appears to be at odds with teaching and learning goals and many ASLE items' correct interpretation appears counterintuitive, suggesting that using ASLE survey responses alone to optimise experiment design may be problematic.

5 *Conclusions and future opportunities*

5.1 How ASLE survey data should be analysed

5.1.1 Use of integer scoring methodology

The first primary hypothesis of this thesis, reproduced below, concerned the validity of quantitative techniques typically applied to ASLE survey response data. Given the observations made over the course of this research, appropriateness of the usual integer value scoring system and hence reliability of past conclusions can now be evaluated.

> ***Hypothesis 1:***
> *Conclusions drawn from the ASLE survey data using typical scoring techniques resemble conclusions drawn using sample independent, interval scale measures extracted from the same data.*

The integer scoring system typically used to analyse ASLE survey data is generally a reasonable approach for most practical purposes. Despite concerns regarding this methodology raised in statistics literature, integer value scores appear not to have any discernible advantage over more 'optimised' score values for the response categories (Figure 30, p. 93), and calculation of mean scores, standard deviations and parametric statistics is not inherently inappropriate (Figure 31, p. 94). This observation of a specific case where scoring or ordered categorical data for the purposes of using parametric statistical methods is relevant to survey-based research in a wide array of disciplines where these controversial techniques are common, including medicine, psychology and education. The problem with using integer score values, in this case, lies far more with their interpretation.

The response categories available for any given item of the ASLE survey are not equidistant (, Figure 27, p. 90), and the range of student perceptions gathered by each response category of the rating scale differs from item to item (Figure 28, p. 91). This not only removes the ability to gauge whether student perceptions are broadly positive or negative based on the sign of the mean score value, but also the ability to compare different items of the survey. Additionally, magnitude of any differences in mean ASLE score values observed cannot be treated as if they are proportional in magnitude to any difference in student perceptions. The fact that any linear relationship between the ASLE mean score and student perception is only approximate (Figure 29, p. 92) means that whilst researchers may be able to determine that an improvement has occurred, they cannot precisely quantify the change based purely on score data. It is also not necessarily valid to claim no change has occurred simply because the ASLE mean scores have failed to change substantially (Figure S 48, p259).

The problem is compounded by the fact that student predisposition towards positive response changes from occasion to occasion, even if the identical students are surveyed (Figure 32, p. 102). These changes in student bias can generate significant mean score differences even if the experiment is unchanged, or exaggerate or mask genuine differences (see section 3.2, pp. 61-83 for multiple examples). Variation between individual students in a given sample (within sample variance) may be controlled for by gathering larger student samples, but variation across independently gathered samples of students (between sample variance) (Figure 33, p. 105) may remain independent of individual sample sizes (Figure 44, p. 128). This means that an inherent margin of error margin of approximately 0.1 score units on average (Figure 45, p. 129) is always expected to exist in any ASLE mean scores obtained using data gathered on a single occasion, in addition to random errors introduced through small sample sizes. Even large

mean score differences ($\cong 0.7$) may be the result of students evaluating an experiment unfairly (Figure 43, p. 127) rather than a genuine difference in experiment quality.

The resulting implication for researchers is that small differences in ASLE scored results (< 0.1) should be dismissed as expected in all cases, even if the differences appear statistically significant. Larger differences, however, cannot be known to reflect experiment quality if using scored data alone, and hence should always be corroborated using open response comments received (see section 3.1, pp. 51-59 for an example). As a rough guideline, differences ≥ 0.25 mean score units are likely to be genuine at sufficient sample sizes (Figure 45, p. 129) and this could be used as a "rule of thumb" for most rough, exploratory studies using ASELL survey data. However, the use of mean scores for more precise and quantitative investigations is more questionable.

Variation in student predisposition toward positive response is by far the primary source of variation in individual ASLE survey responses (see section 3.3.3.1, pp. 89-90), meaning one person's rating of "good" is not guaranteed to reflect better experiment quality than another person's rating of "poor". The same conclusion is supported by the broad overlap of likely categories of response to be observed for any given student perception (Figure 27, p. 90). This calls into question the appropriateness of using non-parametric rank-based methods such as the Wilcoxon rank sum test, often used to avoid the controversies associated with integer scoring. Further, the validity of correlating scored ASLE response data is threatened, since variance in individual responses is far more reflective of correlations within students' own biases than it is of correlations between perceived aspects of the learning experience (section 4.1.6.3, p. 108 and section 4.3.4.3, p. 139). Averaging the scores associated with many student responses before correlating the data may make scores reflect experiment-specific qualities more closely, but again the issues of a non-linear relationship between mean score and the underlying perception remain (see section 3.3.4, pp. 95-96). Correlating scored ASLE survey data should therefore be avoided, and any conclusions drawn from such studies should be revised using alternate analysis.

The heavy conflation between student dependent and student independent factors, as well as non-linear relationship between mean scores and true interval scale measures of the analogous student perception make Hypothesis 1 above strictly false. This does not imply that scored data cannot be used; they are far simpler to compute than Rasch measures and therefore much more practical for most educators. Limitations in the scoring methodology simply restrict score-based research to more exploratory purposes, which may then be supported further by additional research (for example from qualitative comments received on open response items of the survey). Quantitative analysis using score data is inappropriate, but support (section 4.1.3, pp. 101-104) for the second primary hypothesis of this research:

> ***Hypothesis 2:***
> *Student independent contributions to the ASLE survey responses exist and are measurable.*

implies that mean scores may act as indicators of more generalizable "objective" laboratory learning experience quality, and are useful for that reason. Researchers simply need to be conscious that these measurable properties of the experiment evaluated contribute minimally to the variance in individual ASLE responses, necessitating larger samples of data for meaningful investigations.

5.1.2 Interpretation of ASLE survey results

Hypothesis 2 above appears true for most survey items, but not all. The poor construct validity of items 11, 12 and 13 (concerning benefits of teamwork, responsibility for own learning and time availability respectively) may imply a lack of generalisability of these qualities beyond the perspective of individuals (Table 19, p. 106). It may not be valid, for example, to make a general claim that one experiment is perceived as having more time availability than another: perceived time availability may vary between individuals far too widely for any group level measurements to be meaningful. Alternately, however, these three items may interact with student predispositions differently to the other items of the survey. Separating these two possible conclusions from one-another requires further investigation.

Reason exists to suspect that even the measures of experiment quality which are more generalizable depend on the student cohort to which the activity is presented. That is, even for items with good construct validity, many factors underpinning ASLE survey responses are likely dependent on the abilities of the student cohort (section 4.4.4, pp. 148-152). The conclusion to be drawn from this is not that assigning quantitative measures to ASLE survey items is inappropriate, but that measures of experiment quality must be viewed as dependent on the student audience. An acknowledgement should also be made that any conclusions drawn from group-level statistics apply only to the "bulk" of the student sample, but may not apply to extreme high or low achievers, or to students otherwise atypical to the average. The *unique* best way to design experiments, or the *definitive* most prominent factors in a positive learning experience may not exist: these things may change entirely for students of different learning styles, different levels of prior experience or possibly even different cultural backgrounds. Even the placement of the activity in the broader course context may alter measures of experiment quality (section 4.4.2, pp. 143-146), limiting the generalisability of any ASLE survey results for individual experiments. This emphasises the importance of detailing the context and audience for any experiments included in the ASELL database, and for any experiments submitted to ASELL workshops for evaluation. Connection to lecture content appears as one of the most prominent factors in student perceptions of their experiment (Table 21, p. 130 and Table 23, p. 132) meaning this should be a prominent consideration in experiment design and evaluation.

Correctly interpreting ASLE survey response data is problematic for reasons independent of independent of any issues with generalisation of the results or whether the integer scoring system is used, however. Over the course of the numerous investigations presented in this thesis, a number of validity issues have been noted with items on the ASLE survey, summarised in Table 29 (see below). Multiple survey items appear to be either best interpreted differently than intended, appear unable to be assigned any measure able to be generalised as true for most students, or would result in issues for student learning if their responses were optimised.

Table 29: Validity issues for items of the ASLE survey

#	Item	Noted validity issues (relevant section in parentheses)
1	This experiment helped me to develop my data interpretation skills	None identified
2	This experiment helped me to develop my laboratory skills	None identified
3	I found this to be an interesting experiment	Consequential validity issues: potentially counterproductive to learning (4.4.7; pp. 156-159)
4	It was clear to me how this laboratory exercise would be assessed	None identified
5	It was clear to me what I was expected to learn from completing this experiment	None identified
6	Completing this experiment has increased my understanding of chemistry	None identified
7	Sufficient background information, of an appropriate standard, is provided in the introduction	None identified
8	The demonstrators offered effective supervision and guidance	Potentially increased in problematic experiments due to students appreciating demonstrators "rescuing" them (4.4.7; pp. 156-159).
9	The experimental procedure was clearly explained in the lab manual or notes	None identified
10	I can see the relevance of this experiment to my chemistry studies	None identified
11	Working in a team to complete this experiment was beneficial	Poor construct validity (4.1.5; Table 19, p. 106). Receives a binary response reflecting whether students worked in pairs or individually (4.3.3.4; pp. 133-135).
12	The experiment provided me with the opportunity to take responsibility for my own learning	Poor construct validity (4.1.5; Table 19, p. 106).
13	I found the time available to complete this experiment was	Poor construct validity (4.1.5; Table 19, p. 106).
14	Overall, as a learning experience, I would rate this experiment as	Not a "summary" item as thought: majority of variance is explained by factors other than those in ASLE survey items. May not take "learning" into account as the question states. Consequential validity issues: counterproductive to learning (4.4.7; Equation 47, pp. 141, 156-159).

This once again highlights the need to take alternate data into account when using ASLE rating scale responses, to confirm that researcher interpretations of the evident issues are accurate. The third major hypothesis investigated in this thesis (below) has a complex answer: in most cases ASLE survey data does reflect measurable properties of the experiment evaluated, but those properties are not necessarily the ones the researcher might expect from the way the items are phrased.

> ***Hypothesis 3:***
> *Student independent measures obtained from ASLE survey data reflect qualities of the experiment evaluated.*

The more pressing issue for researchers using the ASLE survey is the fact that structuring experiments to optimise ASLE survey responses may threaten the educational value of the activity. Given that demonstrators may be rated more positively in problematic experiments, it may be inadvisable to strive for a positive demonstrator rating as judged by the ASLE survey. Similarly, the "boredom" response observed as a consequence of clear connections to lecture content and the irrelevance of increased understanding on reported overall "learning" experience highlight the risks associated with catering to student desires. A well-structured

learning experience, from the perspective of the teacher, necessarily appears to include many elements that students are liable to dislike. Educators therefore must remain conscious of this fact whenever ASLE survey responses are being utilised: whilst the surveys may help improve the student experience, they do not ensure teaching objectives are maintained. A balance needs to be struck between maintaining learning goals and appealing to the preferences of the learners.

The implication of this is that features of the ASELL review process other than the ASLE surveys are critical. Workshop sessions allow much needed feedback from other teachers about the educational merits of the activity, which students at home institutions evidently may be unconcerned with. Completion of the Educational Template document for experiments submitted to the ASELL review process is also a valuable and necessary step in ensuring that ASELL reviewed experiments have educational merit as well as the appeal from the student perspective desired. It would be highly inadvisable to conduct any ASLE survey research with the intent of improving experiment design without explicitly including a consideration of the educational goals of experiments, judged from the teacher perspective rather than the student perspective. Analysis of the open responses provided to item 16 of the survey: “What did you think was the main lesson to be learnt from the experiment?” may assist with this goal. The low number of responses typically gained for open response items in individual experiment analyses may be insufficient, however, re-emphasising the need for the wider ASELL review process incorporating the Educational Template document and feedback from educators at ASELL workshops.

5.1.3 Recommended research methodology

These results provide an opportunity to evaluate the appropriate “best practice” for using ASLE survey rating scale data. Suggested methodology for several different research purposes are summarised below.

For evaluating the merits of an individual experiment:

Assigning integer score values to response categories and calculating mean values is reasonable. However, it should not be assumed that the score assigned to the neutral category of the scale reflects a neutral perception and scores should not be compared between different items of the survey. Interpreting ASLE survey results as if they reflect the precise quality described in the exact phrasing of the survey item should be avoided (see Table 29) and any conclusions drawn from rating scale response data should be corroborated by further research or feedback provided on open response items of the survey. ASELL workshop feedback and data contained in the Educational Template document is critical for any review of experiment quality to ensure the educational aims of the experiment are not compromised by the appeal to student preferences. It should be recognised that results apply only to the bulk of the particular student sample surveyed, in the specific course context in which the experiment was presented.

For comparing perceptions of quality for two or more experiments

Assigning integer value scores to the rating scale categories is reasonable. Mean scores may be calculated for each item, treating these mean scores as rough indicators only. Statistical significance or otherwise of results is often uninformative, so little is gained using rigorous statistical testing, parametric or otherwise (though parametric tests may be preferred to rank-based tests). Small score differences (<0.1) should be dismissed as expected regardless of

sample size. As a rough guideline, score differences of approximately 0.25 units may usually be taken as significant when using sufficient sample sizes ($n \geq 30$). For more precise calculation of error margins in the mean score, Equation 46 (reproduced below) can be used to obtain the corrected standard error value at a particular sample size.

$$SE(A_{fair}) = \sqrt{\frac{\sigma_A^2}{n} + 0.01} \qquad \textbf{46}$$

For large score differences, any and all conclusions regarding the magnitude of the difference in perception should be avoided if based on the scores alone. Differences suggested by mean score data should be affirmed using analysis of qualitative response data in all cases, or through subsequent research. Concluding that an experiment is "improved" or "better" in a broad sense simply because student perception appears more positive is inadvisable: learning goals of the activity may still have been compromised, despite student approval. This should be checked. Student cohorts evaluating the experiments to be compared should be roughly similar, notably in terms of their level of background knowledge. It should also be acknowledged that some perceptions are dependent on the experiment's positioning in relation to the lecture content, so course sequencing differences may contribute to changes observed. It is possible that students may be caused to rate experiments highly unfairly in some extreme circumstances, and this should be taken into consideration when interpreting results.

For investigating broad trends in student perceptions across many experiments

Scored data is not suited to this purpose, especially for producing correlations. Rasch modelling or other methods which separate student dependent and independent factors are more appropriate for quantitative analysis in this case, modelling data to account for the fact student biases differ between occasions. In the case of Rasch modelling, data connectivity can be achieved by identifying students whose measures appear not to vary and ensuring they evaluate multiple experiments in the data set. Connectivity could alternately be achieved through models such as the LLTM, though fit to the model should be confirmed. Bulk population level statistics may be misleading, given that the relevant factors contributing to student perceptions may change depending on student prior knowledge, positioning of experiments in wider course contexts and other factors easily overlooked in bulk analyses.

In the case of interpreting past analyses already conducted using scored data, patterns in individual responses should be interpreted as revealing patterns in students' own internal biases rather than patterns in the qualities of experiments themselves. An elaboration on this point is presented in section 4.3.4.3: "Student independence of results". Patterns in mean scores prominently reflect patterns in students' internal biases when small samples are used to obtain the mean values, but increasingly reflect patterns in the qualities of experiments themselves as sample sizes used to calculate the mean scores are increased.

5.2 Issues in the design of learning activities

5.2.1 Key factors in student perception

A major conclusion in this thesis was estimation of a Linear Logistic Test Model for the ASLE survey data, and hence the identification of key factors underpinning ASLE survey responses. That is, the factors underpinning student perceptions of their laboratory learning experiences. Individual student biases were controlled for during these analyses, meaning that the factors obtained are only those which may be considered "objective"; the factors obtained are components of the laboratory learning experience whose relative quality can be assumed as generalizable to all students, and therefore can be controlled by the teacher through design of the activity. These key factors, presented in descending order of impact on the "objective" qualities of the learning experience, are as follows (see sections 4.3 and 4.4):

- Whether the activity is clearly seen by students to be reminiscent of content previously covered in lectures.
- The quality and appropriateness of the instructional material provided to students
- The understanding of theoretical content gained through collaboration with others.
- The extent to which data interpretation skills are developed through the activity
- Whether students work in groups or individually
- The students' reliance on their demonstrators
- Other unidentified or complex factors

Examining the interactions of these factors revealed a number of trends in student perception (see section 4.4). Perceptions of increased understanding were so strongly associated with perceived benefits of teamwork that the two manifested as one singular indistinguishable factor, reminiscent of suggested benefits of collaborative learning suggested in education literature.[314-318, 320, 321, 326] The exception to this was the development of technical skills, which were instead seen to benefit from individual work and interpretation of data. Student preference was observed to lean towards these more "skills-based" activities, with a strong "boredom" response associated with the inclusion of lecture content. High quality instructional notes were seen to be broadly beneficial.

These factors did not explain the entirety of variance in student perceptions seen in this study, however. Most notably, the "overall learning experience" appears to involve a substantial contribution from features of the learning experience not addressed within the ASLE survey. Identification of these factors is a goal for future research, and could potentially build upon the model presented here.

5.2.2 The need for compromise between students and teachers

Some of the validity issues discovered within the ASLE survey items have wider implications for chemistry education broadly. Conclusions of this research conducted in the context of laboratory learning specifically highlight issues that may also exist in learning activities beyond the lab, or which must be acknowledged if effective pedagogy is to be implemented.

Most prominent of these is the fact that student preferences often appear counterproductive to the attainment of learning outcomes. The contribution of the various factors in laboratory learning to student perspectives of a positive learning experience were seen to be biased against the inclusion of content reminiscent of lecture material, instead preferring more "skills-based" activities. Structuring learning activities by using appeal to the learner as the primary

guiding principle of design could therefore be catastrophic for the learning goals of the activity, conceivably even beyond the laboratory context. Teachers appear to have to choose between two types of learning activity:

Activities conducive to learning (see Equations 48 and 49)	**Activities appealing to students** (see Equations 50 and 51)
Understanding of theory through collaboration with others and connection to lecture content. Less reliance on demonstrators/ teachers.	Skills development through individual work and data interpretation, without "boring" lecture content. More reliance on demonstrators/teachers.

The fact that the ASELL project was created with the explicit goal of restructuring activities to appeal to students (in order to raise enrolment and retention in chemistry), implies that the project may contribute to a widespread decrease in standards of learning and teaching if this dilemma is not recognised. Whilst appealing to student preferences assists in meeting some goals of educational institutions, it hinders the achievement of others. Substantial weight needs to be given to the review of experiments by educators, for example using the Educational Template document and discussion at ASELL workshop, avoiding an exclusive appeal to the student perspective.

In this way a "compromise" needs to be made between the students, who wish to enjoy their learning experiences, and the teachers, who wish students to retain theoretical understandings. Recognising the student perspective has been increasingly viewed as important in education,[330] but the data discussed here suggest that neglecting the teacher perspective could potentially be damaging. Maximising student retention and enrolment in chemistry degrees at the cost of knowledge and understanding, as the ASELL project may have inadvertently pursued, would be a hollow victory.

The fact that student preference is often counterproductive to effective teaching also has further implications. Student evaluation data regarding the quality of courses or teachers could be subject to similar issues, implying that such data may not in fact reflect informative measures of quality (as may be assumed). In fact, highly positive ratings of practical demonstrators were seen in this thesis not to reflect high quality teaching, but instead were an indicator of a problematic teaching activity. Poorly designed learning activities may prompt students to be more appreciative when the teacher "rescues" them, whilst the teacher may go relatively unnoticed in a well-designed activity. This observation illustrates the point that appreciation for the teacher does not necessarily indicate high teacher competence, and that evaluating teacher competency based solely or primarily on student feedback may fail to account for any actual learning gains by the students. Again, this does not imply the student perspective should be ignored, merely that it should not be treated as the sole guiding factor in evaluating the quality of teachers or courses. Further research needs to be conducted to explore the impacts of catering to student preferences on quality of learning and knowledge retention.

5.2.3 There is no single best way to design a learning activity

Whilst the Linear Logistic Test Model derived in the final sections of this thesis reveals many patterns in student perceptions, closer examination also reveals the model does not provide a "one size fits all" solution to the question of how to design an activity. The key factors in the

laboratory learning experience are seldom always positive or always negative. Rather, different factors typically work towards some desirable outcomes, but against others. Additionally, factors which are highly relevant to some desirable outcomes may be much less relevant to others.

Through the LLTM, it was observed that the best way to design a laboratory activity depends on:

1. **The purpose of the activity.** As discussed previously (section 5.2.2), appeal to student interest is often at odds with gains in understanding and relevance to the course. In designing learning activities, therefore, teachers need to weigh and prioritise these two options. Skills-based activities have far greater student appeal, but teachers may need to instead prioritise the reinforcement of theoretical course content.

2. **The key learning objectives.** A strong difference was observed between factors contributing to skills development and factors contributing to theoretical understanding (sections 4.4.3 and 4.4.5). Activities where the primary learning objectives are practical skills are best designed without strong connection to lecture theory, instead focusing on data analysis with a strong responsibility placed on students for their own learning. Conversely, activities in which the key learning objectives are reminiscent of lecture content are best structured as collaborative group work activities, naturally including clear connection to the relevant lecture theory.

3. **The background knowledge of the students.** Evidence was shown (section 4.4.4) that students with differing levels of prior experience may perceive the identical activity differently. Features of the task such as the instructional material may need to be pitched at a level suitable to the audience, and factors such as time availability or the ability to understand the theoretical content may differ based on prior experience.

4. **The learning styles of individuals**. The data show that the more extreme perceptions of some students could not all be explained well by the broad scale patterns described (Table 26, p151). Factors relevant to the quality of the learning experience of some individuals may differ entirely to those which apply to the majority, meaning appropriate teaching methods may not only differ from class to class, but also from individual to individual. General trends in best practice will only ever suit the bulk of the student population, with students at the extreme "edges" best catered for with what could be entirely different pedagogy.

The complexity involved in structuring laboratory activities described above suggests that searching for a single "optimal" way to design activities may be misguided. Correlational studies such as the Barrie e*t al.* paper[61] which have the objective of revealing common themes in student perception may well be informative, but they will always only be representative of the "average" of the whole student sample used. If practicals delivered for different purposes, with different learning objectives, pitched to audiences of different levels of background knowledge or different learning styles are merged into a single analysis, the results may not adequately apply to any one of them taken individually. Correlations may simply reflect what was most often the case within the set of experiments used.

This thesis, for example, exclusively used data from first year undergraduate chemistry students in Australia, limiting the ability to generalise the LLTM formulated beyond that context. Further, these two different cohorts of students with different levels of background knowledge were merged into a single data set, meaning this potentially confounds the results obtained. By no means does this suggest the results obtained are uninformative – they suggest a range of important considerations for future research and teaching practice – it simply implies that perceptions reported by different student groups may behave in a somewhat different manner to the way student perceptions were observed to operate in this study. Testing the more general applicability of the LLTM derived here, or even the estimation of a new LLTM for a wider data set, is therefore a worthwhile pursuit.

5.3 Achievements in measurement

5.3.1 Reaffirmation of the advantages of Rasch methodology

A substantial component of this thesis naturally involved a comparison of Rasch measurement and scoring techniques more akin to classic test theory. Whilst these methods have been contrasted at length previously in the literature (see section 1.3.1), results here reaffirm many established conclusions, this time for the ASLE survey specifically. Despite this, the vast majority of practitioners are unlikely to adopt Rasch measurement techniques for the study of ASLE project data, principally because of small sample sizes in isolated analyses, data connectivity issues in wider scale studies and simply the fact that few have knowledge of how to estimate or interpret the results of Rasch modelling. For this reason, the evaluation of the validity of common scoring techniques presented in this thesis is of value, despite the fact scoring techniques are not ideal. However, should the issues with implementing Rasch analysis be overcome, Rasch measurement has numerous advantages:

1. **Rasch analysis provides interval scale measures**

   Rasch measurement yields genuine interval scale data, fit for parametric statistical methods. Whilst the calculation of means and estimation of standard error margins was seen to be reasonable for scores, Rasch measures are additionally fit for correlational work due to their known proportionality to the latent trait of interest. By contrast, correlational work was shown to be invalid using scored data, given their non-linear relation to the traits they are desired to reflect. (see section 3.3)

2. **Rasch measures are sample independent**

   A second primary limitation in scoring methodology was the heavy conflation between student dependent and student independent effects, leading to perturbation in scores received depending on the occasion and students surveyed (see sections 3.2 and 4.1.4). Rasch measurement techniques allow separation of generalizable qualities of the experiment itself from fluctuations in student biases, even allowing comparability of otherwise isolated data sets by using models such as the LLTM (see section 4.3.3.2).

3. **Rasch models are highly versatile**

   Scoring methods were seen in this thesis to have limitations for the analysis of larger data sets (see section 5.1.3). These issues are substantially alleviated when using Rasch measurement, which is readily amenable to analysing very large numbers of responses. Further, a wide range of different conceptions of how responders interact with the survey may be encapsulated within Rasch models (see sections 2.2, 4.1.2.2 and 4.3.3.1), in contrast to the tendency to assume one singular (simplistic) means by which students interact with surveys when applying scoring analyses (see section 3.3.2.1). The model is also made explicit within Rasch analyses in construction of the appropriate Rasch model, whereas it is often tacitly presumed in score-based analyses.

4. **Rasch models are testable**

   In contrast with the variety of validity assumptions required in score-based analysis (see for example section 1.2.2), as well as assumptions regarding the way surveys operate (see for example section 3.3.2.1), aspects of validity may be directly tested when using Rasch techniques. This is true for the construct validity associated with individual models (for example see sections 4.1.5 and 4.4.4) or even the validity of one

explanatory model of the data in contrast with others (for example see sections 4.1.3 and 4.3.3.1). These issues are far less readily investigated using scoring techniques, leading them to have been overlooked in the past. Validity assumptions made within Rasch analyses are, however, often directly testable through the fit statistics reported alongside Rasch models.

5. Rasch methodology is amenable to scientific investigation

Having been initially conceived in part to ensure the objectivity required for genuine scientific comparisons[129] (see section 1.3.1), Rasch modelling is an ideal tool for scientific inquiry. The additional ability to formulate nearly any hypothesised explanation of data variance as a Rasch model, coupled with the Rasch model's quantifiable and therefore testable predictions enables research to span far beyond the qualitative and exploratory. Model selection techniques such as use of the corrected Akaike Information Criterion (AICc) allow for various hypotheses, encapsulated within Rasch models, to be supported or refuted based on empirical observation and measurement. An iterative process of scientific inquiry utilising these advantages is shown in Figure 50.

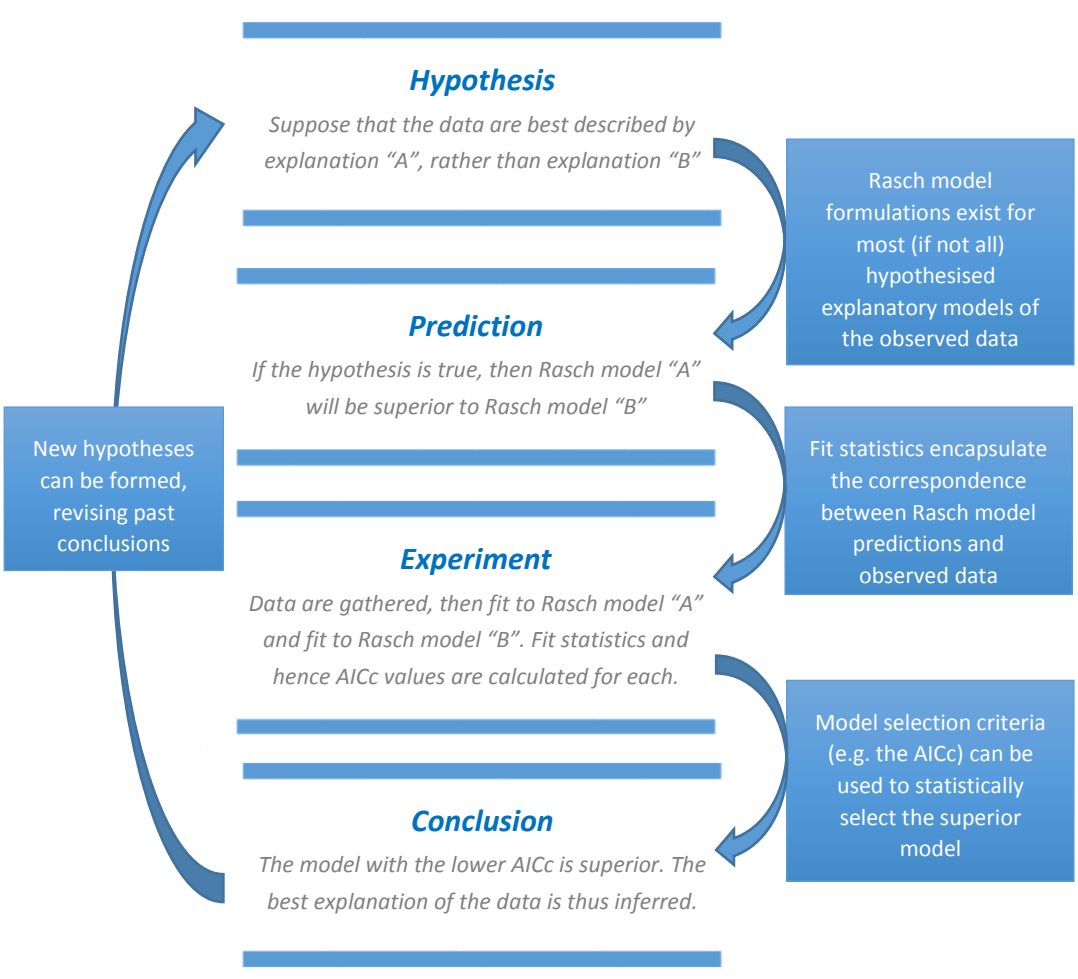


Figure 50: The scientific method applied using Rasch measurement techniques

All of these advantages were critical in drawing the primary conclusions of this thesis. Notably, methodology directly mapping to the scheme in Figure 50 above was implemented in this thesis to improve on models of the laboratory learning experience, without which the vast

majority of pedagogical conclusions discussed could not have been drawn. First, the interaction between students and surveys presumed by typical analyses (modelled by Equation 32, see section 3.3.2.1) was improved upon by including a variation in student dispositions from occasion to occasion (modelled by Equation 35, see section 4.1.1). Then, this generalised model was further improved by identifying the basic elements of the laboratory experience which give rise to perceptions reported on all 14 ASLE survey rating scale items (modelled by the LLTM, see sections 4.2 and 4.4). Future research could easily involve testing new refinements to the existing best explanation of the data using a similar process, leading to better understandings of laboratory-based teaching and learning.

5.3.2 Novel approaches to measurement problems

Whilst instrumental to this thesis, the scheme presented in Figure 50 is somewhat atypical of Rasch measurement practices. As previously described in the introductory material, Rasch models are usually used in a confirmatory manner, establishing that data fit to a model the researcher already has in mind (see section 2.3.2). By contrast, this thesis has used Rasch modelling in a more exploratory manner, using model fit as a tool to select the best explanation of the data from an array of proposed models.

The reason for this approach is largely due to the original design of the ASLE survey itself. The ASLE survey was never intended to be unidimensional; rather than measuring a single latent trait, different items of the survey were designed to address an array of (possibly independent) features of the learning experience which may or may not have influenced the overall quality. It was immediately recognised that this purpose of the survey invalidated commonly used unidimensional models, leading to the need to formulate a model which would be more in keeping with the intent of the survey. Section 3.3 therefore modelled the data without one single unidimensional Rasch model as would be typical, but instead using fourteen models: one for each separate dimension of the instrument. Rather than modifying the survey itself to more adequately address one single dimension (as would be more typical), the structure (and therefore purpose) of the survey was kept, instead tailoring the analysis to suit the survey itself.

This was a novel approach: the “items” within Rasch models estimated in this stage of research were not a set of many survey questions asked within a single context (as would be typical), they were instead a set of many contexts in which the same question was asked. There is nothing inherently “wrong” with this approach; “item” facets constructed in this manner may appear commonly in many facet Rasch models (see section 2.2.2) and models such as this can easily be derived from first principles (see section 7.4.1 of the supporting information for a derivation of all models used). However, typical methodology would instead suggest the survey be altered such that the series of items formed a single valid dimension of measurement. Altering the survey was not an option in this case: the major objectives of the thesis were to address validity concerns regarding the survey *as it had been presented previously* and as such the survey construct itself needed to remain in its current form.

Beyond the choice to amend the Rasch model used rather than amending the survey to fit a pre-established model, atypical approaches were also implemented when investigating the survey data’s qualitative meanings. It became clear early on that even the more suitable Rasch model selected was not necessarily the best option. Again, the goal here was to explore which interpretation of the existing survey was appropriate; the goal was not to amend the survey to fit a desired construct. Theoretical expectations and observed behaviour of the data do not

necessarily match in all cases, and this led to the exploration of which model best suited the data based on observation, not based on theory (section 4.1). This was a major shift in the use of Rasch models from confirmatory applications, instead now for exploratory purposes. Such an approach is far more typical of item response theory than it is of Rasch measurement (see section 2.3.2).

This exploratory approach allowed for a range of conclusions regarding both the survey's validity and implications of the data received. Contrasting an array of possible interactions between student and survey instrument (see section 4.1) allowed for a determination of the optimal qualitative understanding of the survey's use. Not only did this reveal crucial issues in comparability of survey results gathered in different occasions, it also enabled an explicit test of whether it was valid to assume features of the laboratory experience were objective and measurable. Having determined this to be reasonable based on data, not merely based on presumption, relationships between these measurable traits were able to be explored, finally generating the Linear Logistic Test Model.

Formulation of the Linear Logistic Test Model itself was also far from typical. In keeping with the usual confirmatory applications of Rasch measurement, Q-matrices for Linear Logistic Test models are typically stipulated a priori, not derived from observational data. This usually arises from a very deliberate and careful design of the survey or test at the outset, or else through consultation with experts in the relevant field (see section 2.2.2). However, neither of these approaches were possible in this case; the survey had not been designed with a specific Q-matrix construct in mind, nor was there sufficient expert knowledge available regarding how student perceptions precisely interact to stipulate an appropriate Q-matrix based on theory. Again a novel approach was taken to rectify the problem: the technique of factor analysis was was merged with Rasch modelling techniques to identify contributors to the Q-matrix, as well as determine the weights of those contributions on each survey item.

Factor analysis and Rasch analysis have not been combined in the precise manner implemented in this thesis previously, likely because of the atypical exploratory purpose of the analysis and the unconventional multidimensional structure of the survey itself. Moreover, the means of determining the appropriate number of factors to be retained was again atypical. Rather than using common techniques involving eigenvalues or scree plots in conjunction with what "makes sense" to the analyst (see section 2.4.6), objective Rasch model fit statistics and model selection criteria were used to choose the most appropriate number of factors. Confirmatory factor analysis in itself is not a new concept, but the marriage of exploratory factor analysis and the confirmatory statistics of Rasch modelling is an entirely new technique. The technique allowed for an objective selection of the factors to be included in the model uninfluenced by any expectations of the researcher, and further produced an LLTM with near perfect approximation to non-LLTM estimates (Figure 41, p 125). Analogous methods could easily be applied to other data sets, identifying variables which underpin the observed data and quantifying their contributions. This could be achievable for any research area making use of survey data, for which the survey items are not intended to comprise a singular dimension.

Though the fit to the data appear to suggest a resounding success of this estimation technique, commentary should be provided on the likely validity of the factors extracted. Most factors extracted were able to be assigned a reasonable real world interpretation by examining two pieces of information: the factor loadings on each of the initial ASLE survey items and patterns in the final measure values for each factor. This in itself lends some initial degree of credibility to the factor labels assigned, but far more compelling evidence of validity arises from the

predictions made by the model itself. Well known and extensively researched concepts in education were evident in the model under the factor identities assigned, including the strong relationship between collaboration and understanding as well as the importance of sequencing activities within the lecture course. This encapsulation of established theory was in no way built into the model itself or the estimation procedure, suggesting that it emerges as a consequence of a genuine correspondence between the model and real world patterns.

It should be noted that assigning factor character in this way necessarily draws upon what each factor correlates with. Or more specifically, what the factor is *observed* to correlate with in the data available. This creates a problem: features which form part of the *identity* of a factor will naturally correlate to its measure values, but so too will the features of the laboratory experience merely *influenced* by that factor. This conflation between identity and influence of the basic factors extracted may mean the labels attributed to them and descriptions of their character (see Table 25, section 4.4.1) could be in error.

Moreover, the factors modelled to underpin the experiences targeted by the survey items are themselves necessarily composed of the original survey item dimensions. That is, any identifiable defining characteristics of the factors extracted are restricted to being one of the topics addressed in the survey items the model seeks to explain. If there are any key factors underpinning survey responses which cannot directly be mapped back to a particular survey item (or items), then this technique simply will not be capable of identifying them. This issue posed a notable problem in the LLTM breakdown of the "overall learning experience" item: the primary contributor to measures for this item was a factor which itself only strongly correlated to "overall learning experience", failing to be assigned any other fundamental identity. Further, it was noted that the student cohort may play a role in factor identity, affecting which features were observed to correlate within the data and therefore which factor dimensions were constructed by the model (see section 4.4.4).

A revealing insight into validity of the factor characterisations assigned is that of factor 8 of the LLTM. This factor was seen to have no clear correlation to any item of the ASLE survey, making its identity a mystery initially, if indeed it was even a valid factor to extract. However, it was observed to have a very clear, albeit unconventional real world interpretation: correction of measures for item 11 (regarding benefit of teamwork) back to a binary response (reflecting whether students worked in pairs or not). Given its lack of correlation to any particular item of the survey, this factor would almost certainly have been discarded in a conventional factor analysis. Its retention here, however, was critical. It was also by definition a perfectly valid factor, since its measures mapped directly and exactly proportionally to a real world phenomenon.

Despite factor 8's validity, which lends credence to the notion that this technique does indeed yield at least some validly interpretable factors, its behaviour also exposes a problem with the method. The observed measures for item 11 of the survey are effectively one of two values if factor 8 is discounted: one value if students worked in pairs, but a different value of they worked individually. Clearly the binary option of whether students worked in pairs or not should be the singular primary factor defining this behaviour in an ideal LLTM, but this was not the estimated model. Instead, the model expressed this component of item 11's behaviour as an artificial combination of various teamwork-related features of the experience. This happened because a single binary item of whether students worked in pairs or not was not included on the survey, and so could not serve as the identity of any factor extracted. This is plainly not the most parsimonious solution, and poses an open goal for future research. A

method is needed to define a known factor into the Q-matrix at the start of the procedure, such that it will be included in the final Q-matrix estimated.

5.3.3 In pursuit of a specification equation

Whilst the LLTM formulation obtained reveals a wide range of connections between student experience and aspects of the activity design, making a direct mathematical connection between objective experiment quality and the real world features of experiment design has yet to be seen. Such a connection was described as a longer term goal for this research in the introductory material (see section 1.4.2) and there are a small number of indicators in the results discussed which suggest such a relationship may be attainable in future.

The measure for $\eta_{data\ interpretation}$ appears to be a direct function of the mode in which data is presented to students. Values for this factor took on low values for all cases in which the PASCO GLX Explorer handheld data logger was used, and consistently took on an equivalent higher value measure ($\eta_{data\ interpretation} \cong 0.28\ Logits$) when this was changed to a laptop computer equipped with PASCO Data Studio software (see section 4.4.5). The fact that the equivalent value was observed for all cases of using the laptop interface indicates a direct connection between activity design and the precise number value of the measure, as does the fact that shifting from data logger to laptop was consistently an improvement (though to a different extent in different experimental contexts). Similarly, the binary measure outcome for survey item 11; $\delta \cong -1.92\ Logits$ if students worked individually and $\delta \cong\ 0.08\ Logits$ if students worked in pairs, also suggests a direct mathematical connection between activity design and the precise number value of the measure.

A simplistic way to encapsulate the design of the experiment mathematically is to express it as a vector of many elements, each element of the vector pertaining to a different possible inclusion in the experiment design. Such a vector could have countless (even infinite) elements, corresponding to the countless different ways to design experiments: each element of the vector could take on a value of 1 or 0, corresponding to the inclusion or lack of inclusion of a specific possible feature in the design of the task respectively.

$$\vec{\omega}_m = [\omega_1 \quad \omega_2 \quad \cdots]\ ;\ \omega_z = \begin{cases} 1, & attribute\ z\ is\ true\ of\ experiment\ m \\ 0, & attribute\ z\ is\ not\ true\ of\ experiment\ m \end{cases} \qquad \textbf{51}$$

Conceptualising such a vector is useful, since it can be further imagined that each of the observed basic factor measures of the LLTM are some direct mathematical function of the experiment design vector. That is, the factor measures are resultant of the experiment design:

$$\eta_{f,m} = \Omega_f(\vec{\omega}_m) \qquad \textbf{52}$$

The survey item measures could also be expressed as a function of experiment design. Using this type of notation, a specification equation for survey item 11's measure could be constructed as follows:

$$\delta_{11\ (teamwork),\ m} = \Omega_{teamwork}(\vec{\omega}_m) = \begin{bmatrix} \vdots \\ -1.92 \\ 0.08 \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} \vdots \\ \omega_{Students\ work\ individually} \\ \omega_{Students\ work\ in\ pairs} \\ \vdots \end{bmatrix} \qquad \textbf{53}$$

Here the measure for item 11 (pertaining to perceived benefit of teamwork) for the mth experiment ($\delta_{11\ (teamwork),\ m}$) is expressed as a direct function ($\Omega_{teamwork}$) of the design of

the experiment ($\vec{\omega}_m$). The relevant features of the experiment design are students' requirement to work individually or in pairs. In Equation 53 above, truth or falsity of each of these options is expressed as a value of 1 or 0 respectively for the $\omega_{Students\ work\ individually}$ and $\omega_{Students\ work\ in\ pairs}$ terms.

Table 30 (page 178) summarises the full theoretical connection between experiment design and observed ASLE survey data, as explored in this research. Relevant equation numbers, as used in prior sections of this thesis, are listed to the left where appropriate. It can be seen that whilst a large portion of this connection has been revealed through the work in this thesis, the explanation of basic laboratory learning experience factors as a direct mathematical function of experiment design attributes remains largely undetermined. Solving these mathematical connections may serve as the goal of future work, potentially via iterated refinement of existing models (see Figure 50 presented previously).

Determining a precise specification equation which could be used by all teachers in all circumstances may not be possible, however. Evidence has been presented and discussed at length that different relationships exist within the measures obtained depending on the student cohort to which experiments and surveys are presented. A specification equation which appears consistently true for one student cohort may be incorrect for another. It is unlikely, therefore, that Rasch measurement for experiment design could ever be used to achieve what "Lexiles" have for reading (see section 1.3.2).

In the case of reading, the objective reading difficulty of a text appears to be consistent across the student population broadly, allowing "Lexiles" (a measure of reading difficulty derived from Rasch measurement) to be meaningful for any audience. This means that Lexiles can be calculated for an array of texts and disseminated to schools and educators as a useful tool in selecting texts appropriate for various readers. In the case of ASLE survey-derived Rasch measures for laboratory exercises, however, any objective measurements obtained may be specific to particular student audiences only. Consequently, measures associated with specific experiments could never be widely disseminated in this way without specifying the precise student audience which was used to obtain them. The measures are therefore far less simple to interpret, and very restricted in their utility. The practice of Rasch measurement and the derivation of any specification equations for the ASLE data in future is likely of far greater use in education research than it is for the purposes of widespread dissemination to teachers.

Table 30: Full model connecting experiment design to observed ASLE survey data

ASELL MEAN SCORES FOR EACH SURVEY QUESTION, FOR EACH EXPERIMENT		
28	$A_{i,m} = \left(\frac{1}{N_{i,m}}\right)\vec{a} \cdot \vec{X}_{i,m}\ ;\ \vec{a} = [-2 \quad -1 \quad +0 \quad +1 \quad +2]$	ASELL mean scores arise from averaging scored observed survey responses
OBSERVED SURVEY RESPONSE FREQUENCIES		
-	$\vec{X}_{i,m} = [c_{i,m,1} \quad c_{i,m,2} \quad \cdots \quad c_{i,m,K}]$	Each survey item, for each experiment has a total count of responses received in each rating scale category
34	$c_{i,m,k} \approx N_{i,m} \times P^*_{i,m}(x_k, \delta_{i,m})$	Observed response category counts arise from population level response category probabilities
33	$P^*_{i,m}(x_k, \delta_{i,m}) = \int_{-\infty}^{\infty} P_{n,m,i}(X = x_k) \times P(\beta_E) \,.\, d\beta_E$	Population level response category probabilities arise from summating probabilities for individual students
35	$ln\left[\frac{P_{n,m,i}(X = x_k)}{P_{n,m,i}(X = x_{k-1})}\right] = \beta_{E_{n,m}} - \delta_{PCM_{i,m}} - \tau_{i,k}$	Interaction between occasion specific student biases, survey question specific experiment quality measures and the response category structure interact to predict response probabilities for individual students
MEASURABLE LATENT VARIABLES/ EXPERIMENT QUALITIES UNDERPINNING RESPONSE		
42	$\delta_{LLTM_{i,m}} \approx -\delta_{PCM_{i,m}} + \gamma_s$	Disconnects between different subsets of the data may offset experiment quality measure estimates
40	$\delta_{LLTM_{i,m}} = \sum_{f=1}^{F} q_{i,f}\eta_{f,m} + \mu_i$	Survey question specific experiment quality measures are a linear combination of more basic experiment specific factors
52	$\eta_{f,m} = \Omega_f(\vec{\omega}_m)$	Measures for basic experiment specific factors are each a direct function of the attributes of experiment design
EXPERIMENT DESIGN ATTRIBUTES		
51	$\vec{\omega}_m = [\omega_1 \quad \omega_2 \quad \cdots]\ ;\ \omega_z = \begin{cases} 1, & attribute\ z\ is\ true\ of\ experiment\ m \\ 0, & attribute\ z\ is\ not\ true\ of\ experiment\ m \end{cases}$	

Relevant equation numbers used in the body of this thesis are shown at the far left of the table.

VARIABLES: ASELL mean score (A), total student responses (N), observed response category count (c), population level probability (P*), individual student probability (P), observed response (X), response category (x), student dependent measure (β_E), student independent measure (δ_{PCM} for Partial Credit Model, δ_{LLTM} for Linear Logistic Test Model), category threshold (τ), measurement subset offset (γ), basic factor weighting (q), basic factor measure (η), survey question relative location (μ), specification equation function (Ω), experiment design description vector ($\vec{\omega}$)

VARIABLE INDICES: n^{th} person, m^{th} experiment, i^{th} survey question, k^{th} response category (of K), s^{th} measurement subset, f^{th} basic factor (of F), z^{th} experiment design attribute

5.4 Future investigation with the Linear Logistic Test Model

5.4.1 Uniting the broader ASELL database

Given the breadth of conclusions able to be drawn from the Linear Logistic Test Model generated from the data set used in this thesis, as well as the questions existing regarding its generality to other data sets, it is desirable to be able to test whether other data sets exhibit the same patterns. The ability to test the applicability or otherwise of a model such as the LLTM is one of the primary strengths of Rasch measurement techniques, making this objective entirely within reach. The ASELL project also has a vast array of data from over 120 evaluated experiments at its disposal,[57] providing an ideal opportunity to test the applicability of the LLTM generated here for a larger and more diverse data set.

Unlike using scored response data, the LLTM inherently separates student dependent and student independent factors. This naturally avoids effects whereby the conclusions drawn reflect patterns in students' inherent predispositions more than patterns in generalizable ways to improve the learning experience (see section 4.3.4.3). The LLTM also naturally connects data sets gathered from different experiments, as the same underlying factors contribute to the perceptions observed for each (see section 4.3.3.2). The LLTM derived here also has the advantage that it was derived objectively from the data alone, without stipulation of which factors to include in the model using subjective researcher judgement (see section 4.3.4.4). These three benefits: the isolation of student independent trends, connection of data sets and objectivity of the model's derivation provide additional motive to estimate parameters associated with experiments in the ASELL database, using the LLTM structure.

In this study, the *Facets* Rasch measurement software was used to estimate parameters for the ASLE data LLTM. Whilst this was achievable, structuring the Q-matrix within the *Facets* software specification file is by no means simple, and cannot be achieved without the aid of other technology such as *Microsoft Excel*. It would be more convenient to utilise more capable Rasch measurement software such as *ConQuest*,[151, 156] within which matrix weighting coefficients can be more simply stipulated. This would also allow the use of non-integer values in the Q-matrix, making the resultant model more accurate. The procedure used in this thesis to generate a Linear Logistic Test Model within the *Facets* software is presented in the supporting information (see section 7.6.1), as are both the integer value and non-integer value forms of the Q-matrix (see section 7.6.3).

The hypothesis of whether the LLTM generated in this research applies to the wider ASELL data set can simply be tested using the corrected Akaike Information Criterion (AICc). The same data could be fit to the LLTM and an analogous non-LLTM model, calculating the AICc value for each. The LLTM can be deemed the best explanatory model of the data if it has the lower AICc value. Further elaboration on AICc value interpretation is presented in section 2.5.4.2.

Because student dependent parameters are best modelled as constant for a singular occasion, but varying between different occasions (see section 4.1.3), estimation of an analogous non-LLTM model of the ASELL data would result in myriad isolated subsets of data. An alternate means of calculating the fit of the data to a non-LLTM analogue is therefore preferable. Once the LLTM has been estimated within Rasch measurement software (such as *Facets*), non-LLTM analogues of all survey item measures estimated can be obtained simply by adding "displacement measures" back to the δ_{LLTM} measures obtained. A Partial Credit model of the

data can then be structured by anchoring all δ_{PCM} measures to equal these δ_{LLTM} + *displacement* values. The fit statistics of this Partial Credit Model would then reflect the fit of the data to a non-LLTM model, which can be contrasted with the fit observed to the LLTM.

AICc values can then be calculated for each of the two alternate models using Equation 31 (reproduced below), where n is the total number of data points gathered (14 per survey, if all surveys are complete) and the $-2\ln(\mathcal{L})$ term is the log-likelihood chi squared value quoted in the *Facets* software as a measure of global model fit ($\mathcal{L}$ is the likelihood value for the full model, which may be quoted in other Rasch measurement software).

$$AICc = -2\ln(\mathcal{L}) + 2k + \frac{2k(k+1)}{n-k-1} \qquad \textbf{31}$$

The number of free parameters (k) to be used in the equation above for the (non-LLTM) Partial Credit Model can be simply calculated using the number of experiments conducted (X) and the number of surveys gathered (N), as shown below.

k	=	56	τ measures	(4 Rasch-Andrich thresholds × 14 items)
		+ 14X	δ measures	(14 ASLE item measures per experiment)
		+ N	β_E measures	(one per survey occasion, even for the same students)
		− 1		(one facet centred at zero)

For the LLTM, a different set of parameters is involved and therefore the number of free parameters must be computed differently. The total number of free parameters (k) depends on the number of factors (F) contributing to the final LLTM approximations to each survey item measure. In this thesis, a 12 factor model (F=12) was computed to be optimal. However, because the factor loading values were rounded to integers, the twelfth factor was rendered not to contribute, leaving only 11 factors in the model for which results are presented (F=11).

k	=	56	τ measures	(4 Rasch-Andrich thresholds × 14 items)
		+ 14	μ measures	(Defining relative location of measures for the 14 survey items)
		+ F. X	η measures	(F factor measures per experiment)
		+ N	β_E measures	(one per survey occasion, even for the same students)
		− 1		(one facet centred at zero)

Having calculated AICc values for the LLTM and non-LLTM models of the data, the preferable model can be concluded (the model of lowest AICc value).

Should the procedure described above reveal that the LLTM presented in this thesis is not a suitable explanation of the wider ASLE data set, it may not be possible to emulate the procedure described within section 4 of this thesis to obtain a better LLTM, due to issues of data connectivity. Currently, ASLE surveys are typically anonymous, meaning any given experiment evaluated with the ASLE survey has no connection to other analyses. In analyses described within this thesis, students responding to multiple experiments were used to 'equate' different isolated subsets of data, allowing comparability between measures and subsequently permitting factor analysis and Q-matrix estimation. A similar procedure would be necessary to equate all experiments in the ASLE survey database if a new LLTM were to be estimated. This is not possible given respondent anonymity, however.

One possible solution to the problem could be to designate a series of "calibration" responders: specific individuals who complete a number of experiments and evaluate them using ASLE surveys, whose responses may be used to connect the wider database. An immediately apparent option for who these individuals may be is ASELL workshop attendees. If ASELL workshop attendees completed ASLE surveys for each experiment they conduct at the workshop, and if multiple surveys completed by the same individual could be tracked, it may be possible to use those individuals to unite the data sets associated with experiments submitted by different institutions. This may at least unite the set of experiments conducted at workshops, but the same could not be said for the data gathered at home institutions: again, responders common to multiple experiments, whose biases toward positive response can be assumed not to change, must be present to unite separate datasets.

5.4.2 Improving the current LLTM

Multiple observations over the course of this research have indicated that the LLTM estimated here may need to be further refined if it is to be applicable to a wider data set. The limited diversity of the student audience for these studies substantially restricts the generality of any precise mathematical patterns observed. Further, the model may conflate the identity and influences of key factors underpinning student perceptions (see section 5.3.2), meaning the LLTM estimated could have been perturbed by features which correlate by chance in the sample of experiments selected.

Substantial contributions by factors of unknown identity have also been revealed, suggesting a target for future research. Measures for the "overall learning experience" item particularly have a majority of variance explained by factors other than those targeted by the ASLE survey or identified within the LLTM. Identifying these factors is pivotal not only to refinement of the LLTM as a model, but also to the understanding of what makes students view their laboratory experiences positively.

The aspects of the laboratory experience currently included on the survey do not encompass all considerations relevant to student perceptions. Were these factors identified, the current model could be substantially improved. A simple solution to this could be to structure a new survey, including some items present on the current ASLE survey, but others designed to investigate different features of the student experience. New question could be designed to target features of the activity design which appeared relevant to the measure values obtained here, but were not explicitly targeted previously. Rating scale items which might be included on a revised survey are suggested below.

Items addressing appropriateness to the prior learning of the audience:

- I found the theory in this experiment to be (above my level, appropriate to my level, below my level)
- I found the technical skills in this experiment to be (above my level, appropriate to my level, below my level)

Items which may reveal the learning style the experiment is best directed towards:

- This activity involves hands-on interaction with chemical concepts
- This activity allows me to visualise chemistry in action
- This activity requires me to use difficult or complex symbols / mathematics
- This activity requires me to remember and apply chemical concepts

Structuring a new survey could also allow an opportunity to target the basic factors seen to underpin ASLE survey response more directly. Doing so could not only serve as a tool for more direct measurements of fundamental components of the laboratory learning experience, but could also be used as a tool to confirm the precise defining characteristics of the six known factors identified in the LLTM (see Table 25). Some rating scale items which may assist in these objectives are suggested below.

For factor 1 (theory focus):

- This activity is strongly connected to the lecture course content
- The main purpose of this activity is to develop my technical laboratory skills
- The main purpose of this activity is to reinforce my understanding of lecture theory

For factor 2 (instructions):

- The instructional material provided is clear and sufficient
- I easily understood the material provided to me in the notes for this activity
- The instructional notes for this activity provided all information I needed

For factor 3 (collaborative understanding):

- This activity has increased my understanding of chemistry theory
- This activity involves collaboration with others
- This activity has increased my understanding of chemistry theory through collaboration with others

For factor 4 (data interpretation):

- This activity allows me to improve my data interpretation skills
- This activity involved the use of technology
- The technology used in this activity was simple to operate

For factor 5 (independent learning):

- This activity allows me to learn independently
- In this experiment I worked (in pairs or in a group / individually)

For factor 6 (demonstrators):

- In this activity I needed help from my demonstrator

If a new survey were structured including questions such as these, and similar methods to those discussed over the course of this thesis were applied, a LLTM obtained would be likely to better encapsulate the factors contributing to student experiences than the model presented here. If such a study were conducted at the University of Adelaide, gathering data for the identical experiments studied already, measures obtained for any new survey items could be included alongside measures obtained from research in this thesis to estimate a vastly improved LLTM for the ASLE data. This may be a far more viable solution to reformulation of the LLTM than uniting the wider ASLE database.

Refinement of the understandings gained through the course of these works, or identification of entirely new factors contributing to student perceptions of laboratory learning experiences in this way could continue to build upon the array of knowledge revealed by the model

presented in this thesis. Not only could this continue to inform effective pedagogy of science, but also serve to reveal key questions for future research in science education and teaching in laboratories.

6 *References*

1. Rowland, H. A. The Physical Laboratory in Modern Education. *Science* 1886, 7, 573-575.
2. Hofstein, A.; Lunetta, V. N. The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research* 1982, 52, 201-217.
3. Hawkes, S. J. Chemistry Is Not a Laboratory Science. *Journal of Chemical Education* 2004, 81, 1257.
4. Hofstein, A.; Lunetta, V. N. The Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research* 1982, 52, 201-217.
5. Hofstein, A.; Lunetta, V. N. The Laboratory in Science Education: Foundation for the 21st Century. *Science Education* 2004, 88, 28-54.
6. Hofstein, A.; Mamlok-Naaman, R. The Laboratory in Science Education: The State of the Art. *Chemistry Education Research and Practice* 2007, 8, 105-107.
7. Hofstein, A. The Laboratory in Chemistry Education: Thirty Years of Experience with Developments, Implementation and Research. *Chemistry Education Research and Practice* 2004, 5, 247-264.
8. Psillos, D.; Niedderer, H. *Teaching and Learning in the Science Laboratory*. Kluwer: Dordrecht, 2002.
9. Reid, N.; Shah, I. The role of laboratory work in university chemistry. *Chemistry Education Research and Practice* 2007, 8, 172-185.
10. Seymour, E.; Hunter, A. B.; Laursen, S. L.; DeAntoni, T. Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education* 2004, 88, 493-534.
11. Schwab, J. J. The Teaching of Science as Inquiry. *Bulletin of the Atomic Scientists* 1958, 14, 374-379.
12. Romey, W. D. *Inquiry techniques for teaching science*. Pretence-Hall: Englewood Cliffs, New Jersey, 1969.
13. Tsaparlis, G. Learning at the Macro Level: The Role of Practical Work. In *Multiple representations in Chemical Education*, Gilbert, J. K.; Treagust, D., Eds. Springer Netherlands: 2009; pp 109-136.
14. Hegarty-Hazel, E. *The Student laboratory and the science curriculum*. Routledge: London, 1990.
15. Moore, J. W. Let's Go for an A in Lab. *Journal of Chemical Education* 2006, 83, 519.
16. Moskovitz, C.; Kellogg, D. Inquiry-Based Writing in the Laboratory Course. *Science* 2011, 332, 919-920.
17. Hidi, S.; Renninger, K. A. The four-phase model of interest development. *Educational psychologist* 2006, 41, 111-127.
18. Hidi, S. Interest and its contribution as a mental resource for learning. *Review of Educational research* 1990, 60, 549-571.
19. George, B.; Wystrach, V. P.; Perkins, R. Why do students choose chemistry as a major? *Journal of Chemical Education* 1985, 62, 501.

20. Weaver, G. C.; Russell, C. B.; Wink, D. J. Inquiry-based and research-based laboratory pedagogies in undergraduate science. *Nat Chem Biol* 2008, 4, 577-580.

21. Hippel, W. v.; Lerner, J. S.; Gregerman, S. R.; Nagda, B. A.; Jonides, J. Undergraduate student-faculty research partnerships affect student retention. *The Review of Higher Education* 1998, 22, 55-72.

22. Johnstone, A. H.; Letton, K. M. Teaching the Large Course: Is Practical Work Practicable? *Journal of College Science Teaching* 1989, 18, 190-92.

23. Johnstone, A.; Al-Shuaili, A. Learning in the laboratory; some thoughts from the literature. *University Chemistry Education* 2001, 5, 42-51.

24. Barrie, S. C.; Buntine, M. A.; Jamie, I. M.; Kable, S. H. APCELL: The Australian Physical Chemistry Enhanced Laboratory Learning Project. *Australian Journal of Education in Chemistry* 2001, 57, 6-12.

25. Boud, D.; Dunn, J.; Hegarty-Hazel, E. *Teaching in laboratories*. Guildford, Surrey England : Society for Research into Higher Education & NFER-Nelson: Guildford, Surrey [England], 1986.

26. Gibbs, G.; Gregory, R.; Moore, I. *Labs and Practicals: With More Students and Fewer Resources*. Oxford Centre for Staff Development: 1997.

27. Barrie, S. C.; Buntine, M. A.; Jamie, I. M.; Kable, S. H. In *APCELL: Developing better ways of teaching in the laboratory*, Proceedings of Research and Development into University Science Teaching and Learning Workshop, Sydney, NSW, Fernandez, A., Ed. UniServe Science: Sydney, NSW, 2001; pp 23-28.

28. Barrie, S.; Buntine, M.; Jamie, I.; Kable, S. Physical Chemistry in the Lab. *Chemistry in Australia* 2001, 68, 36-37.

29. Lim, K. F. The Australian journal of education in chemistry. *Chemistry in Australia* 2009, 76, 20.

30. O'Grady, B. The Determination of the Dissociation Constant of a Weak Acid by Titration: An APCELL Experiment. *Australian Journal of Education in Chemistry* 2001, 57, 13-17.

31. Barnett, V. Laser-based liquid prism sucrosemeter: An APCELL experiment. *Australian Journal of Education in Chemistry* 2002, 59, 5-10.

32. Lim, K. F. Inhibition of the reaction kinetics of the enzyme o-diphenol oxidase: An APCELL experiment. *Australian Journal of Education in Chemistry* 2002, 59, 11-16.

33. McNaughton, D. The identification of drugs by infrared and Raman spectroscopy: An APCELL experiment. *Australian Journal of Education in Chemistry* 2002, 60, 5-8.

34. Williamson, B. E.; Taylor, K. C. Group Theory and the Near-ultraviolet Absorption Spectrum of Gas-phase Benzene: An APCELL Experiment. *Australian Journal of Education in Chemistry* 2002, 58, 13-20.

35. Barnett, V. Reactions in Non-Ideal Solution-The Effect of Ionic Strength on the Rate of Reactions between Ions in Aqueous Solution (The Kinetic Salt Effect). *Australian Journal of Education in Chemistry* 2003, 62, 5-8.

36. Gascooke, J.; Shapter, J. G. Electronic Spectra of Benzene: An APCELL Experiment. *Australian Journal of Education in Chemistry* 2003, 61, 26-29.

37. Lim, K. An IR investigation of the CO dipole direction and other properties. *Australian Journal of Education in Chemistry* 2004, 64, 24-28.

38. Metha, G.; Buntine, M.; Kable, S. The emission spectroscopy of C2 produced in a hydrocarbon/oxygenflame: An APCELL experiment. *Australian Journal of Education in Chemistry* 2004, 63, 21-25.

39. Wajrak, M.; Rummey, J. Determination of silver by differential pulse anodic stripping voltammetry: An APCELL experiment. *Australian Journal of Education in Chemistry* 2004, 63, 26-30.

40. Lim, K. F.; Dyson, G. A.; Mitchell, C. Phosphorus and selenium NMR spectroscopy: an APCELL experiment. *Australian Journal of Education in Chemistry* 2007, 68, 23-26.

41. Price, W. E.; Griffith, D. W.; Wilson, S. R. Thermodynamics of the NO2-N2O4 Equilibrium by FTIR: An APCELL Experiment. *Australian Journal of Education in Chemistry* 2007, 67, 14-17.

42. Read, J. R. The Australian Chemistry Enhanced Laboratory Learning Project. *Chemistry in Australia* 2006, 73, 3-5.

43. Read, J. R.; Barrie, S. C.; Bucat, R. B.; Buntine, M. A.; Crisp, G. T.; George, A. V.; Jamie, I. M.; Kable, S. H. Achievements of an ACELL workshop. *Chemistry in Australia* 2006, 73, 17-20.

44. Read, J. R.; Buntine, M. A.; Crisp, G. T.; Barrie, S. C.; George, A. V.; Kable, S. H.; Bucat, R. B.; Jamie, I. M. In *The ACELL project: Student participation, professional development, and improving laboratory learning*, Symposium Proceedings: Assessment in Science Teaching and Learning, Sydney, NSW, UniServe Science.: Sydney, NSW, 2006; pp 113-119.

45. Buntine, M. A.; Read, J. R.; Barrie, S. C.; Bucat, R. B.; Crisp, G. T.; George, A. V.; Jamie, I. M.; Kable, S. H. Advancing Chemistry by Enhancing Learning in the Laboratory (ACELL): a model for providing professional and personal development and facilitating improved student laboratory learning outcomes. *Chemistry Education Research & Practice* 2007, 8, 232-254.

46. Jamie, I. M.; Read, J. R.; Barrie, S. C.; Bucat, R. B.; Buntine, M. A.; Crisp, G. T.; George, A. V.; Kable, S. H. From APCELL to ACELL and beyond : expanding a multi-institution project for laboratory-based teaching and learning. *Australian Journal of Education in Chemistry* 2007, 1-17,23.

47. Buntine, M. A.; Read, J. R. Guide to content analysis *Available from http://www.asell.org/Educational-Information/Guide-to-Content-Analysis* [Online], 2007.

48. Pyke, S. M.; Yeung, A.; Kable, S. H.; Sharma, M. D.; Barrie, S. C.; Buntine, M. A.; Da Silva, K. B.; Lim, K. F. In *The Advancing Science by Enhancing Learning in the Laboratory (ASELL) Project: The Next Chapter*, 16th UniServe Science Annual Conference, 2010.

49. Wajrak, M.; Boyce, M. The determination of the best separation conditions for a mixture of preservatives of varying polarity using HPLC: An ACELL experiment. *Australian Journal of Education in Chemistry* 2005, 65, 20-23.

50. Read, J. R.; Kable, S. H. Educational analysis of the first year chemistry experiment 'Thermodynamics Think-In': an ACELL experiment. *Chemistry Education Research & Practice* 2007, 8, 255-273.

51. Crisp, M. G.; Kable, S. H.; Read, J. R.; Buntine, M. A. A disconnect between staff and student perceptions of learning: an ACELL educational analysis of the first year undergraduate chemistry experiment 'investigating sugar using a home made polarimeter'. *Chemistry Education Research and Practice* 2011, 12, 469-477.

52. Wilson, K.; Mills, D.; Sharma, M.; Kirkup, L.; Mendez, A.; Scott, D. In *ACELL for Physics?*, Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference), 2012.

53. Yeung, A.; Pyke, S. M.; Sharma, M. D.; Barrie, S. C.; Buntine, M. A.; Da Silva, K. B.; Kable, S. H.; Lim, K. F. The Advancing Science by Enhancing Learning in the Laboratory (ASELL) Project: The first Australian multidisciplinary workshop. *International Journal of Innovation in Science and Mathematics Education* 2011, 19, 51 - 72.

54. Kable, S.; Buntine, M.; Yeung, A.; Sharma, M.; Lim, K.; Pyke, S.; Da Silva, K. B.; Barrie, S. *Advancing Science by Enhancing Learning in the Laboratory (ASELL) Final Report 2012*; Australian Learning and Teaching Council: 2012.

55. Heifer, E. ASELL Schools. *Science Education News* 2015, 64, 65-66.

56. ASELL Advancing Science by Enhancing Learning in the Laboratory website. http://www.asell.org (accessed August 2014).

57. Yeung, A.; Sharma, M.; Kable, S.; Lim, K.; Sutherland, L.; Buntine, M.; Dawson, V.; Southam, D.; Maynard, N. In *Enhancing student engagement in laboratory learning using inquiry-based activities: Expanding ASELL into schools*, The International Chemical Congress of Pacific Basin Societies, Honolulu, Hawaii, USA, Honolulu, Hawaii, USA, 2015.

58. Buntine, M.; Burke da Silva, K.; Lim, K.; Pyke, S.; Read, J.; Sharma, M.; Yeung, A.; Kable, S. In *Student perceptions and staff misconceptions about the undergraduate laboratory learning experience*, The International Chemical Congress of Pacific Basin Societies, Honolulu, Hawaii, USA, Honolulu, Hawaii, USA, 2015.

59. George, A. V.; Read, J. R.; Barrie, S. C.; Bucat, R. B.; Buntine, M. A.; Crisp, G. T.; Jamie, I. M.; Kable, S. H. What Makes a Good Laboratory Learning Exercise? Student Feedback from the ACELL Project. In *Chemistry Education in the ICT Age*, Gupta-Bhowon, M.; Kam Wah, H.; Jhaumeer-Laulloo, S.; Ramasami, P., Eds. Springer Netherlands: 2009; pp 363-376.

60. Southam, D. C.; Shand, B.; Buntine, M. A.; Kable, S. H.; Read, J. R.; Morris, J. C. The timing of an experiment in the laboratory program is crucial for the student laboratory experience: acylation of ferrocene as a case study. *Chemistry Education Research and Practice* 2013, 14, 476-484.

61. Barrie, S. C.; Bucat, R. B.; Buntine, M. A.; Burke da Silva, K.; Crisp, G. T.; George, A. V.; Jamie, I. M.; Kable, S. H.; Lim, K. F.; Pyke, S. M.; Read, J. R.; Sharma, M. D.; Yeung, A. Development, Evaluation and Use of a Student Experience Survey in Undergraduate Science Laboratories: The Advancing Science by Enhancing Learning in the Laboratory Student Laboratory Learning Experience Survey. *International Journal of Science Education* 2015, 37, 1795-1814.

62. Palmer, D. H. Student interest generated during an inquiry skills lesson. *Journal of Research in Science Teaching* 2009, 46, 147-165.

63. Likert, R. A technique for the measurement of attitudes. *Archives of Psychology* 1932, 22.

64. Carifio, J.; Perla, R. J. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences* 2007, 3.

65. Carifio, J.; Perla, R. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 2008, 42, 1150-1152.

66. Delucchi, K. L. The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin* 1983, 94, 166-176.

67. Marascuilo, L. A.; Dagenais, F. Planned and Post Hoc Comparisons for Tests of Homogeneity where the Dependent Variable is Categorical and Ordered. *Educational and Psychological Measurement* 1982, 42, 777-781.

68. Campbell, M. J.; Julious, S. A.; Altman, D. G. Estimating Sample Sizes For Binary, Ordered Categorical, And Continuous Outcomes In Two Group Comparisons. *BMJ: British Medical Journal* 1995, 311, 1145-1148.

69. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* 1900, 50, 157-175.

70. Kruskal, W. H. Historical Notes on the Wilcoxon Unpaired Two-Sample Test. *Journal of the American Statistical Association* 1957, 52, 356-360.

71. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1945, 1, 80-83.

72. Mann, H. B.; Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 1947, 18, 50-60.

73. Conover, W. J. *Practical nonparametric statistics*. 2nd ed.; Wiley: New York, 1980.

74. Kruskal, W. H.; Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* 1952, 47, 583-621.

75. Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* 1938, 30, 81-93.

76. Spearman, C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 1904, 15, 72-101.

77. Student. The Probable Error of a Mean. *Biometrika* 1908, 6, 1-25.

78. Welch, B. L. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938, 29, 350-362.

79. Welch, B. L. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika* 1947, 34, 28-35.

80. Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* 1946, 2, 110-114.

81. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 1919, 52, 399-433.

82. Fisher, R. A. On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *JSTOR* 1921, 1, 3-32.

83. Welch, B. L. On the comparison of several mean values: an alternative approach. *Biometrika* 1951, 330-336.

84. Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London* 1895, 58, 240-242.

85. Jamieson, S. Likert scales: how to (ab)use them. *Medical Education* 2004, 38, 1217-1218.

86. Clason, D. L.; Dormody, T. J. Analyzing Data Measured by Individual Likert-Type Items. *Journal of Agricultural Education* 1994, 35, 31-35.

87. Sisson, D. A.; Stocker, H. R. Analysing and interpereting Likert-type survey data. *The Delta Pi Epsilon Journal* 1989, 31, 81-85.

88. Glass, G. V.; Peckham, P. D.; Sanders, J. R. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research* 1972, 42, 237-288.

89. Havlicek, L. L.; Peterson, N. L. Robustness of the Pearson Correlation Against Violations of Assumptions. *Perceptual and Motor Skills* 1976, 43, 1319-1334.

90. Graubard, B. I.; Korn, E. L. Choice of Column Scores for Testing Independence in Ordered 2 x K Contingency Tables. *Biometrics* 1987, 43, 471-476.

91. Norman, G. Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education* 2010, 15, 625-632.

92. Kuzon Jr, W. M.; Urbanchek, M. G.; McCabe, S. The seven deadly sins of statistical analysis. *Annals of plastic surgery* 1996, 37, 265-272.

93. Kelley, T. L. *Interpretation of educational measurements*. Macmillan: New York, 1927.

94. Cronbach, L. J.; Meehl, P. E. Construct validity in psychological tests. *Psychological bulletin* 1955, 52, 281.

95. Cronbach, L. J. Five perspectives on validity argument. In *Test validity*, 1988; pp 3-17.

96. Messick, S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 1995, 50, 741.

97. Newton, P. E. Clarifying the Consensus Definition of Validity. *Measurement: Interdisciplinary Research and Perspectives* 2012, 10, 1-29.

98. Kane, M. T. Current concerns in validity theory. *Journal of Educational measurement* 2001, 38, 319-342.

99. Kane, M. T. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 2013, 50, 1-73.

100. Newton, P.; Shaw, S. *Validity in Educational and Psychological Assessment*. SAGE: 2014.

101. Newton, P. E.; Shaw, S. D. Standards for talking and thinking about validity. 2013.

102. Borsboom, D.; Mellenbergh, G. J.; van Heerden, J. The concept of validity. *Psychological review* 2004, 111, 1061.

103. Crocker, L.; Algina, J. *Introduction to classical and modern test theory*. ERIC: 1986.

104. Alagumalai, S.; Curtis, D. D. Classical test theory. In *Applied Rasch measurement: A book of exemplars*, Springer: 2005; pp 1-14.

105. Hambleton, R. K.; Swaminathan, H.; Cook, L. L.; Eignor, D. R.; Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research* 1978, 467-510.

106. Lord, F. M. *Applications of item response theory to practical testing problems*. Routledge: 1980.

107. Hambleton, R. K. *Fundamentals of item response theory*. Sage publications: 1991; Vol. 2.

108. Fan, X. Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement* 1998, 58, 357-381.

109. Güler, N.; Uyanık, G. K.; Teker, G. T. Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education* 2014, 2, 1-6.

110. Macdonald, P.; Paunonen, S. V. A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement* 2002, 62, 921-943.

111. Progar, S.; Socan, G.; Slovejija, M. An empirical comparison of item response theory and classical test theory. *Horizons of Psychology* 2008, 17, 5-24.

112. Xu, T.; Stone, C. A. Using IRT Trait Estimates Versus Summated Scores in Predicting Outcomes. *Educational and Psychological Measurement* 2012, 72, 453-468.

113. Wright, B. D.; Linacre, J. M. Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation* 1989, 70, 857-860.

114. Wright, B. D. Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement* 1977, 14, 97-116.

115. Wright, B. D.; Mok, M. Understanding Rasch measurement: Rasch models overview. *Journal of applied measurement* 2000.

116. Boone, W. J.; Townsend, J. S.; Staver, J. Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education* 2011, 95, 258-280.

117. Harwell, M. R.; Gatti, G. G. Rescaling Ordinal Data to Interval Data in Educational Research. *Review of Educational Research* 2001, 71, 105-131.

118. Kathryn, S.; William, J. B. Designing Tests and Surveys for Chemistry Education Research. In *Nuts and Bolts of Chemical Education Research*, American Chemical Society: 2008; Vol. 976, pp 149-169.

119. Curtis, D. D.; Boman, P. X-Ray Your Data with Rasch. *International Education Journal* 2007, 8, 249-259.

120. Wei, S.; Liu, X.; Jia, Y. Using Rasch Measurement to Validate the Instrument of Students' Understanding of Models in Science (SUMS). *Int J of Sci and Math Educ* 2013, 1-16.

121. Fortus, D.; Vedder-Weiss, D. Measuring students' continuing motivation for science learning. *Journal of Research in Science Teaching* 2014, 51, 497-522.

122. Scantlebury, K.; Boone, W.; Kahle, J. B.; Fraser, B. J. Design, validation, and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching* 2001, 38, 646-662.

123. Rasch, G. An Item Analysis which Takes Individual Differences Into Account. *British Journal of Mathematical and Statistical Psychology* 1966, 19, 49-57.

124. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research: Copenhagen, 1960.

125. Wright, B.; Panchapakesan, N. A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement* 1969, 29, 23-48.

126. Wright, B.; Masters, G. N. *Rating Scale Analysis*. MESA Press: Chicago, 1982.

127. Roskam, E. E.; Jansen, P. G. W. A New Derivation of the Rasch Model. In *Advances in Psychology*, Degreef, E.; Buggenhaut, J. V., Eds. North-Holland: 1984; Vol. Volume 20, pp 293-307.

128. Fischer, G. H. Derivations of the Rasch Model. In *Rasch Models: Foundations, recent developments and applications*, Fischer, G.; Molenaar, I., Eds. Springer New York: 1995; pp 15-38.

129. Wright, B. D.; Douglas, G. A. The rating scale model for objective measurement. University of Chicago: 1986.

130. Stenner, A. J.; Stone, M. H.; Burdick, D. S. The Concept of a Measurement Mechanism. *Rasch Measurement Transactions* 2009, 23, 1204-1206.

131. Stenner, A. J.; Fisher Jr, W. P.; Stone, M. H.; Burdick, D. S. Causal Rasch models. *Frontiers in psychology* 2013, 4.

132. Stenner, A. J.; Smith, M. I.; Burdick, D. S. Toward a Theory of Construct Definition. *Journal of Educational Measurement* 1983, 20, 305-316.

133. Stenner, A.; Burdick, H.; Sanford, E.; Burdick, D. How accurate are Lexile text measures? *Journal of Applied Measurement* 2005, 7, 307-322.

134. Burdick, H.; Stenner, A. J. Theoretical prediction of test items. *Rasch Measurement Transactions* 1996, 10, 475.

135. Stenner, A. J.; Burdick, D. S. The Objective Measurement of Reading Comprehension: In Response to Technical Questions Raised by the California Department of Education Technical Study Group. MetaMetrics: Durham, NC, 1997.

136. Masters, G. A rasch model for partial credit scoring. *Psychometrika* 1982, 47, 149-174.

137. Andrich, D. An Expanded Derivation of the Threshold Structure of the Polytomous Rasch Model That Dispels Any “Threshold Disorder Controversy”. *Educational and Psychological Measurement* 2013, 73, 78-124.

138. Andrich, D. A Rating Formulation for Ordered Response Categories. *Psychometrika* 1978, 43, 561-573.

139. Linacre, J. M.; Wright, B. D. Construction of measures from many-facet data. *Journal of Applied Measurement* 2002.

140. Lunz, M. E.; Wright, B. D.; Linacre, J. M. Measuring the Impact of Judge Severity on Examination Scores. *Applied Measurement in Education* 1990, 3, 331-345.

141. Kelderman, H. Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement* 1996, 20, 155-168.

142. Fischer, G. H. The linear logistic test model as an instrument in educational research. *Acta psychologica* 1973, 37, 359-374.

143. Fischer, G. H. The linear logistic test model. In *Rasch models*, Springer: 1995; pp 131-155.

144. Kubinger, K. D. Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement* 2009, 69, 232-244.

145. Tatsuoka, K. K. Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement* 1983, 20, 345-354.

146. Barnes, T. In *The q-matrix method: Mining student response data for knowledge*, American Association for Artificial Intelligence 2005 Educational Data Mining Workshop, 2005.

147. Buck, G.; VanEssen, T.; Tatsuoka, K.; Kostin, I.; Lutz, D.; Phelps, M. Development Selection and Validation of a set of Cognitive and Linguistic Attributes for the SAT I Verbal: Analogy Section1. *ETS Research Report Series* 1998, 1998, i-25.

148. Li, H.; Suen, H. K. Constructing and Validating a Q-Matrix for Cognitive Diagnostic Analyses of a Reading Test. *Educational Assessment* 2013, 18, 1-25.

149. Rupp, A. A.; Templin, J. L. Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art. *Measurement: Interdisciplinary Research and Perspectives* 2008, 6, 219-262.

150. Adams, R. J.; Wilson, M.; Wang, W.-c. The multidimensional random coefficients multinomial logit model. *Applied psychological measurement* 1997, 21, 1-23.

151. Wu, M. L.; Adams, R. J.; Wilson, M. R. *ACER ConQuest: Generalised item response modelling software*. ACER press: 1998.

152. Briggs, D. C.; Wilson, M. An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement* 2003, 4, 87-100.

153. Linacre, J. M. *Winsteps® Rasch measurement computer program*, Version 3.80.0; Beaverton, Oregon, 2013.

154. Linacre, J. M. *Facets computer program for many-facet Rasch measurement*, 3.71.4; Winsteps.com: Beaverton, Oregon, 2013.

155. Kline, T. L.; Schmidt, K. M.; Bowles, R. P. Using LinLog and FACETS to model item components in the LLTM. *J Appl Meas* 2006, 7, 74-91.

156. Wu, M.; Adams, R.; Wilson, M. *ConQuest: Multiaspect test software*, Australian Council for Educational Research: Camberwell, Vic., 1997.

157. Linacre, J. M. PROX with missing data, or known item or person measures. *Rasch Measurement Transactions* 1994, 8, 378.

158. Cohen, L. Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology* 1979, 32, 113-120.

159. Andrich, D. Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical care* 2004, 42, I-7.

160. Andrich, D. Understanding Rasch measurement: Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *Journal of Applied Measurement* 2002.

161. Smith, R. M. Fit analysis in latent trait measurement models. *J Appl Meas* 2000, 1, 199-218.

162. Curtis, D. D. Person misfit in attitude surveys: influences, impacts and implications. 2004.

163. Hohensinn, C.; Kubinger, K. D.; Reif, M. On robustness and power of the likelihood-ratio test as a model test of the linear logistic test model. *J Appl Meas* 2014, 15, 252-66.

164. Wright, B. D. Estimating Rasch measures for extreme scores. *Rasch Measurement Transactions* 1998, 12, 632-633.

165. Linacre, J. M. Disconnected Subsets, Guttman Patterns and Data Connectivity. *Rasch Measurement Transactions* 2013, 27, 1415-1417.

166. Wolfe, E. W. Equating and item banking with the Rasch model. *J Appl Meas* 2000, 1, 409-34.

167. Skaggs, G.; Wolfe, E. W. Equating designs and procedures used in Rasch scaling. *J Appl Meas* 2010, 11, 182-95.

168. Shaffer, J. P. Multiple Hypothesis Testing. *Annual Review of Psychology* 1995, 46, 561-584.

169. Fischer, H. *A history of the central limit theorem: from classical to modern probability theory*. Springer: 2010.

170. Arnold, B. C.; Groeneveld, R. A. Measuring Skewness with Respect to the Mode. *The American Statistician* 1995, 49, 34-38.

171. Balanda, K. P.; MacGillivray, H. Kurtosis: a critical review. *The American Statistician* 1988, 42, 111-119.

172. Cochran, W. G. *Sampling Techniques, 3Rd Edition*. Wiley India Pvt. Limited: 2007.

173. Boos, D. D.; Hughes-Oliver, J. M. How Large Does n Have to be for Z and t Intervals? *The American Statistician* 2000, 54, 121-128.

174. Kolmogorov, A. Sulla determinazione empirica di una leggi di distribuzione," G. Inst. Ital. Attuari, vol. 4, 1933, translated in English in Breakthroughs in Statistics, by Kotz and Johnson. Springer-Verlag: 1992.

175. Smirnov, N. Table for Estimating the Goodness of Fit of Empirical Distributions. 1948, 279-281.

176. Lilliefors, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 1967, 62, 399-402.

177. Shapiro, S. S.; Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* 1965, 591-611.

178. Wilk, M.; Gnanadesikan, R. Probability Plotting Methods for the Analysis of Data. *Biometrika* 1968, 1-17.

179. Ghasemi, A.; Zahediasl, S. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism* 2012, 10, 486-489.

180. Razali, N. M.; Wah, Y. B. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics* 2011, 2, 21-33.

181. Brown, L. D.; Cai, T. T.; DasGupta, A. Interval Estimation for a Binomial Proportion. 2001, 101-133.

182. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927, 22, 209-212.

183. Zimmerman, D. W. A note on preliminary tests of equality of variances. *British Journal of Mathematical & Statistical Psychology* 2004, 57, 173-181.

184. Zimmerman, D. W.; Zumbo, B. D. Hazards in Choosing Between Pooled and Seperate-Variances t Tests. *Psicológica* 2009, 30, 371-390.

185. Wilson, E. B.; Hilferty, M. M. The Distribution of Chi-Square. *Proceedings of the National Academy of Sciences of the United States of America* 1931, 17, 684-688.

186. Brown, G. W.; Mood, A. M. In *On median tests for linear hypotheses*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, The Regents of the University of California: 1951.

187. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 1922, 87-94.

188. Dodge, Y. Simple Linear Regression. In *The Concise Encyclopedia of Statistics*, Springer: New York, 2008; pp 491-497.

189. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficients in Samples from an Indefinitely Large Population. *Biometrika* 1915, 10, 507-521.

190. Fisher, R. A. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *JSTOR* 1921, 1, 3-32.

191. Clogg, C. C.; Petkova, E.; Haritou, A. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology* 1995, 100, 1261-1293.

192. Cohen, A. Comparing Regression Coefficients Across Subsamples. *Sociological Methods & Research* 1983, 12, 77-94.

193. Jolliffe, I. *Principal component analysis*. Wiley Online Library: 2002.

194. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 1987, 2, 37-52.

195. Abdi, H.; Williams, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010, 2, 433-459.

196. Thompson, B. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association: 2004.

197. Bryant, F. B.; Yarnold, P. R. Principal-components analysis and exploratory and confirmatory factor analysis. 1995.

198. Harman, H. H. Modern factor analysis. 1960.

199. Fabrigar, L. R.; Wegener, D. T.; MacCallum, R. C.; Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 1999, 4, 272-299.

200. Kendall, M. G.; Lawley, D. N. The Principles of Factor Analysis. *Journal of the Royal Statistical Society. Series A (General)* 1956, 119, 83-84.

201. Suhr, D. D. Principal component analysis vs. exploratory factor analysis.

202. Bandalos, D. L.; Boehm-Kaufman, M. R. Four common misconceptions in exploratory factor analysis. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* 2009, 61-87.

203. Rao, C. R. Estimation and tests of significance in factor analysis. *Psychometrika* 1955, 20, 93-111.

204. Thurstone, L. L. Multiple factor analysis. 1947.

205. Kaiser, H.; Caffrey, J. Alpha factor analysis. *Psychometrika* 1965, 30, 1-14.

206. Guttman, L. Image theory for the structure of quantitative variates. *Psychometrika* 1953, 18, 277-296.

207. Dillon, W. R.; Kumar, A.; Mulani, N. Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin* 1987, 101, 126.

208. Jöreskog, K. G. Efficient estimation in image factor analysis. *Psychometrika* 1969, 34, 51-75.

209. Zwick, W. R.; Velicer, W. F. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin* 1986, 99, 432-442.

210. Kaiser, H. F. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* 1960, 20, 141-151.

211. Cattell, R. B. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research* 1966, 1, 245-276.

212. Darton, R. A. Rotation in factor analysis. *The statistician* 1980, 167-194.

213. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 1958, 23, 187-200.

214. Neuhaus, J. O.; Wrigley, C. The Quartimax Method. *British Journal of Statistical Psychology* 1954, 7, 81-91.

215. Jennrich, R. I. Admissible values of γ in direct oblimin rotation. *Psychometrika* 1979, 44, 173-177.

216. Hendrickson, A. E.; White, P. O. Promax: A Quick Method for Rotation to Oblique Simple Structure. *British Journal of Statistical Psychology* 1964, 17, 65-70.

217. Dziuban, C. D.; Shirkey, E. C. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin* 1974, 81, 358.

218. Kaiser, H. F. A second generation little jiffy. *Psychometrika* 1970, 35, 401-415.

219. Cerny, B. A.; Kaiser, H. F. A Study Of A Measure Of Sampling Adequacy For Factor-Analytic Correlation Matrices. *Multivariate Behavioral Research* 1977, 12, 43-47.

220. Bartlett, M. S. Tests of Significance in Factor Analysis. *British Journal of Statistical Psychology* 1950, 3, 77-85.

221. Bartlett, M. S. A Note on the Multiplying Factors for Various χ^2 Approximations. *Journal of the Royal Statistical Society. Series B (Methodological)* 1954, 16, 296-298.

222. Linacre, J. M. Communicating Examinee Measures as Expected Ratings. *Rasch Measurement Transactions* 1997, 11, 550-551.

223. Olsson, U.; Drasgow, F.; Dorans, N. The polyserial correlation coefficient. *Psychometrika* 1982, 47, 337-347.

224. Schulz, M. Standardization of mean-squares. *Rasch Measurement Transactions* 2002, 16, 879.

225. Smith, R. M.; Suh, K. K. Rasch fit statistics as a test of the invariance of item parameter estimates. *J Appl Meas* 2003, 4, 153-63.

226. Gustafsson, J. E. Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology* 1980, 33, 205-233.

227. Wright, B. D.; Linacre, J. M.; Gustafson, J.; Martin-Lof, P. Reasonable mean-square fit values. *Rasch measurement transactions* 1994, 8, 370.

228. Linacre, J. M. Rasch Power Analysis: Size vs. Significance: Standardized Chi-Square Fit Statistic. *Rasch Measurement Transactions* 2003, 17, 918.

229. Masters, G. N. Item discrimination: When more is worse. *Journal of Educational Measurement* 1988, 25, 15-29.

230. Andersen, E. B. A goodness of fit test for the rasch model. *Psychometrika* 1973, 38, 123-140.

231. Douglas, G. Issues in the Fit of Data to Psychometric Models. *Education Research and Perspectives* 1982, 9, 32-43.

232. Linacre, J. M. Data Variance: Explained, Modeled and Empirical. *Rasch Measurement Transactions* 2003, 17, 942-943.

233. Wright, B. D.; Masters, G. N. Number of Person or Item Strata: (4*Separation + 1)/3. *Rasch Measurement Transactions* 2002, 16, 888.

234. Wright, B. Reliability and separation. *Rasch Measurement Transactions* 1996, 9, 472.

235. Andrich, D. An Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. *Education Research and Perspectives* 1982, 9, 95-104.

236. Fisher, W. P., Jr. The Cash Value of Reliability. *Rasch Measurement Transactions* 2008, 22, 1160-1163.

237. Linacre, J. M. Estimating 50% Cumulative Probability Thresholds. *Rasch Measurement Transactions* 2003, 16, 901.

238. Linacre, J. M. Predicting Measures from Rating Scale or Partial Credit Categories for Samples and Individuals. *Rasch Measurement Transactions* 2004, 18, 972.

239. Linacre, J. M. Optimizing rating scale category effectiveness. *Journal of Applied Measurement* 2002, 3, 85-106.

240. Mantel, N. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* 1963, 58, 690-700.

241. Mantel, N.; Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. 1959.

242. Donoghue, J. R.; Allen, N. L. Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics* 1993, 18, 131-154.

243. Linacre, J. M.; Wright, B. D. Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions* 1989, 3, 52-53.

244. Engelhard Jr., G. An Empirical Comparison of Mantel-Haenszel and Rasch Procedures for Studying Differential Item Functioning on Teacher Certification Tests. 1989.

245. Schulz, E. M. DIF detection: Rasch versus Mantel-Haenszel. *Rasch Measurement Transactions* 1990, 4, 107.

246. Schulz, E. M.; C., P.; W.K., R.; B.D., W. An Empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In *Objective measurement: Theory into practice*, Engelhard Jr., G.; Wilson, M., Eds. Ablex: Norwood, NJ, 1996; Vol. 3, pp 65-82.

247. Smith, R.; Hedges, L. Comparison of likelihood ratio χ^2 and pearsonian χ^2 tests of fit in the rasch model. *Education Research and Perspectives* 1982, 9, 1-44.

248. Smith, R. M.; Plackner, C. The family approach to assessing fit in Rasch measurement. *Journal of applied measurement* 2008, 10, 424-437.

249. Andrich, D.; Hagquist, C. Real and Artificial Differential Item Functioning. *Journal of Educational and Behavioral Statistics* 2012, 37, 387-416.

250. Wilks, S. S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. 1938, 60-62.

251. Alexandrowicz, R. W. Statistical and practical significance of the Likelihood Ratio Test of the linear logistic test model versus the Rasch model. *Educational Research and Evaluation* 2011, 17, 335-350.

252. Akaike, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 1974, 19, 716-723.

253. Hurvich, C. M.; Tsai, C.-L. Regression and Time Series Model Selection in Small Samples. *Biometrika* 1989, 76, 297-307.

254. Burnham, K. P.; Anderson, D. R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 2004, 33, 261-304.

255. Kullback, S.; Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* 1951, 22, 79-86.

256. Priest, S. J.; Pyke, S. M.; Williamson, N. M. Student Perceptions of Chemistry Experiments with Different Technological Interfaces: A Comparative Study. *Journal of Chemical Education* 2014, 91, 1787-1795.

257. Johnstone, A. H. Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning* 1991, 7, 75-83.

258. Johnstone, A. H. The development of chemistry teaching: A changing response to changing demand. *Journal of Chemical Education* 1993, 70, 701.

259. Gilbert, J. K.; Treagust, D. *Multiple representations in chemical education*. Springer Netherlands: 2009.

260. Johnstone, A. H. You Can't Get There from Here! *Journal of Chemical Education* 2010, 87, 22-29.

261. Cantu, L. L.; Herron, J. D. Concrete and formal piagetian stages and science concept attainment. *Journal of Research in Science Teaching* 1978, 15, 135-143.

262. Linn, M. Technology and science education: Starting points, research programs, and trends. *International Journal of Science Education* 2003, 25, 727-758.

263. Durick, M. A. The Study of Chemistry by Guided Inquiry Method Using Microcomputer-Based Laboratories. *Journal of Chemical Education* 2001, 78, 574.

264. Nyasulu, F.; Barlag, R. Thermokinetics: Iodide-Catalyzed Decomposition Kinetics of Hydrogen Perioxide. An Initial-Rate Approach. *Journal of Chemical Education* 2009, 86, 1231.

265. Struck, W.; Yerrick, R. The Effect of Data Acquisition-Probeware and Digital Video Analysis on Accurate Graphical Representation of Kinetics in a High School Physics Class. *Journal of Science Education and Technology* 2010, 19, 199-211.

266. Amrani, D. Determination of absolute zero using a computer-based laboratory. *Physics Education* 2007, 42, 304.

267. Vannatta, M. W.; Richards-Babb, M.; Solomon, S. D. Personal Multifunctional Chemical Analysis Systems for Undergraduate Chemistry Laboratory Curricula. *Journal of Chemical Education* 2010, 87, 770-772.

268. Chebolu, V.; Storandt, B. C. Stoichiometry of the Reaction of Magnesium with Hydrochloric Acid. *Journal of Chemical Education* 2003, 80, 305.

269. Choi, M. M. F.; Wong, P. S.; Yiu, T. P.; Mark, C. Application of datalogger in observing photosynthesis. *Journal of Chemical Education* 2002, 79, 980.

270. Thornton, R. K.; Sokoloff, D. R. Learning motion concepts using real-time microcomputer-based laboratory tools. *American Journal of Physics* 1990, 58, 858-867.

271. Mokros, J. R.; Tinker, R. F. The impact of microcomputer-based labs on children's ability to interpret graphs. *Journal of Research in Science Teaching* 1987, 24, 369-383.

272. Trumper, R.; Gelbman, M. A Microcomputer-Based Contribution to Scientific and Technological Literacy. *Journal of Science Education and Technology* 2001, 10, 213-221.

273. Linn, M. C.; Layman, J. W.; Nachmias, R. Cognitive consequences of microcomputer-based laboratories: Graphing skills development. *Contemporary Educational Psychology* 1987, 12, 244-253.

274. Rogers, L. T. Computer as an Aid for Exploring Graphs. *School Science Review* 1995, 76, 31-39.

275. Redish, E. F.; Saul, J. M.; Steinberg, R. N. On the effectiveness of active-engagement microcomputer-based laboratories. *American Journal of Physics* 1997, 65, 45-54.

276. BouJaoude, S.; Jurdak, M. Integrating physics and math through microcomputer-based laboratories (MBL): effects on discourse type, quality, and mathematization. *Int J of Sci and Math Educ* 2010, 8, 1019-1047.

277. Metcalf, S.; Tinker, R. Probeware and Handhelds in Elementary and Middle School Science. *Journal of Science Education and Technology* 2004, 13, 43-49.

278. Stringfield, J. K. Using Commercially Available, Microcomputer-Based Labs in the Biology Classroom. *The American Biology Teacher* 1994, 56, 106-108.

279. Russell, D. W.; Lucas, K. B.; McRobbie, C. J. Role of the microcomputer-based laboratory display in supporting the construction of new understandings in thermal physics. *Journal of Research in Science Teaching* 2004, 41, 165-185.

280. Solomon, J.; Bevan, R.; Frost, A.; Reynolds, H.; Summers, M.; Zimmerman, C. Can pupils learn through their own movements? A study of the use of a motion sensor interface. *Physics Education* 1991, 26, 345.

281. Trumper, R. Learning Kinematics with a V-Scope: A Case Study. *Journal of Computers in Mathematics and Science Teaching* 1997, 16, 91-110.

282. Thornton, R. K. Tools for scientific thinking-microcomputer-based laboratories for physics teaching. *Physics Education* 1987, 22, 230.

283. Espinoza, F. The Use of Graphical Analysis with Microcomputer-Based Laboratories to Implement Inquiry as the Primary Mode of Learning Science. *Journal of Educational Technology Systems* 2007, 35, 315-335.

284. Espinoza, F.; Quarless, D. An Inquiry-Based Contextual Approach as the Primary Mode of Learning Science with Microcomputer-Based Laboratory Technology. *Journal of Educational Technology Systems* 2009, 38, 407-426.

285. Weller, H. G. Assessing the impact of computer-based learning in science. *Journal of Research on Computing in Education* 1996, 28, 461.

286. Rogers, L.; Wild, P. Data-logging: effects on practical science. *Journal of Computer Assisted Learning* 1996, 12, 130-145.

287. Newton, L. R. Data-logging in practical science: research and reality. *International Journal of Science Education* 2000, 22, 1247-1259.

288. Adams, D. D.; Shrum, J. W. The effects of microcomputer-based laboratory exercises on the acquisition of line graph construction and interpretation skills by high school biology students. *Journal of Research in Science Teaching* 1990, 27, 777-787.

289. Nakhleh, M. B.; Krajcik, J. S. *Journal of Research in Science Teaching* 1994, 31, 1077.

290. Atar, H. Y. Examining students' and teacher's perceptions of microcomputer base laboratories(MBLs) in a high school chemistry classes. Masters Thesis, Florida State University, 2001.

291. Tortosa, M. The use of microcomputer based laboratories in chemistry secondary education: Present state of the art and ideas for research-based practice. *Chemistry Education Research and Practice* 2012, 13, 161-171.

292. Thomas, G. P.; Man-wai, P. F.; Po-keung, E. T. Students' perceptions of early experiences with microcomputer-based laboratories (MBL). *British Journal of Educational Technology* 2004, 35, 669-671.

293. Aksela, M. Supporting meaningful chemistry learning and higher-order thinking through computer-assisted inquiry: A design research approach. Academic Dissertation, University of Helsinki, 2005.

294. Roth, W.-M.; Woszczyna, C.; Smith, G. Affordances and constraints of computers in science education. *Journal of Research in Science Teaching* 1996, 33, 995-1017.

295. Atar, H. Y. *Chemistry Students' Challenges in Using MBL's in Science Laboratories*. Distributed by ERIC Clearinghouse: Washington, D.C., 2002.

296. Kay, R.; Knaack, L. Evaluating the Use of Learning Objects for Secondary School Science. *Journal of Computers in Mathematics and Science Teaching* 2007, 26, 261-289.

297. Lowry, R. VassarStats: Website for Statistical Computation. www.vassarstats.net.

298. Linacre, J. M. Sample Size and Item Calibration Stability. *Rasch Measurement Transactions* 1994, 7, 328.

299. Riley, B. Considering Large Group Differences in Ability in DIF Analysis. *Rasch Measurement Transactions* 2011, 251, 1326.

300. Linacre, J. M. Investigating rating scale category utility. *Journal of outcome measurement* 1999, 3, 103-22.

301. Rice, J. A. *Mathematical Statistics and Data Analysis*. 3 ed.; BROOKS/ COLE CENGAGE Learning: 2007.

302. *MATLAB R2011b*, 7.13.0.564; The Mathworks, Inc.: Natick, Massachusetts, United States, 2011.

303. Muthén, L. K.; Muthén, B. O. How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal* 2002, 9, 599-620.

304. Taber, K. S. Revisiting the chemistry triplet: drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education. *Chemistry Education Research and Practice* 2013, 14, 156-168.

305. Johnstone, A. H. Teaching of Chemistry - Logical or Psychological? *Chemistry Education: Research and Practice in Europe* 2000, 1, 9-15.

306. Johnstone, A. H. Macro and microchemistry. *School Science review* 1982, 64, 377-379.

307. Piaget, J.; Inhelder, B. *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Routledge: 2013; Vol. 84.

308. Piaget, J. *The Science of Education and the Psychology of the Child*. Orion Press: New York, 1970.

309. Johnstone, A. H. Chemical education research: where from here. *University Chemistry* 2000.

310. Dalton, R.; Tasker, R. In *Research into practice: Using molecular representations as a learning strategy in chemistry*.

311. Justi, R. Teaching and learning about chemical change. *Chemical Education: Towards Research-based Practice* 2002, 269-270.

312. Weerawardhana, A.; Ferry, B.; Brown, C. A. In *Developing conceptual understanding of chemical equilibrium through the use of computer-based visualization software*, 9th International Conference on Sri Lanka Studies, 2003; pp 28-30.

313. Hodson, D. A critical look at practical work in school science. *School Science Review* 1990, 71, 33-40.

314. Vygotsky, L. S. *Mind in society: The development of higher psychological processes*. Harvard university press: 1980.

315. Vygotsky, L. Interaction between learning and development. *Readings on the development of children* 1978, 23, 34-41.

316. Sharan, S. Cooperative learning in small groups: Recent methods and effects on achievement, attitudes, and ethnic relations. *Review of educational research* 1980, 50, 241-271.

317. Curşeu, P. L.; Pluut, H. Student groups as learning entities: The effect of group diversity and teamwork quality on groups' cognitive complexity. *Studies in Higher Education* 2013, 38, 87-103.

318. Macpherson, A. Cooperative Learning Group Activities for College Courses. 2015.

319. Kuhlthau, C. C.; Maniotes, L. K.; Caspari, A. K. *Guided Inquiry: Learning in the 21st Century: Learning in the 21st Century*. ABC-CLIO: 2015.

320. Moog, R. *Process oriented guided inquiry learning*. Washington University Libraries: 2014.

321. Williamson, N. M.; Huang, D. M.; Bella, S. G.; Metha, G. F. Guided Inquiry Learning in an Introductory Chemistry Course. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)* 2016, 23.

322. Savery, J. R. Overview of problem-based learning: Definitions and distinctions. *Essential readings in problem-based learning: Exploring and extending the legacy of Howard S. Barrows* 2015, 5-15.

323. Savery, J. R.; Duffy, T. M. Problem based learning: An instructional model and its constructivist framework. *Educational technology* 1995, 35, 31-38.

324. Barrows, H. S. A taxonomy of problem-based learning methods. *Medical education* 1986, 20, 481-486.

325. Allen, D. E.; Duch, B. J.; Groh, S. E. The power of problem-based learning in teaching introductory science courses. *New directions for teaching and learning* 1996, 1996, 43-52.

326. Dolmans, D. H.; Wolfhagen, I. H.; Van Der Vleuten, C. P.; Wijnen, W. H. Solving problems with group work in problem-based learning: hold on to the philosophy. *Medical education* 2001, 35, 884-889.

327. Carpendale, J. I. M.; Müller, U.; Bibok, M. B. Piaget's Theory of Cognitive Development. In *Encyclopedia of Educational Psychology*, Sage Publications, Inc: Thousand Oaks, CA, 2008; pp 798-804.

328. Biggs, J. Enhancing teaching through constructive alignment. *Higher education* 1996, 32, 347-364.

329. Biggs, J. Aligning teaching and assessment to curriculum objectives. *Imaginative Curriculum Project, LTSN Generic Centre* 2003, 12.

330. Schunk, D. H.; Meece, J. L. *Student perceptions in the classroom*. Routledge: 1992.

7 Supporting Information

7.1 Information provided to participants

7.1.1 Excluding the option to provide student identification number

PARTICIPANT INFORMATION SHEET

Project Title: ***Representations of scientific concepts and learning experience of first year undergraduate university students***

THE UNIVERSITY of ADELAIDE

This project will gather information regarding student perceptions of laboratory exercises, and hence has the ability to evaluate the efficacy of each experiment as a learning exercise from the student perspective. Benefits of this investigation include contribution to knowledge of effective pedagogy in sciences and also contributions to the improvement and/or sustained quality of first year undergraduate laboratory programs in chemistry, biology and physics courses at the University of Adelaide. Researchers involved in the project include Mr. Sam Priest, A/Prof. Simon Pyke, Dr. Natalie Williamson and Dr. John Willison.

The aim of this project is to investigate factors contributing to positive learning experience from a student perspective. Of primary interest is the investigation of any trends in student perception of experiments as related to the modes in which concepts are represented. Specifically, macro (accessible to the senses), sub-micro (not accessible to the senses), and symbolic representations (those using equations, symbols, diagrams etc) of concepts will be considered.

Data for this project is collected via surveys distributed during laboratory sessions. These surveys are modelled on those used for the ALTC funded Advancing Science by Enhancing Learning in the Laboratory (ASELL) project (see http://www.asell.org), which have been used at numerous institutions Australia wide for more than 10 years. We ask that you complete these surveys at the end of your laboratory sessions, and that any feedback you provide is honest.

Your participation is entirely voluntary and your feedback is anonymous. Neither whether you choose to respond to these surveys, nor any feedback you provide, will have any influence on your progress, results or grades in any subject. Researchers gathering the data requested will have no direct role in assessment of the laboratory activities concerned. Filling out the survey will constitute consent for its use for research purposes.

If you are willing to participate, please complete the surveys made available to you during your practical sessions, upon completion of your experiment. Thank you for your cooperation. Please direct any queries regarding this research to one of the contacts below:

Mr. Sam Priest

BSc. (Hons I), PhD student

School of Chemistry and physics

Email: [redacted]

A/Prof. Simon Pyke

Associate Dean (Learning & Quality) – Faculty of Sciences

School of Chemistry and Physics

Tel: [redacted]

Email: [redacted]

Dr. Natalie Williamson

First Year Coordinator (Discipline of Chemistry)

School of Chemistry and Physics

Tel: [redacted]

Email: [redacted]

Dr. John Willison

Senior Lecturer

School of Education

Tel: [redacted]

Email: [redacted]

7.1.2 Including the option to provide student identification number

PARTICIPANT INFORMATION SHEET

Project Title: ***Representations of scientific concepts and learning experience of first year undergraduate university students***

This project will gather information regarding student perceptions of laboratory exercises, and hence has the ability to evaluate the efficacy of each experiment as a learning exercise from the student perspective. Benefits of this investigation include contribution to knowledge of effective pedagogy in sciences and also contributions to the improvement and/or sustained quality of first year undergraduate laboratory programs in chemistry, biology and physics courses at the University of Adelaide. Researchers involved in the project include Mr. Sam Priest, A/Prof. Simon Pyke, Dr. Natalie Williamson and Dr. John Willison.

The aim of this project is to investigate factors contributing to positive learning experience from a student perspective. Of primary interest is the investigation of any trends in student perception of experiments as related to the modes in which concepts are represented. Specifically, macro (accessible to the senses), sub-micro (not accessible to the senses), and symbolic representations (those using equations, symbols, diagrams etc) of concepts will be considered.

Data for this project is collected via surveys distributed during laboratory sessions. These surveys are modelled on those used for the ALTC funded Advancing Science by Enhancing Learning in the Laboratory (ASELL) project (see http://www.asell.org), which have been used at numerous institutions Australia wide for more than 10 years. We ask that you complete these surveys at the end of your laboratory sessions, and that any feedback you provide is honest.

Your participation is entirely voluntary and your feedback is anonymous. Neither your response to these surveys, nor any feedback you provide, will have any influence on your progress, results or grades in any subject. Researchers gathering the data requested will have no direct role in assessment of the laboratory activities concerned. Filling out the survey will constitute consent for its use for research purposes.

The surveys provide you with the opportunity to include your student ID number. This is entirely optional, and you may elect to still complete a survey without including this. Should you choose to provide your student identification number, this is at no stage intended to be linked to your name. The optional provision of your identification number is included solely for the purpose of identifying surveys which have been completed by the same person, and also relating perceptions of experiments to the different subjects students study. This facilitates investigation regarding whether perceptions and learning experiences associated with one science discipline area influence those of another.

If you are willing to participate, please complete the surveys made available to you during your practical sessions, upon completion of your experiment. Thank you for your cooperation. Please direct any queries regarding this research to one of the contacts below:

Mr. Sam Priest

BSc. (Hons I), PhD student

School of Chemistry and physics

Email: [redacted]

A/Prof. Simon Pyke

Associate Dean (Learning & Quality) – Faculty of Sciences

School of Chemistry and Physics

Tel: [redacted]

Email: [redacted]

Dr. Natalie Williamson

First Year Coordinator (Discipline of Chemistry)

School of Chemistry and Physics

Tel: [redacted]

Email: [redacted]

Dr. John Willison

Senior Lecturer

School of Education

Tel: [redacted]

Email: [redacted]

7.2 Supporting information for sections 3.1 and 3.2

7.2.1 Responses to Likert-type items

For each of the following tables, a number of symbols are used. Response categories A through E represent the most positive to least positive response options respectively. The total number of responses received for that survey item (the sample size) is labelled as n ; m is the mean response score for that item; s is the standard deviation of response scores about the mean score; SE(m) is the standard error in the mean value, calculated as the standard deviation (s) divided by the square root of the sample size (n).

Table S 1: Likert type item response data for the Biological Buffers experiment

2011 responses (datalogger interface)										
Survey Item		**Response frequencies**					**Count**	**Sampling Distribution**		
		A	B	C	D	E	*n*	*m*	*s*	*Se (m)*
1	data interpretation skills	20	86	18	3	6	133	0.83	0.87	0.08
2	laboratory skills	21	80	23	4	6	134	0.79	0.90	0.08
3	interest	19	46	50	13	6	134	0.44	1.00	0.09
4	clear assessment	26	68	35	3	2	134	0.84	0.81	0.07
5	clear expected learning	30	67	26	8	3	134	0.84	0.92	0.08
6	increased understanding	28	70	27	5	4	134	0.84	0.90	0.08
7	background information	36	61	24	9	5	135	0.84	1.01	0.09
8	demonstrators	67	51	16	1	1	136	1.34	0.77	0.07
9	clear procedure	35	68	21	6	5	135	0.90	0.96	0.08
10	relevance to discipline	51	64	17	3	1	136	1.18	0.79	0.07
11	benefit of teamwork	68	48	13	3	3	135	1.30	0.90	0.08
12	responsibility for own learing	31	72	26	2	3	134	0.94	0.83	0.07
13	time availability	2	9	113	8	1	133	0.02	0.47	0.04
14	overall learning experience	11	76	32	12	3	134	0.60	0.85	0.07
2012 responses (laptop interface)										
Survey Item		**Response frequencies**					**Count**	**Sampling Distribution**		
		A	B	C	D	E	*n*	*m*	*s*	*Se (m)*
1	data interpretation skills	15	53	12	0	0	80	1.04	0.58	0.07
2	laboratory skills	19	45	14	2	0	80	1.01	0.72	0.08
3	interest	13	33	28	4	2	80	0.64	0.90	0.10
4	clear assessment	19	45	11	4	1	80	0.96	0.83	0.09
5	clear expected learning	18	46	8	4	4	80	0.88	0.99	0.11
6	increased understanding	15	43	18	1	2	79	0.86	0.83	0.09
7	background information	18	35	17	7	3	80	0.73	1.03	0.12
8	demonstrators	45	28	4	2	1	80	1.43	0.81	0.09
9	clear procedure	13	29	21	13	4	80	0.43	1.10	0.12
10	relevance to discipline	28	38	12	2	0	80	1.15	0.76	0.09
11	benefit of teamwork	48	24	6	2	0	80	1.48	0.75	0.08
12	responsibility for own learing	21	42	15	2	0	80	1.03	0.75	0.08
13	time availability	1	11	65	3	0	80	0.13	0.46	0.05
14	overall learning experience	12	50	15	2	1	80	0.88	0.74	0.08

Table S 2: Likert type item response data for the Vapour Pressure experiment

2011 responses (datalogger interface)										
Survey Item		Response frequencies					Count	Sampling Distribution		
		A	B	C	D	E	*n*	*m*	*s*	*Se (m)*
1	data interpretation skills	10	37	18	11	8	84	0.36	1.15	0.13
2	laboratory skills	16	32	16	14	6	84	0.45	1.19	0.13
3	interest	7	22	24	13	18	84	-0.15	1.27	0.14
4	clear assessment	7	33	27	9	8	84	0.26	1.08	0.12
5	clear expected learning	9	35	20	8	11	83	0.28	1.19	0.13
6	increased understanding	12	32	22	8	10	84	0.33	1.20	0.13
7	background information	13	43	15	8	5	84	0.61	1.05	0.11
8	demonstrators	39	27	14	3	1	84	1.19	0.92	0.10
9	clear procedure	7	24	27	17	9	84	0.04	1.12	0.12
10	relevance to discipline	15	38	21	5	5	84	0.63	1.04	0.11
11	benefit of teamwork	47	25	8	3	1	84	1.36	0.89	0.10
12	responsibility for own learing	15	30	30	5	4	84	0.56	1.01	0.11
13	time availability	2	1	39	34	8	84	-0.54	0.78	0.09
14	overall learning experience	3	28	27	16	10	84	-0.02	1.08	0.12
2012 responses (laptop interface)										
Survey Item		Response frequencies					Count	Sampling Distribution		
		A	B	C	D	E	*n*	*m*	*s*	*Se (m)*
1	data interpretation skills	24	63	13	1	1	102	1.06	0.70	0.07
2	laboratory skills	25	58	16	3	1	103	1.00	0.78	0.08
3	interest	13	45	28	10	7	103	0.46	1.06	0.10
4	clear assessment	15	56	18	10	4	103	0.66	0.98	0.10
5	clear expected learning	21	59	17	5	1	103	0.91	0.81	0.08
6	increased understanding	26	56	15	5	1	103	0.98	0.83	0.08
7	background information	26	51	18	5	3	103	0.89	0.94	0.09
8	demonstrators	49	45	6	0	3	103	1.33	0.83	0.08
9	clear procedure	20	31	33	11	8	103	0.43	1.15	0.11
10	relevance to discipline	18	53	24	6	2	103	0.77	0.88	0.09
11	benefit of teamwork	56	36	5	4	1	102	1.39	0.83	0.08
12	responsibility for own learing	21	65	13	3	1	103	0.99	0.73	0.07
13	time availability	1	6	84	9	2	102	-0.05	0.51	0.05
14	overall learning experience	11	58	26	7	1	103	0.69	0.79	0.08

Table S 3: Likert-type item response data for the Copper(II) Ion Concentration experiment

2011 responses (datalogger interface)										
Survey Item		Response frequencies					Count	Sampling Distribution		
		A	B	C	D	E	*n*	*m*	*s*	*Se (m)*
1	data interpretation skills	24	78	23	0	0	125	1.01	0.62	0.06
2	laboratory skills	35	71	18	1	1	126	1.10	0.72	0.06
3	interest	25	62	30	7	2	126	0.80	0.88	0.08
4	clear assessment	40	60	23	3	0	126	1.09	0.77	0.07
5	clear expected learning	41	65	17	3	0	126	1.14	0.73	0.07
6	increased understanding	31	64	23	6	1	125	0.94	0.84	0.07
7	background information	39	65	20	2	0	126	1.12	0.72	0.06
8	demonstrators	80	39	5	1	0	125	1.58	0.61	0.05
9	clear procedure	54	55	13	3	0	125	1.28	0.75	0.07
10	relevance to discipline	34	58	30	4	0	126	0.97	0.80	0.07
11	benefit of teamwork	73	46	5	1	0	125	1.53	0.62	0.06
12	responsibility for own learing	34	67	23	1	1	126	1.05	0.75	0.07
13	time availability	8	14	102	2	0	126	0.22	0.58	0.05
14	overall learning experience	14	89	23	0	0	126	0.93	0.54	0.05
2012 responses (laptop interface)										
Survey Item		Response frequencies					Count	Sampling Distribution		
		A	B	C	D	E	*n*	*m*	*s*	*Se (m)*
1	data interpretation skills	24	68	28	0	0	120	0.97	0.66	0.06
2	laboratory skills	26	74	19	1	0	120	1.04	0.64	0.06
3	interest	31	59	23	7	0	120	0.95	0.83	0.08
4	clear assessment	41	57	19	3	0	120	1.13	0.77	0.07
5	clear expected learning	41	58	18	3	0	120	1.14	0.76	0.07
6	increased understanding	27	56	32	4	0	119	0.89	0.79	0.07
7	background information	49	59	9	3	0	120	1.28	0.71	0.07
8	demonstrators	81	31	7	0	1	120	1.59	0.68	0.06
9	clear procedure	47	59	13	1	0	120	1.27	0.68	0.06
10	relevance to discipline	30	67	19	4	0	120	1.03	0.74	0.07
11	benefit of teamwork	65	40	11	2	2	120	1.37	0.85	0.08
12	responsibility for own learing	28	61	29	1	0	119	0.97	0.72	0.07
13	time availability	2	23	94	1	0	120	0.22	0.47	0.04
14	overall learning experience	17	81	22	0	0	120	0.96	0.57	0.05

7.2.2 Comparative tests for the Biological Buffers experiment data

The total number of surveys collected for the Biological Buffers experiment in 2011 and 2012 were 136 and 80 surveys respectively. In this case, the significance level (α) is 0.05 and the number of statistical tests conducted (n) is 137.Therefore, p values below $\alpha/n = 3.65\times10^{-4}$ are shaded to indicate refutation of the relevant null hypothesis, controlling for family-wise error.

Quantitative comparisons in this section compare mean scored Likert-type item responses, using the t-test for unequal variances. $p < \alpha / n$ refutes the null hypothesis that mean scores are equal for the two data sets. Qualitative comparisons test the significance of the association between the data set sampled, and content and/or nature of the comments received using Fisher's exact test. $p < \alpha / n$ refutes the null hypothesis that response content is independent of the student data set sampled (data logger or laptop).

Table S 4: Quantitative comparisons for the Biological Buffers experiment

Survey item and topic		$m_{data\ logger}$	m_{laptop}	df	t	p
1	data interpretation skills	0.83	1.04	208.6	-2.03	4.34×10^{-2}
2	laboratory skills	0.79	1.01	194.7	-1.98	4.95×10^{-2}
3	interest	0.44	0.64	179.7	-1.48	1.40×10^{-1}
4	clear assessment	0.84	0.96	162.8	-1.02	3.08×10^{-1}
5	clear expected learning	0.84	0.88	156.8	-0.23	8.15×10^{-1}
6	increased understanding	0.84	0.86	174.8	-0.14	8.86×10^{-1}
7	background information	0.84	0.73	163.8	0.83	4.10×10^{-1}
8	demonstrators	1.34	1.43	159.7	-0.77	4.40×10^{-1}
9	clear procedure	0.90	0.43	148.7	3.23	1.52×10^{-3}
10	relevance to discipline	1.18	1.15	170.2	0.31	7.57×10^{-1}
11	benefit of teamwork	1.30	1.48	190.3	-1.57	1.18×10^{-1}
12	responsibility for own learing	0.94	1.03	180.3	-0.77	4.42×10^{-1}
13	time availability	0.02	0.13	168.7	-1.56	1.20×10^{-1}
14	overall learning experience	0.60	0.88	185.3	-2.52	1.26×10^{-2}

7.2.2.1 General perceptions of "Biological Buffers"

Table S 5: General nature of responses to item 15: "Did you enjoy doing the experiment? Why or why not?" for the Biological Buffers experiment

Are responses more positive?			Are responses more negative?		
	Data logger	Laptop		Data logger	Laptop
Positive	28	36	Negative	24	19
Not positive	23	18	Not negative	26	35
	p =	2.36×10^{-1}		p =	2.33×10^{-1}

Table S 6: Topic referenced in response to item 15: "Did you enjoy doing the experiment? Why or why not?" for the Biological Buffers experiment

		Data logger		Laptop		p
Code		Positive	Negative	Positive	Negative	
T	Time availability	0	1	4	0	2.00×10^{-1}
C	Relation to the course/ lectures	4	0	2	0	1.00×10^{0}
P	Aspects of the procedure	5	0	4	6	4.40×10^{-2}
M	Manual or answer book	0	2	0	7	1.00×10^{0}
I	level of interest	2	2	4	0	4.29×10^{-1}
R	results obtained	2	1	3	2	1.00×10^{0}
L	new learning achieved	4	0	3	0	1.00×10^{0}
E	equipment, apparatus or technology	4	21	2	4	5.67×10^{-1}
F	level of familiarity or relevance	0	0	3	0	1.00×10^{0}
U	level of understanding	6	0	3	3	1.82×10^{-1}
O	others in the lab (students/ demonstrators)	2	0	2	0	1.00×10^{0}
S	level of simplicity	3	0	8	0	1.00×10^{0}
X	uncategorised	7	0	9	3	2.63×10^{-1}

Table S 7: Content referenced in response to item 16: "What did you think was the main lesson to be learned from the experiment" for the Biological Buffers experiment

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
E	Equivalence points/ pKa	6	32	10	39	7.81×10^{-1}
R	Effective range of buffers	11	27	8	41	1.95×10^{-1}
H	Henderson Hasselbalch equation	6	32	4	45	3.22×10^{-1}
P	pH	7	31	11	38	7.91×10^{-1}
X	none of the above	17	21	25	24	6.66×10^{-1}

7.2.2.2 Positive perceptions of “Biological Buffers”

Table S 8: Reasons cited for enjoying the Biological buffers experiment considered amongst all comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p	
Code		Coded	Not coded	Coded	Not coded		
T	Time availability	0	51	4	50	1.18	$\times 10^{-1}$
C	Relation to the course/ lectures	4	47	2	52	4.28	$\times 10^{-1}$
P	Aspects of the procedure	5	46	4	50	7.37	$\times 10^{-1}$
M	Manual or answer book	0	51	0	54	1.00	$\times 10^{0}$
I	level of interest	2	49	4	50	6.79	$\times 10^{-1}$
R	results obtained	2	49	3	51	1.00	$\times 10^{0}$
L	new learning achieved	4	47	3	51	7.11	$\times 10^{-1}$
E	equipment, apparatus or technology	4	47	2	52	4.28	$\times 10^{-1}$
F	level of familiarity or relevance	0	51	3	51	2.43	$\times 10^{-1}$
U	level of understanding	6	45	3	51	3.11	$\times 10^{-1}$
O	others in the lab (students/ demonstrators)	2	49	2	52	1.00	$\times 10^{0}$
S	level of simplicity	3	48	8	46	2.03	$\times 10^{-1}$
X	uncategorised	7	44	9	45	7.88	$\times 10^{-1}$

Table S 9: Reasons cited for enjoying the Biological Buffers experiment considered only amongst other reasons cited for liking the experiment in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p	
Code		Coded	Not coded	Coded	Not coded		
T	Time availability	0	28	4	32	1.25	$\times 10^{-1}$
C	Relation to the course/ lectures	4	24	2	34	3.91	$\times 10^{-1}$
P	Aspects of the procedure	5	23	4	32	4.88	$\times 10^{-1}$
M	Manual or answer book	0	28	0	36	1.00	$\times 10^{0}$
I	level of interest	2	26	4	32	6.88	$\times 10^{-1}$
R	results obtained	2	26	3	33	1.00	$\times 10^{0}$
L	new learning achieved	4	24	3	33	6.89	$\times 10^{-1}$
E	equipment, apparatus or technology	4	24	2	34	3.91	$\times 10^{-1}$
F	level of familiarity or relevance	0	28	3	33	2.50	$\times 10^{-1}$
U	level of understanding	6	22	3	33	1.63	$\times 10^{-1}$
O	others in the lab (students/ demonstrators)	2	26	2	34	1.00	$\times 10^{0}$
S	level of simplicity	3	25	8	28	3.22	$\times 10^{-1}$
X	uncategorised	7	21	9	27	1.00	$\times 10^{0}$

Table S 10: Features cited as the most enjoyable and interesting aspects of the Biological Buffers experiment considered amongst all responses to item 17: "What aspects of the experiment did you find most enjoyable and interesting?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	0	31	0	36	1.00	$\times 10^{0}$
C	Relation to the course/ lectures	0	31	1	35	1.00	$\times 10^{0}$
P	Aspects of the procedure	8	23	17	19	8.23	$\times 10^{-2}$
M	Manual or answer book	1	30	0	36	4.63	$\times 10^{-1}$
I	level of interest	0	31	0	36	1.00	$\times 10^{0}$
R	results obtained	10	21	9	27	5.92	$\times 10^{-1}$
L	new learning achieved	0	31	0	36	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	7	24	10	26	7.80	$\times 10^{-1}$
F	level of familiarity or relevance	0	31	1	35	1.00	$\times 10^{0}$
U	level of understanding	0	31	0	36	1.00	$\times 10^{0}$
O	others in the lab (students/ demonstrators)	0	31	2	34	4.95	$\times 10^{-1}$
S	level of simplicity	0	31	0	36	1.00	$\times 10^{0}$
X	uncategorised	5	26	2	34	2.36	$\times 10^{-1}$

Table S 11: Features cited as the most enjoyable and interesting aspects of the Biological Buffers experiment considered amongst only other positive responses to item 17: "What aspects of the experiment did you find most enjoyable and interesting?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	0	25	0	31	1.00	$\times 10^{0}$
C	Relation to the course/ lectures	0	25	1	30	1.00	$\times 10^{0}$
P	Aspects of the procedure	8	17	17	14	1.10	$\times 10^{-1}$
M	Manual or answer book	1	24	0	31	4.46	$\times 10^{-1}$
I	level of interest	0	25	0	31	1.00	$\times 10^{0}$
R	results obtained	10	15	9	22	4.11	$\times 10^{-1}$
L	new learning achieved	0	25	0	31	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	7	18	10	21	7.77	$\times 10^{-1}$
F	level of familiarity or relevance	0	25	1	30	1.00	$\times 10^{0}$
U	level of understanding	0	25	0	31	1.00	$\times 10^{0}$
O	others in the lab (students/ demonstrators)	0	25	2	29	4.97	$\times 10^{-1}$
S	level of simplicity	0	25	0	31	1.00	$\times 10^{0}$
X	uncategorised	5	20	2	29	2.23	$\times 10^{-1}$

7.2.2.3 Negative perceptions of "Biological Buffers"

Table S 12: Reasons cited for not enjoying the Biological Buffers experiment considered amongst all comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

Code		Data logger		Laptop		p
		Coded	Not coded	Coded	Not coded	
T	Time availability	1	50	0	54	4.86×10^{-1}
C	Relation to the course/ lectures	0	51	0	54	1.00×10^{0}
P	Aspects of the procedure	0	51	6	48	2.72×10^{-2}
M	Manual or answer book	2	49	7	47	1.62×10^{-1}
I	level of interest	2	49	0	54	2.34×10^{-1}
R	results obtained	1	50	2	52	1.00×10^{0}
L	new learning achieved	0	51	0	54	1.00×10^{0}
E	equipment, apparatus or technology	21	30	4	50	5.96×10^{-5}
F	level of familiarity or relevance	0	51	0	54	1.00×10^{0}
U	level of understanding	0	51	3	51	2.43×10^{-1}
O	others in the lab (students/ demonstrators)	0	51	0	54	1.00×10^{0}
S	level of simplicity	0	51	0	54	1.00×10^{0}
X	uncategorised	0	51	3	51	2.43×10^{-1}

Table S 13: Reasons cited for not enjoying the Biological Buffers experiment considered amongst only other negative comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

Code		Data logger		Laptop		p
		Coded	Not coded	Coded	Not coded	
T	Time availability	1	23	0	19	1.00×10^{0}
C	Relation to the course/ lectures	0	24	0	19	1.00×10^{0}
P	Aspects of the procedure	0	24	6	13	4.45×10^{-3}
M	Manual or answer book	2	22	7	12	3.04×10^{-2}
I	level of interest	2	22	0	19	4.95×10^{-1}
R	results obtained	1	23	2	17	5.75×10^{-1}
L	new learning achieved	0	24	0	19	1.00×10^{0}
E	equipment, apparatus or technology	21	3	4	15	2.44×10^{-5}
F	level of familiarity or relevance	0	24	0	19	1.00×10^{0}
U	level of understanding	0	24	3	16	7.85×10^{-2}
O	others in the lab (students/ demonstrators)	0	24	0	19	1.00×10^{0}
S	level of simplicity	0	24	0	19	1.00×10^{0}
X	uncategorised	0	24	3	16	7.85×10^{-2}

Table S 14: Areas of potential improvement cited for the Biological Buffers experiment considered amongst all comments received in response to item 18: "What aspects of the experiment need improvement and what changes would you suggest?"

Code		Data logger		Laptop		p
		Coded	Not coded	Coded	Not coded	
T	Time availability	0	35	2	30	2.39×10^{-1}
C	Relation to the course/ lectures	0	35	0	32	1.00×10^{0}
P	Aspects of the procedure	1	34	2	30	6.03×10^{-1}
M	Manual or answer book	3	32	17	15	1.10×10^{-4}
I	level of interest	1	34	0	32	1.00×10^{0}
R	results obtained	0	35	0	32	1.00×10^{0}
L	new learning achieved	1	34	0	32	1.00×10^{0}
E	equipment, apparatus or technology	25	10	6	26	2.16×10^{-5}
F	level of familiarity or relevance	0	35	0	32	1.00×10^{0}
U	level of understanding	0	35	0	32	1.00×10^{0}
O	others in the lab (students/ demonstrators)	1	34	1	31	1.00×10^{0}
S	level of simplicity	0	35	0	32	1.00×10^{0}
X	uncategorised	2	33	0	32	4.93×10^{-1}

Table S 15: Areas of potential improvement cited for the Biological Buffers experiment considered amongst only other negative comments received in response to item 18: "What aspects of the experiment need improvement and what changes would you suggest?"

Code		Data logger		Laptop		p
		Coded	Not coded	Coded	Not coded	
T	Time availability	0	31	2	23	1.95×10^{-1}
C	Relation to the course/ lectures	0	31	0	25	1.00×10^{0}
P	Aspects of the procedure	1	30	2	23	5.81×10^{-1}
M	Manual or answer book	3	28	17	8	1.11×10^{-5}
I	level of interest	1	30	0	25	1.00×10^{0}
R	results obtained	0	31	0	25	1.00×10^{0}
L	new learning achieved	1	30	0	25	1.00×10^{0}
E	equipment, apparatus or technology	25	6	6	19	3.38×10^{-5}
F	level of familiarity or relevance	0	31	0	25	1.00×10^{0}
U	level of understanding	0	31	0	25	1.00×10^{0}
O	others in the lab (students/ demonstrators)	1	30	1	24	1.00×10^{0}
S	level of simplicity	0	31	0	25	1.00×10^{0}
X	uncategorised	2	29	0	25	4.97×10^{-1}

7.2.3 Comparative tests for the Vapour Pressure experiment data

The total number of surveys collected for the Vapour Pressure experiment in 2011 and 2012 were 84 and 103 surveys respectively. In this case, the significance level (α) is 0.05 and the number of statistical tests conducted (n) is 140.Therefore, p values below $\alpha/n = 3.57\times10^{-4}$ are shaded to indicate refutation of the relevant null hypothesis, controlling for family-wise error.

Quantitative comparisons in this section compare mean scored Likert-type item responses, using the t-test for unequal variances. $p < \alpha / n$ refutes the null hypothesis that mean scores are equal for the two data sets. Qualitative comparisons test the significance of the association between the data set sampled, and content and/or nature of the comments received using Fisher's exact test. $p < \alpha / n$ refutes the null hypothesis that response content is independent of the student data set sampled (data logger or laptop).

Table S 16: Quantitative comparisons for the Vapour Pressure experiment

item		$m_{data\ logger}$	m_{laptop}	df	t	p
1	data interpretation skills	0.36	1.06	131.7	-4.90	2.76×10^{-6}
2	laboratory skills	0.45	1.00	137.9	-3.64	3.88×10^{-4}
3	interest	-0.15	0.46	161.5	-3.53	5.33×10^{-4}
4	clear assessment	0.26	0.66	169.5	-2.62	9.50×10^{-3}
5	clear expected learning	0.28	0.91	138.4	-4.15	5.76×10^{-5}
6	increased understanding	0.33	0.98	142.9	-4.21	4.58×10^{-5}
7	background information	0.61	0.89	168.0	-1.94	5.41×10^{-2}
8	demonstrators	1.19	1.33	169.0	-1.07	2.85×10^{-1}
9	clear procedure	0.04	0.43	179.1	-2.34	2.02×10^{-2}
10	relevance to discipline	0.63	0.77	162.8	-0.95	3.41×10^{-1}
11	benefit of teamwork	1.36	1.39	172.7	-0.28	7.84×10^{-1}
12	responsibility for own learing	0.56	0.99	147.6	-3.27	1.35×10^{-3}
13	time availability	-0.54	-0.05	138.2	-4.89	2.74×10^{-6}
14	overall learning experience	-0.02	0.69	149.0	-5.06	1.22×10^{-6}

7.2.3.1 General perceptions "Vapour Pressure"

Table S 17: General nature of responses to item 15: "Did you enjoy doing the experiment? Why or why not?" for the Vapour Pressure experiment

Are responses more positive?			Are responses more negative?		
	Data logger	Laptop		Data logger	Laptop
Positive	18	54	Negative	39	31
Not positive	37	25	Not negative	16	48
	p =	5.34×10^{-5}		p =	4.09×10^{-4}

Table S 18: Topic referenced in response to item 15: "Did you enjoy doing the experiment? Why or why not?" for the Vapour Pressure experiment

		Data logger		Laptop		p
Code		Positive	Negative	Positive	Negative	
T	Time availability	1	1	1	1	1.00×10^{0}
C	Relation to the course/ lectures	0	0	1	0	1.00×10^{0}
P	Aspects of the procedure	0	4	3	12	1.00×10^{0}
M	Manual or answer book	0	12	1	5	3.33×10^{-1}
I	level of interest	9	5	12	8	1.00×10^{0}
R	results obtained	0	0	1	0	1.00×10^{0}
L	new learning achieved	6	0	4	0	1.00×10^{0}
E	equipment, apparatus or technology	3	6	7	3	1.79×10^{-1}
F	level of familiarity or relevance	1	1	15	0	1.18×10^{-1}
U	level of understanding	1	11	6	2	4.44×10^{-3}
O	others in the lab (students/ demonstrators)	1	0	2	0	1.00×10^{0}
S	level of simplicity	0	7	12	0	1.99×10^{-5}
X	uncategorised	2	6	8	5	1.83×10^{-1}

Table S 19: Content referenced in response to item 16: "What did you think was the main lesson to be learned from the experiment" for the Vapour Pressure experiment

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
L	mention Laws - Raoult's law or Dalton's law	14	31	37	34	3.48×10^{-2}
I	mention ideal or non-ideal mixtures	2	43	3	68	1.00×10^{0}
F	mention intermolecular forces	9	36	15	56	1.00×10^{0}
P	mention vapour pressure	14	31	24	47	8.40×10^{-1}
A	mention use of aparatus or equipment	6	39	12	59	7.93×10^{-1}
X	uncategorised as any of the above	16	29	17	54	2.08×10^{-1}
Ap	mention (non)application of laws	3	42	12	59	1.57×11^{-1}
L/F/I	Contain comments coded L,F or I	20	25	47	24	3.33×12^{-2}

7.2.3.2 Positive perceptions of “Vapour Pressure”

Table S 20: Reasons cited for enjoying the Vapour Pressure experiment considered amongst all comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
T	Time availability	1	54	1	78	1.00×10^{0}
C	Relation to the course/ lectures	0	55	1	78	1.00×10^{0}
P	Aspects of the procedure	0	55	3	76	2.69×10^{-1}
M	Manual or answer book	0	55	1	78	1.00×10^{0}
I	level of interest	9	46	12	67	1.00×10^{0}
R	results obtained	0	55	1	78	1.00×10^{0}
L	new learning achieved	6	49	4	75	3.16×10^{-1}
E	equipment, apparatus or technology	3	52	7	72	5.25×10^{-1}
F	level of familiarity or relevance	1	54	15	64	2.20×10^{-3}
U	level of understanding	1	54	6	73	2.39×10^{-1}
O	others in the lab (students/ demonstrators)	1	54	2	77	1.00×10^{0}
S	level of simplicity	0	55	12	67	1.45×10^{-3}
X	uncategorised	2	53	8	71	1.97×10^{-1}

Table S 21: Reasons cited for enjoying the Vapour Pressure experiment considered only amongst other reasons cited for liking the experiment in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
T	Time availability	1	17	1	53	4.40×10^{-1}
C	Relation to the course/ lectures	0	18	1	53	1.00×10^{0}
P	Aspects of the procedure	0	18	3	51	5.68×10^{-1}
M	Manual or answer book	0	18	1	53	1.00×10^{0}
I	level of interest	9	9	12	42	3.63×10^{-2}
R	results obtained	0	18	1	53	1.00×10^{0}
L	new learning achieved	6	12	4	50	1.25×10^{-2}
E	equipment, apparatus or technology	3	15	7	47	7.03×10^{-1}
F	level of familiarity or relevance	1	17	15	39	5.64×10^{-2}
U	level of understanding	1	17	6	48	6.72×10^{-1}
O	others in the lab (students/ demonstrators)	1	17	2	52	1.00×10^{0}
S	level of simplicity	0	18	12	42	2.99×10^{-2}
X	uncategorised	2	16	8	46	1.00×10^{0}

Table S 22: Features cited as the most enjoyable and interesting aspects of the Vapour Pressure experiment considered amongst all responses to item 17: "What aspects of the experiment did you find most enjoyable and interesting?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	0	43	0	69	1.00	$\times 10^{0}$
C	Relation to the course/ lectures	0	43	0	69	1.00	$\times 10^{0}$
P	Aspects of the procedure	7	36	13	56	8.04	$\times 10^{-1}$
M	Manual or answer book	2	41	2	67	6.37	$\times 10^{-1}$
I	level of interest	0	43	1	68	1.00	$\times 10^{0}$
R	results obtained	4	39	15	54	1.21	$\times 10^{-1}$
L	new learning achieved	4	39	2	67	2.01	$\times 10^{-1}$
E	equipment, apparatus or technology	14	29	35	34	7.82	$\times 10^{-2}$
F	level of familiarity or relevance	5	38	3	66	2.56	$\times 10^{-1}$
U	level of understanding	2	41	0	69	1.45	$\times 10^{-1}$
O	others in the lab (students/ demonstrators)	4	39	1	68	7.06	$\times 10^{-2}$
S	level of simplicity	0	43	1	68	1.00	$\times 10^{0}$
X	uncategorised	3	40	4	65	1.00	$\times 10^{0}$

Table S 23: Features cited as the most enjoyable and interesting aspects of the Vapour Pressure experiment considered amongst only other positive responses to item 17: "What aspects of the experiment did you find most enjoyable and interesting?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	0	32	0	63	1.00	$\times 10^{0}$
C	Relation to the course/ lectures	0	32	0	63	1.00	$\times 10^{0}$
P	Aspects of the procedure	7	25	13	50	1.00	$\times 10^{0}$
M	Manual or answer book	2	30	2	61	6.01	$\times 10^{-1}$
I	level of interest	0	32	1	62	1.00	$\times 10^{0}$
R	results obtained	4	28	15	48	2.79	$\times 10^{-1}$
L	new learning achieved	4	28	2	61	1.75	$\times 10^{-1}$
E	equipment, apparatus or technology	14	18	35	28	2.88	$\times 10^{-1}$
F	level of familiarity or relevance	5	27	3	60	1.14	$\times 10^{-1}$
U	level of understanding	2	30	0	63	1.11	$\times 10^{-1}$
O	others in the lab (students/ demonstrators)	4	28	1	62	4.26	$\times 10^{-2}$
S	level of simplicity	0	32	1	62	1.00	$\times 10^{0}$
X	uncategorised	3	29	4	59	6.84	$\times 10^{-1}$

7.2.3.3 Negative perceptions of "Vapour Pressure"

Table S 24: Reasons cited for not enjoying the Vapour Pressure experiment considered amongst all comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
T	Time availability	1	54	1	78	1.00×10^{0}
C	Relation to the course/ lectures	0	55	0	79	1.00×10^{0}
P	Aspects of the procedure	4	51	12	67	1.88×10^{-1}
M	Manual or answer book	12	43	5	74	1.54×10^{-2}
I	level of interest	5	50	8	71	1.00×10^{0}
R	results obtained	0	55	0	79	1.00×10^{0}
L	new learning achieved	0	55	0	79	1.00×10^{0}
E	equipment, apparatus or technology	6	49	3	76	1.60×10^{-1}
F	level of familiarity or relevance	1	54	0	79	4.10×10^{-1}
U	level of understanding	11	44	2	77	1.69×10^{-3}
O	others in the lab (students/ demonstrators)	0	55	0	79	1.00×10^{0}
S	level of simplicity	7	48	0	79	1.55×10^{-3}
X	uncategorised	6	49	5	74	3.58×10^{-1}

Table S 25: Reasons cited for not enjoying the Vapour Pressure experiment considered amongst only other negative comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
T	Time availability	1	38	1	30	1.00×10^{0}
C	Relation to the course/ lectures	0	39	0	31	1.00×10^{0}
P	Aspects of the procedure	4	35	12	19	8.68×10^{-3}
M	Manual or answer book	12	27	5	26	1.75×10^{-1}
I	level of interest	5	34	8	23	2.20×10^{-1}
R	results obtained	0	39	0	31	1.00×10^{0}
L	new learning achieved	0	39	0	31	1.00×10^{0}
E	equipment, apparatus or technology	6	33	3	28	7.21×10^{-1}
F	level of familiarity or relevance	1	38	0	31	1.00×10^{0}
U	level of understanding	11	28	2	29	2.91×10^{-2}
O	others in the lab (students/ demonstrators)	0	39	0	31	1.00×10^{0}
S	level of simplicity	7	32	0	31	1.50×10^{-2}
X	uncategorised	6	33	5	26	1.00×10^{0}

Table S 26: Areas of potential improvement cited for the Vapour Pressure experiment considered amongst all comments received in response to item 18: "What aspects of the experiment need improvement and what changes would you suggest?"

		Data logger		Laptop		p	
Code		Coded	Not coded	Coded	Not coded		
T	Time availability	6	34	3	54	1.55	$\times 10^{-1}$
C	Relation to the course/ lectures	0	40	1	56	1.00	$\times 10^{0}$
P	Aspects of the procedure	7	33	8	49	7.77	$\times 10^{-1}$
M	Manual or answer book	18	22	15	42	8.11	$\times 10^{-2}$
I	level of interest	2	38	1	56	5.67	$\times 10^{-1}$
R	results obtained	0	40	1	56	1.00	$\times 10^{0}$
L	new learning achieved	0	40	0	57	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	4	36	8	49	7.56	$\times 10^{-1}$
F	level of familiarity or relevance	0	40	0	57	1.00	$\times 10^{0}$
U	level of understanding	1	39	0	57	4.12	$\times 10^{-1}$
O	others in the lab (students/ demonstrators)	2	38	0	57	1.68	$\times 10^{-1}$
S	level of simplicity	2	38	0	57	1.68	$\times 10^{-1}$
X	uncategorised	3	37	2	55	4.01	$\times 10^{-1}$

Table S 27: Areas of potential improvement cited for the Vapour Pressure experiment considered amongst only other negative comments received in response to item 18: "What aspects of the experiment need improvement and what changes would you suggest?"

		Data logger		Laptop		p	
Code		Coded	Not coded	Coded	Not coded		
T	Time availability	6	29	3	31	4.77	$\times 10^{-1}$
C	Relation to the course/ lectures	0	35	1	33	4.93	$\times 10^{-1}$
P	Aspects of the procedure	7	28	8	26	7.77	$\times 10^{-1}$
M	Manual or answer book	18	17	15	19	6.32	$\times 10^{-1}$
I	level of interest	2	33	1	33	1.00	$\times 10^{0}$
R	results obtained	0	35	1	33	4.93	$\times 10^{-1}$
L	new learning achieved	0	35	0	34	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	4	31	8	26	2.18	$\times 10^{-1}$
F	level of familiarity or relevance	0	35	0	34	1.00	$\times 10^{0}$
U	level of understanding	1	34	0	34	1.00	$\times 10^{0}$
O	others in the lab (students/ demonstrators)	2	33	0	34	4.93	$\times 10^{-1}$
S	level of simplicity	2	33	0	34	4.93	$\times 10^{-1}$
X	uncategorised	3	32	2	32	1.00	$\times 10^{0}$

7.2.4 Comparative tests for the Copper (II) Ion Concentration experiment data

The total number of surveys collected for the Copper (II) Ion Concentration experiment in 2011 and 2012 were 126 and 120 surveys respectively. In this case, the significance level (α) is 0.05 and the number of statistical tests conducted (n) is 137.Therefore, p values below $\alpha/n = 3.65\times10^{-4}$ are shaded to indicate refutation of the relevant null hypothesis, controlling for family-wise error.

Quantitative comparisons in this section compare mean scored Likert-type item responses, using the t-test for unequal variances. $p < \alpha / n$ refutes the null hypothesis that mean scores are equal for the two data sets. Qualitative comparisons test the significance of the association between the data set sampled, and content and/or nature of the comments received using Fisher's exact test. $p < \alpha / n$ refutes the null hypothesis that response content is independent of the student data set sampled (data logger or laptop).

Table S 28: Quantitative comparisons for the Copper (II) Ion Concentration experiment

item		$m_{data\ logger}$	m_{laptop}	df	t	p
1	data interpretation skills	1.01	0.97	240.1	0.51	6.13×10^{-1}
2	laboratory skills	1.10	1.04	242.9	0.62	5.38×10^{-1}
3	interest	0.80	0.95	244.0	-1.37	1.73×10^{-1}
4	clear assessment	1.09	1.13	243.5	-0.47	6.39×10^{-1}
5	clear expected learning	1.14	1.14	242.4	0.01	9.90×10^{-1}
6	increased understanding	0.94	0.89	242.0	0.51	6.09×10^{-1}
7	background information	1.12	1.28	243.7	-1.80	7.37×10^{-2}
8	demonstrators	1.58	1.59	237.9	-0.09	9.26×10^{-1}
9	clear procedure	1.28	1.27	242.4	0.15	8.84×10^{-1}
10	relevance to discipline	0.97	1.03	243.8	-0.58	5.63×10^{-1}
11	benefit of teamwork	1.53	1.37	216.6	1.70	9.14×10^{-2}
12	responsibility for own learing	1.05	0.97	242.9	0.78	4.37×10^{-1}
13	time availability	0.22	0.22	238.3	0.08	9.34×10^{-1}
14	overall learning experience	0.93	0.96	241.3	-0.42	6.75×10^{-1}

7.2.4.1 General perceptions of "Determination of Copper (II) Ion Concentration"

Table S 29: General nature of responses to item 15: "Did you enjoy doing the experiment? Why or why not?" for the Copper (II) Ion Concentration experiment

Are responses more positive?			Are responses more negative?		
	Data logger	Laptop		Data logger	Laptop
Positive	66	86	Negative	15	9
Not positive	15	9	Not negative	66	86
	p =	1.22×10^{-1}		p =	1.22×10^{-1}

Table S 30: Topic referenced in response to item 15: "Did you enjoy doing the experiment? Why or why not?" for the Copper (II) Ion Concentration experiment

		Data logger		Laptop		p
Code		Positive	Negative	Positive	Negative	
T	Time availability	10	0	6	0	1.00×10^{0}
C	Relation to the course/ lectures	2	0	1	0	1.00×10^{0}
P	Aspects of the procedure	3	2	5	1	5.45×10^{-1}
M	Manual or answer book	3	2	4	2	1.00×10^{0}
I	level of interest	11	5	11	3	6.89×10^{-1}
R	results obtained	5	0	6	1	1.00×10^{0}
L	new learning achieved	8	0	11	0	1.00×10^{0}
E	equipment, apparatus or technology	6	4	10	2	3.48×10^{-1}
F	level of familiarity or relevance	7	0	6	1	1.00×10^{0}
U	level of understanding	8	1	21	0	3.00×10^{-1}
O	others in the lab (students/ demonstrators)	3	1	1	0	1.00×10^{0}
S	level of simplicity	12	1	32	1	4.90×10^{-1}
X	uncategorised	14	2	16	1	6.01×10^{-1}

Table S 31: Content referenced in response to item 16: "What did you think was the main lesson to be learned from the experiment" for the Copper (II) Ion Concentration experiment

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
L	Light's wavelength & colour relationship	4	65	1	82	1.77×10^{-1}
B	Beer's law/ conc. & absorbance relationship	37	32	31	52	5.06×10^{-2}
C	Determination of unknown concentration	3	66	6	77	5.12×10^{-1}
E	Use of equipment	9	60	15	68	5.04×10^{-1}
X	none of the above	22	47	36	47	1.80×10^{-1}

7.2.4.2 Positive perceptions of “Determination of Copper (II) Ion Concentration”

Table S 32: Reasons cited for enjoying the Copper (II) Ion Concentration experiment considered amongst all comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p	
Code		Coded	Not coded	Coded	Not coded		
T	Time availability	10	71	6	89	1.95	$\times 10^{-1}$
C	Relation to the course/ lectures	2	79	1	94	5.95	$\times 10^{-1}$
P	Aspects of the procedure	3	78	5	90	7.27	$\times 10^{-1}$
M	Manual or answer book	3	78	4	91	1.00	$\times 10^{0}$
I	level of interest	11	70	11	84	8.20	$\times 10^{-1}$
R	results obtained	5	76	6	89	1.00	$\times 10^{0}$
L	new learning achieved	8	73	11	84	8.10	$\times 10^{-1}$
E	equipment, apparatus or technology	6	75	10	85	6.01	$\times 10^{-1}$
F	level of familiarity or relevance	7	74	6	89	5.77	$\times 10^{-1}$
U	level of understanding	8	73	21	74	4.05	$\times 10^{-2}$
O	others in the lab (students/ demonstrators)	3	78	1	94	3.35	$\times 10^{-1}$
S	level of simplicity	12	69	32	63	4.98	$\times 10^{-3}$
X	uncategorised	14	67	16	79	1.00	$\times 10^{0}$

Table S 33: Reasons cited for enjoying the Copper (II) Ion Concentration experiment considered only amongst other reasons cited for liking the experiment in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p	
Code		Coded	Not coded	Coded	Not coded		
T	Time availability	10	56	6	80	1.17	$\times 10^{-1}$
C	Relation to the course/ lectures	2	64	1	85	5.80	$\times 10^{-1}$
P	Aspects of the procedure	3	63	5	81	1.00	$\times 10^{0}$
M	Manual or answer book	3	63	4	82	1.00	$\times 10^{0}$
I	level of interest	11	55	11	75	6.43	$\times 10^{-1}$
R	results obtained	5	61	6	80	1.00	$\times 10^{0}$
L	new learning achieved	8	58	11	75	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	6	60	10	76	7.91	$\times 10^{-1}$
F	level of familiarity or relevance	7	59	6	80	5.61	$\times 10^{-1}$
U	level of understanding	8	58	21	65	6.33	$\times 10^{-2}$
O	others in the lab (students/ demonstrators)	3	63	1	85	3.17	$\times 10^{-1}$
S	level of simplicity	12	54	32	54	1.18	$\times 10^{-2}$
X	uncategorised	14	52	16	70	8.37	$\times 10^{-1}$

Table S 34: Features cited as the most enjoyable and interesting aspects of the Copper (II) Ion Concentration experiment considered amongst all responses to item 17: "What aspects of the experiment did you find most enjoyable and interesting?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	1	64	0	83	4.39	$\times 10^{-1}$
C	Relation to the course/ lectures	0	65	0	83	1.00	$\times 10^{0}$
P	Aspects of the procedure	21	44	26	57	1.00	$\times 10^{0}$
M	Manual or answer book	0	65	0	83	1.00	$\times 10^{0}$
I	level of interest	1	64	4	79	3.85	$\times 10^{-1}$
R	results obtained	21	44	30	53	7.28	$\times 10^{-1}$
L	new learning achieved	2	63	0	83	1.91	$\times 10^{-1}$
E	equipment, apparatus or technology	21	44	35	48	2.36	$\times 10^{-1}$
F	level of familiarity or relevance	2	63	0	83	1.91	$\times 10^{-1}$
U	level of understanding	0	65	0	83	1.00	$\times 10^{0}$
O	others in the lab (students/ demonstrators)	1	64	1	82	1.00	$\times 10^{0}$
S	level of simplicity	0	65	0	83	1.00	$\times 10^{0}$
X	uncategorised	5	60	1	82	8.70	$\times 10^{-2}$

Table S 35: Features cited as the most enjoyable and interesting aspects of the Copper (II) Ion Concentration experiment considered amongst only other positive responses to item 17: "What aspects of the experiment did you find most enjoyable and interesting?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	1	61	0	79	4.40	$\times 10^{-1}$
C	Relation to the course/ lectures	0	62	0	79	1.00	$\times 10^{0}$
P	Aspects of the procedure	21	41	26	53	5.94	$\times 10^{-1}$
M	Manual or answer book	0	62	0	79	1.00	$\times 10^{0}$
I	level of interest	1	61	4	75	3.85	$\times 10^{-1}$
R	results obtained	21	41	30	49	7.24	$\times 10^{-1}$
L	new learning achieved	2	60	0	79	1.92	$\times 10^{-1}$
E	equipment, apparatus or technology	21	41	35	44	2.29	$\times 10^{-1}$
F	level of familiarity or relevance	2	60	0	79	1.92	$\times 10^{-1}$
U	level of understanding	0	62	0	79	1.00	$\times 10^{0}$
O	others in the lab (students/ demonstrators)	1	61	1	78	1.00	$\times 10^{0}$
S	level of simplicity	0	62	0	79	1.00	$\times 10^{0}$
X	uncategorised	5	57	1	78	8.68	$\times 10^{-2}$

7.2.4.3 Negative perceptions of "Determination of Copper (II) Ion Concentration"

Table S 36: Reasons cited for not enjoying the Copper (II) Ion Concentration experiment considered amongst all comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
T	Time availability	0	81	0	95	1.00×10^{0}
C	Relation to the course/ lectures	0	81	0	95	1.00×10^{0}
P	Aspects of the procedure	2	79	1	94	5.95×10^{-1}
M	Manual or answer book	2	79	2	93	1.00×10^{0}
I	level of interest	5	76	3	92	4.73×10^{-1}
R	results obtained	0	81	1	94	1.00×10^{0}
L	new learning achieved	0	81	0	95	1.00×10^{0}
E	equipment, apparatus or technology	4	77	2	93	4.16×10^{-1}
F	level of familiarity or relevance	0	81	1	94	1.00×10^{0}
U	level of understanding	1	80	0	95	4.60×10^{-1}
O	others in the lab (students/ demonstrators)	1	80	0	95	4.60×10^{-1}
S	level of simplicity	1	80	1	94	1.00×10^{0}
X	uncategorised	2	79	1	94	5.95×10^{-1}

Table S 37: Reasons cited for not enjoying the Copper (II) Ion Concentration experiment considered amongst only other negative comments received in response to item 15: "Did you enjoy doing the experiment? Why or why not?"

		Data logger		Laptop		p
Code		Coded	Not coded	Coded	Not coded	
T	Time availability	0	15	0	9	1.00×10^{0}
C	Relation to the course/ lectures	0	15	0	9	1.00×10^{0}
P	Aspects of the procedure	2	13	1	8	1.00×10^{0}
M	Manual or answer book	2	13	2	7	1.00×10^{0}
I	level of interest	5	10	3	6	1.00×10^{0}
R	results obtained	0	15	1	8	3.75×10^{-1}
L	new learning achieved	0	15	0	9	1.00×10^{0}
E	equipment, apparatus or technology	4	11	2	7	1.00×10^{0}
F	level of familiarity or relevance	0	15	1	8	3.75×10^{-1}
U	level of understanding	1	14	0	9	1.00×10^{0}
O	others in the lab (students/ demonstrators)	1	14	0	9	1.00×10^{0}
S	level of simplicity	1	14	1	8	1.00×10^{0}
X	uncategorised	2	13	1	8	1.00×10^{0}

Table S 38: Areas of potential improvement cited for the Copper (II) Ion Concentration experiment considered amongst all comments received in response to item 18: "What aspects of the experiment need improvement and what changes would you suggest?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	0	43	1	55	1.00	$\times 10^{0}$
C	Relation to the course/ lectures	1	42	0	56	4.34	$\times 10^{-1}$
P	Aspects of the procedure	3	40	6	50	7.28	$\times 10^{-1}$
M	Manual or answer book	9	34	7	49	2.83	$\times 10^{-1}$
I	level of interest	1	42	0	56	4.34	$\times 10^{-1}$
R	results obtained	0	43	0	56	1.00	$\times 10^{0}$
L	new learning achieved	0	43	0	56	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	18	25	12	44	4.61	$\times 10^{-2}$
F	level of familiarity or relevance	1	42	0	56	4.34	$\times 10^{-1}$
U	level of understanding	1	42	0	56	4.34	$\times 10^{-1}$
O	others in the lab (students/ demonstrators)	0	43	1	55	1.00	$\times 10^{0}$
S	level of simplicity	0	43	2	54	5.04	$\times 10^{-1}$
X	uncategorised	1	42	1	55	1.00	$\times 10^{0}$

Table S 39: Areas of potential improvement cited for the Copper (II) Ion Concentration experiment considered amongst only other negative comments received in response to item 18: "What aspects of the experiment need improvement and what changes would you suggest?"

Code		Data logger		Laptop		p	
		Coded	Not coded	Coded	Not coded		
T	Time availability	0	32	1	25	4.48	$\times 10^{-1}$
C	Relation to the course/ lectures	1	31	0	26	1.00	$\times 10^{0}$
P	Aspects of the procedure	3	29	6	20	2.74	$\times 10^{-1}$
M	Manual or answer book	9	23	7	19	1.00	$\times 10^{0}$
I	level of interest	1	31	0	26	1.00	$\times 10^{0}$
R	results obtained	0	32	0	26	1.00	$\times 10^{0}$
L	new learning achieved	0	32	0	26	1.00	$\times 10^{0}$
E	equipment, apparatus or technology	18	14	12	14	5.98	$\times 10^{-1}$
F	level of familiarity or relevance	1	31	0	26	1.00	$\times 10^{0}$
U	level of understanding	1	31	0	26	1.00	$\times 10^{0}$
O	others in the lab (students/ demonstrators)	0	32	1	25	4.48	$\times 10^{-1}$
S	level of simplicity	0	32	2	24	1.97	$\times 10^{-1}$
X	uncategorised	1	31	1	25	1.00	$\times 10^{0}$

7.3 Supporting information for section 3.3

7.3.1 Sample sizes

Table S 40: Survey responses available prior to data cleaning processes

	Year	Semester	Experiment	ID surveys	Total surveys collected	% surveys with ID provided
Foundations of Chemistry IA/B experiments	2012	2	Aromas	22	103	21%
	2012	2	Analysis of Spinach Extracts	39	107	36%
	2012	2	Activity Series	34	81	42%
	2012	2	Thermochemistry	24	77	31%
	2012	2	Reaction Kinetics	17	74	23%
	2013	1	Introductory Experiment	104	126	83%
	2013	1	Quantitative Techniques	54	61	89%
	2013	1	Determination of Vitamin C Content in Apple juice	44	57	77%
	2013	1	Equilibrium & Le Chatelier	119	137	87%
	2013	1	Absorption Spectrophotometry	71	83	86%
	2013	2	Aromas	229	248	92%
	2013	2	Analysis of Spinach Extracts	182	206	88%
	2013	2	Thermochemistry	185	204	91%
	2013	2	Activity Series	146	161	91%
	2013	2	Reaction Kinetics	72	82	88%
Chemistry IA/B Experiments	2012	2	Biological Buffers	30	80	38%
	2012	2	Melting Points and Recrystallisation	25	70	36%
	2012	2	Reaction Kinetics	54	84	64%
	2012	2	Liquid-Liquid Extraction and TLC	22	72	31%
	2012	2	Synthesis of Aspirin	14	36	39%
	2012	2	Analysis of Spinach Extracts	21	77	27%
	2013	1	Thermochemistry	220	227	97%
	2013	1	Vapour Pressure	140	148	95%
	2013	1	Quantitative Techniques	195	203	96%
	2013	1	Equilibrium & Le Chatelier	167	174	96%
	2013	1	Ion Exchange Chromatography	244	252	97%
	2013	1	Absorption Spectrophotometry	224	232	97%
	2013	2	Analysis of Spinach Extracts	210	218	96%
	2013	2	Synthesis of Aspirin	125	129	97%
	2013	2	Reaction Kinetics	201	205	98%
	2013	2	Melting Points and Recrystallisation	181	182	99%
	2013	2	Biological Buffers	162	170	95%
	2013	2	Liquid-Liquid Extraction and TLC	126	128	98%

7.3.2 Matlab codes for population level expected score distributions

The code presented below is one example of the technique used to derive population level response probability values, and other values used to generate figures showing the population level relationship between score and measure. The five response categories are referred to using A, B, C, D and E, from most to least positive respectively.

txy = is used to specify the Andrich threshold between categories X and Y
meanb = is used to specify the mean student measure
stdevb = is used to specify the standard deviation in student measures

vd = specifies the interval for which values of all functions are defined, and the space between data points defined.

PE, **PD**, **PC**, **PB**, and **PA** are the population level probabilities of observing responses in categories E,D,C,B and A respectively .They are computed using **quad(, ,)**, which specifies a finite range of values the integral term of Equation 33 is computed over (given integration from negative to positive infinity was not possible). These values may require adjustment for ideal calculation. Plots of PE, PD, PC, PB and PA against vd should yield smooth curves to indicate acceptable computation.

ASELLScore gives the population level mean expected score using the ASELL integer scoring system.
ASELLdevlessX and **ASELLdevmoreX** respectively give the lower and upper boundaries for the 95% confidence interval of the distribution of expected mean ASELL scores taken from samples of size X.

```
tab = 7.82;
tbc = 4.26;
tcd = -4.96;
tde = -7.12;
meanb = 0.03;
stdevb = 2.29;

cA = exp(-tab);
cB = exp(-tbc);
cC = exp(-tcd);
cD = exp(-tde);

vd = -15:0.1:15;

PE = zeros(size(vd));
for i = 1:length(vd)
d = vd(i);
funE = @(b) (1./(cA*cB*cC*cD*exp(4*b - 4*d) + cB*cC*cD*exp(3*b-3*d) + cC*cD*exp(2*b-2*d) + cD*exp(b-
d) + 1)).*normpdf(b,meanb,stdevb);
PE(i) = quad(funE,-20,20);
end

PD = zeros(size(vd));
for i = 1:length(vd)
d = vd(i);
funD = @(b) (1./(cA*cB*cC*exp(3*b - 3*d) + cB*cC*exp(2*b-2*d) + cC*exp(b-d) + 1 + (1/cD)*exp(d-
b))).*normpdf(b,meanb,stdevb);
PD(i) = quad(funD,-20,20);
end

PC = zeros(size(vd));
```

```
for i = 1:length(vd)
d = vd(i);
funC = @(b) (1./(cA*cB*exp(2*b - 2*d) + cB*exp(b-d) + 1 + (1/cC)*exp(d-b) + (1/(cC*cD))*exp(2*d-
2*b))).*normpdf(b,meanb,stdevb);
PC(i) = quad(funC,-20,20);
end

PB = zeros(size(vd));
for i = 1:length(vd)
d = vd(i);
funB = @(b) (1./(cA*exp(b - d) + 1 + (1/cB)*exp(d-b) + (1/(cB*cC))*exp(2*d-2*b) +
(1/(cB*cC*cD))*exp(3*d-3*b))).*normpdf(b,meanb,stdevb);
PB(i) = quad(funB,-18,18);
end

PA = zeros(size(vd));
for i = 1:length(vd)
d = vd(i);
funA = @(b) (1./(1 + (1/cA)*exp(d-b) + (1/(cA*cB))*exp(2*d-2*b) + (1/(cA*cB*cC))*exp(3*d-3*b) +
(1/(cA*cB*cC*cD))*exp(4*d-4*b))).*normpdf(b,meanb,stdevb);
PA(i) = quad(funA,-20,20);
end

for i = 1:length(vd)
ASELLScore(i)= 2*PA(i)+PB(i)-PD(i)-2*PE(i);
end

for i = 1:length(vd)
ASELLdev(i) = sqrt(PA(i)*((2-ASELLScore(i))^2)+PB(i)*((1-ASELLScore(i))^2)+PC(i)*((0-
ASELLScore(i))^2)+PD(i)*((-1-ASELLScore(i))^2)+PE(i)*((-2-ASELLScore(i))^2));
end

for i = 1:length(vd)
ASELLdevless10(i) = ASELLScore(i)-2*ASELLdev(i)/sqrt(9);
end

for i = 1:length(vd)
ASELLdevless20(i) = ASELLScore(i)-2*ASELLdev(i)/sqrt(19);
end

for i = 1:length(vd)
ASELLdevless30(i) = ASELLScore(i)-2*ASELLdev(i)/sqrt(29);
end

for i = 1:length(vd)
ASELLdevless50(i) = ASELLScore(i)-2*ASELLdev(i)/sqrt(49);
end

for i = 1:length(vd)
ASELLdevless100(i) = ASELLScore(i)-2*ASELLdev(i)/sqrt(99);
end

for i = 1:length(vd)
ASELLdevmore10(i) = ASELLScore(i)+2*ASELLdev(i)/sqrt(9);
end

for i = 1:length(vd)
ASELLdevmore20(i) = ASELLScore(i)+2*ASELLdev(i)/sqrt(19);
end

for i = 1:length(vd)
ASELLdevmore30(i) = ASELLScore(i)+2*ASELLdev(i)/sqrt(29);
end

for i = 1:length(vd)
ASELLdevmore50(i) = ASELLScore(i)+2*ASELLdev(i)/sqrt(49);
end

for i = 1:length(vd)
ASELLdevmore100(i) = ASELLScore(i)+2*ASELLdev(i)/sqrt(99);
end
```

7.3.3 Equality of response scales between different student cohorts

For each survey item, the equality of response scales between the two different student cohorts was tested. These tests were conducted using statistics gathered from the first Rasch models estimated after the initial removal of disconnected subsets of data. The Corrected Akaike Information Criterion was used to select the preferable explanation of the data, as shown in blue in Figure S 1 below. Results of a typical Likelihood ratio test are also shown (red), though this statistic does not account for parsimony of the explanation and was therefore not preferred as a criterion for model selection.

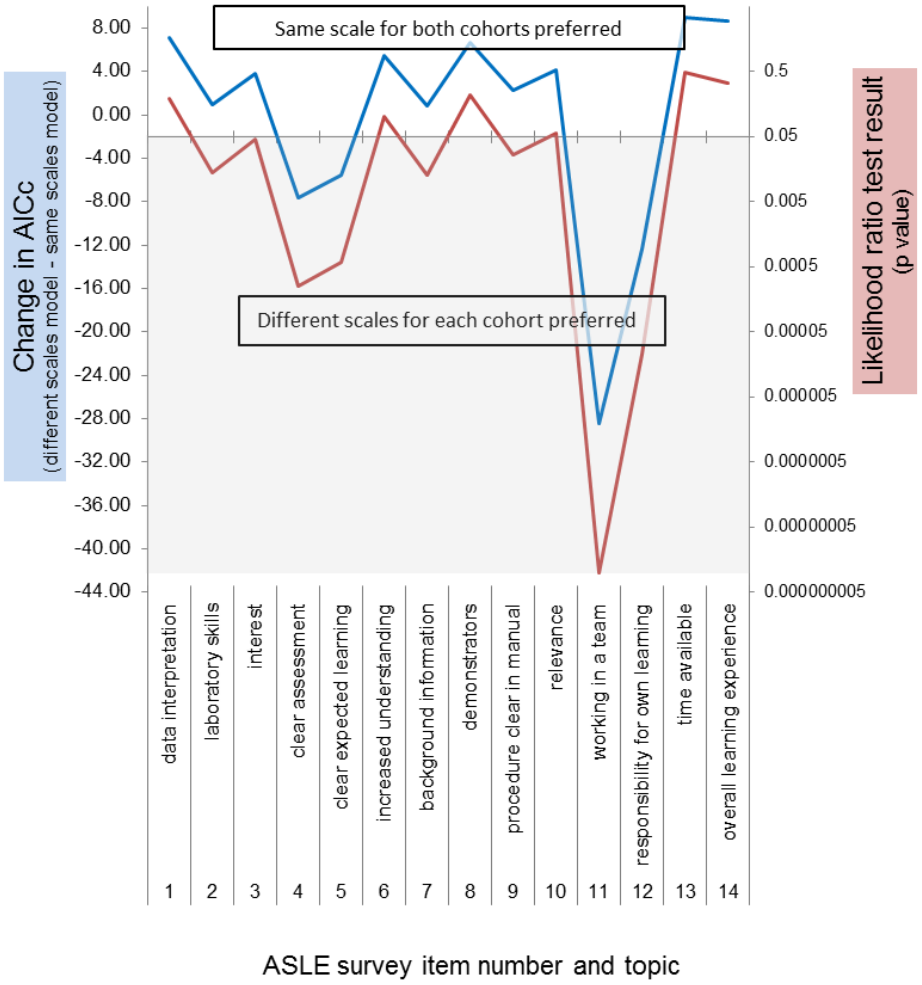


Figure S 1: Model selection for whether Foundations of Chemistry IA/B and Chemistry IA/B student cohorts were assigned different response scales

In the cases where the best explanation of the data is such that the two student cohorts treat the response scale differently (items 3, 4, 11 and 12 as judged above), rating scale associated statistics are reported for each cohort in the material following.

7.3.4 Item 1: "This experiment helped me to develop my data interpretation skills"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 164 persons removed
3. Further connectivity issues resolved: 13 persons removed
4. Misfit issues resolved: 81 persons with z-scores for infit or outfit |z|> 2 removed.
5. Further connectivity issues resolved. 33 persons and 6 items removed.
6. Extreme responses removed: 8 persons removed (final results reported)

Table S 41: Rasch model details for item 1

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label	/	Upper category label	measure	st. error		
		Agree	/	Strongly Agree	5.45	0.07	5.45	0.98
		Neutral	/	Agree	0.45	0.06	0.5	1.02
		Disagree	/	Neutral	-2.41	0.14	-2.22	1.01
		Strongly Disagree	/	Disagree	-3.49	0.4	-3.73	1.07

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C⇒M	Coherence M⇒C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	6.55	5.47	∞	40%	68%	1.03	1.01	0.6718	371
1	Agree	2.95	0.57	5.47	87%	76%	0.99	0.99	0.3223	1714
0	Neutral	-0.91	-2.07	0.57	54%	58%	0.99	0.94	0.5722	550
-1	Disagree	-2.98	-3.98	-2.07	6%	29%	1	1	1.0622	62
-2	Strongly Disagree	-4.78	−∞	-3.98	0%	0%	0.84	0.72	1.5321	7

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures empirical	Raw variance in observed data explained by measures modelled
Persons	828	2.50	1.89	0.40	0.56	0.14	0.24	37.7%	37.7%
Experiments	27	0.00	0.63	1.61	1.67	0.72	0.73	11.0%	11.0%

Data points: 2704 Log-likelihood chi square: 3561.54 df: 1847

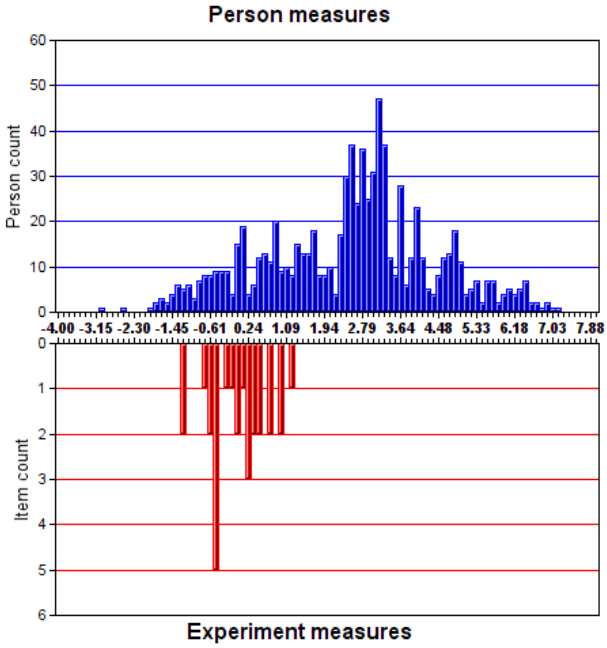


Figure S 2: Measure distributions for item 1

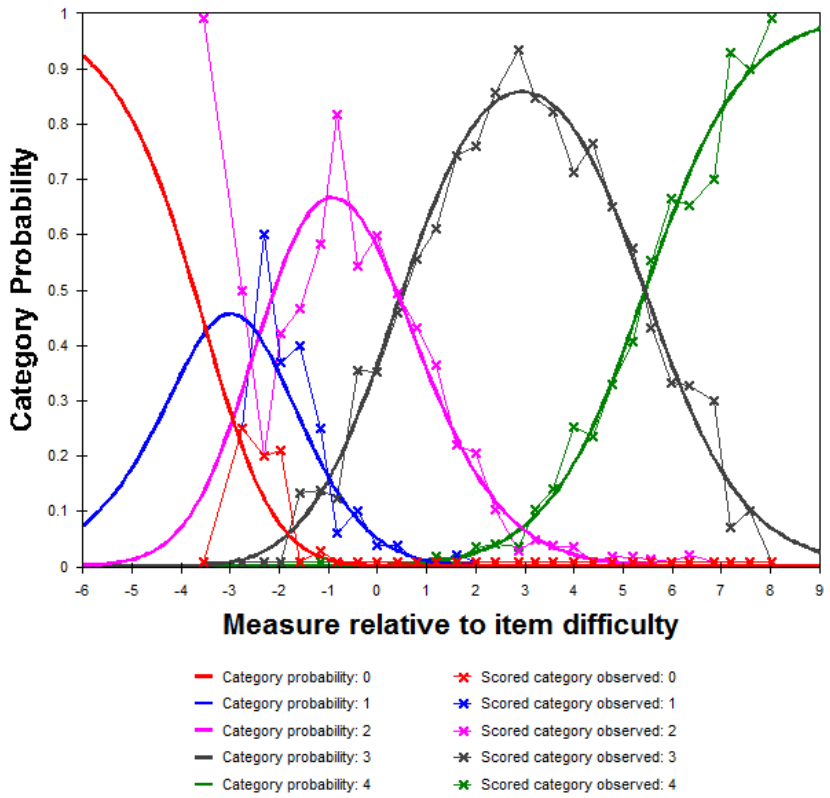


Figure S 3: Category structure for item 1

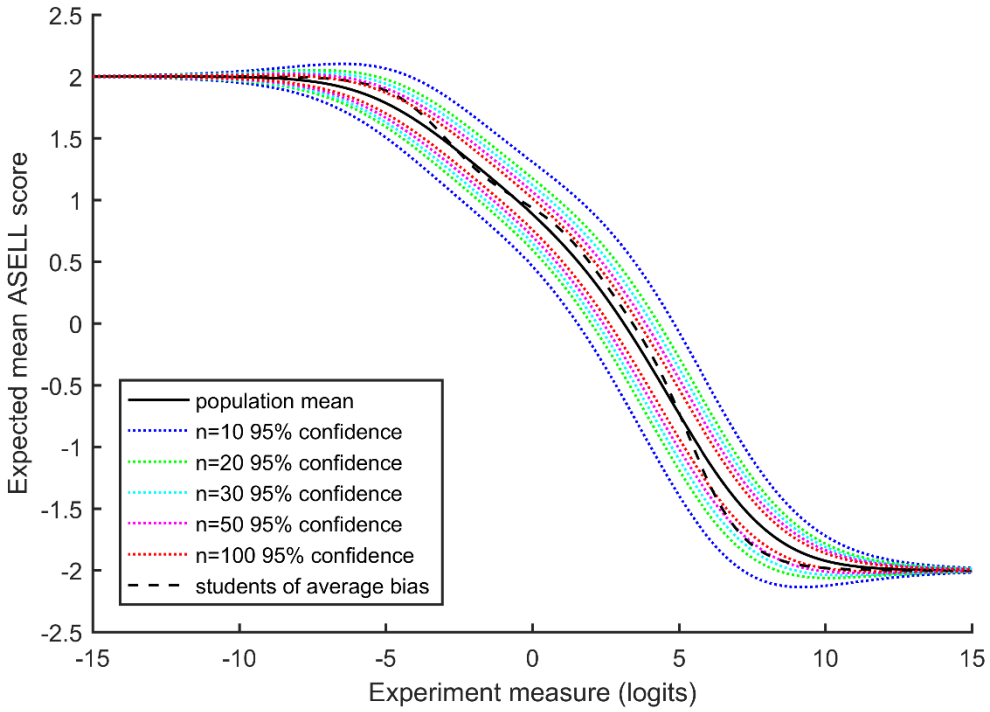


Figure S 4: Expected mean ASELL scores for item 1

7.3.5 Item 2: "This experiment helped me to develop my laboratory skills"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 221 persons and 1 item removed
3. Further connectivity issues resolved: 5 persons removed
4. Misfit issues resolved: 78 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
5. Further connectivity issues resolved. 12 persons removed (final results reported)

Table S 42: Rasch model details for item 2

RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label	/	Upper category label	measure	st. error		
	Agree	/	Strongly Agree	5.15	0.06	5.17	0.94
	Neutral	/	Agree	0.7	0.06	0.75	1.06
	Disagree	/	Neutral	-2.03	0.13	-1.96	0.99
	Strongly Disagree	/	Disagree	-3.82	0.35	-3.96	0.98

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C => M	Coherence M => C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	6.26	5.19	∞	52%	69%	1.06	1.04	0.6026	660
1	Agree	2.93	0.82	5.19	86%	73%	0.94	1.01	0.3356	1722
0	Neutral	-0.63	-1.88	0.82	51%	62%	0.94	0.94	0.6045	492
-1	Disagree	-2.95	-4.14	-1.88	27%	50%	1.04	1.04	0.9512	79
-2	Strongly Disagree	-5.03	−∞	-4.14	9%	50%	0.95	0.92	1.2395	11

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	811	2.94	1.77	0.61	0.76	0.27	0.37	31.4%	31.1%
Experiments	32	0	1.75	3.92	4.08	0.94	0.94	23.8%	23.5%

Data points: 2964 Log-likelihood chi square: 4080.63 df: 2119

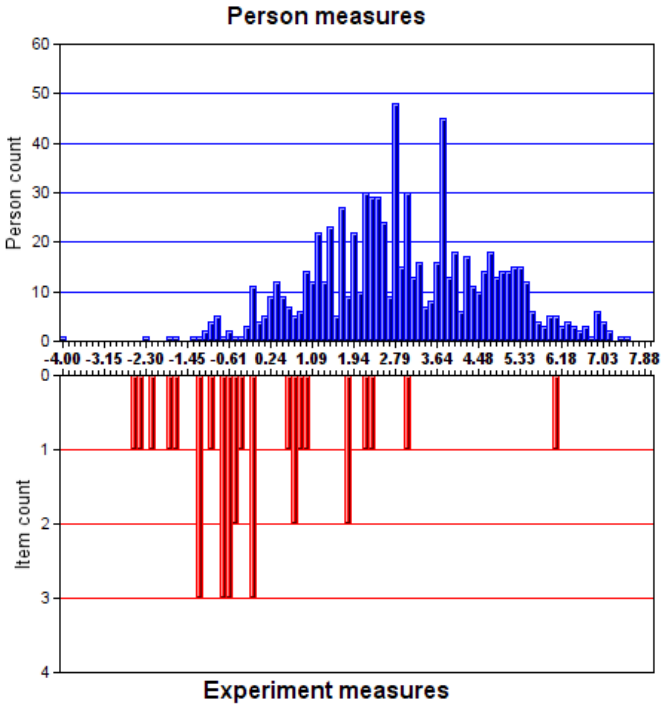


Figure S 5: Measure distributions for item 2

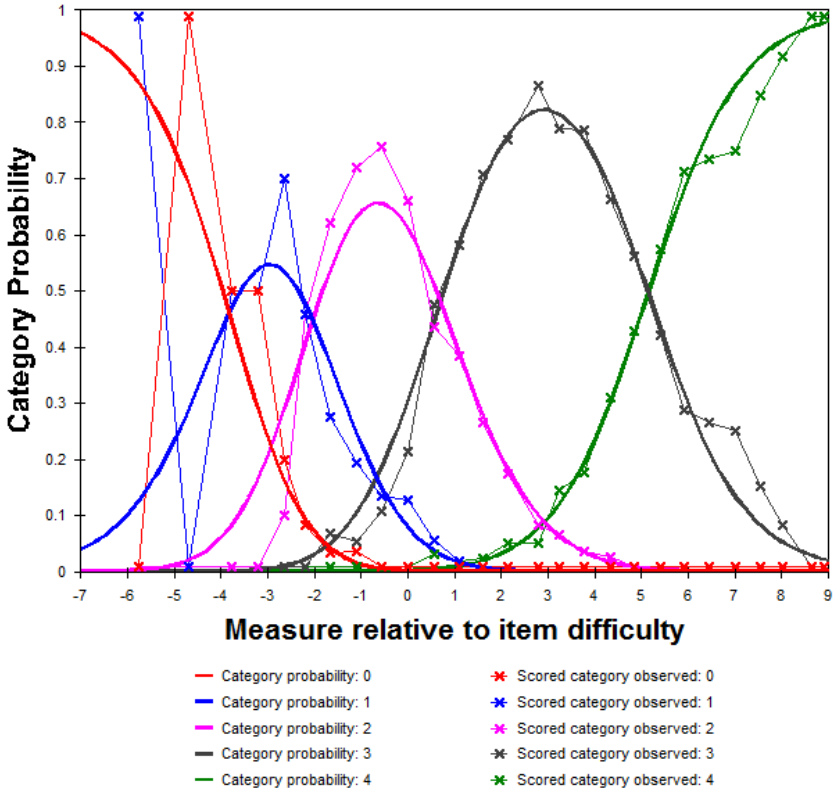


Figure S 6: Category structure for item 2

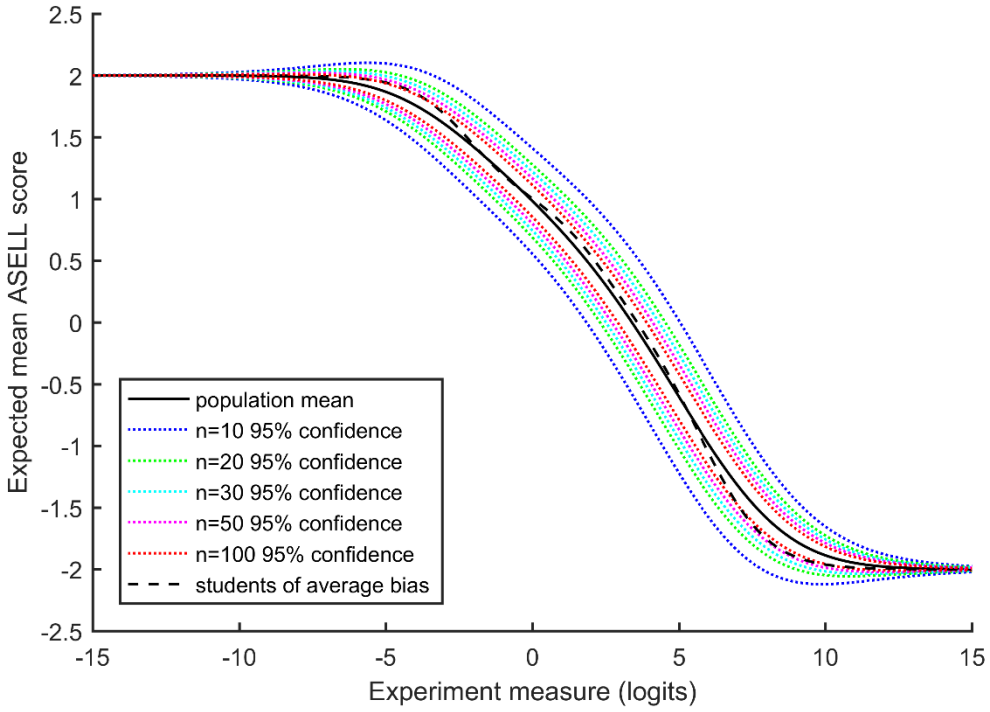


Figure S 7: Expected mean ASELL scores for item 2

7.3.6 Item 3: "I found this to be an interesting experiment"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 173 persons removed
3. Misfit issues resolved: 114 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Remaining data split into two subsets: one containing Foundations of Chemistry experiments, the other containing Chemistry IA experiments. To ensure connectivity, one previously removed misfitting person was added back into analysis. Resulting data had connectivity issues, but with the two cohorts still connected. Connectivity issues were resolved: 3 persons removed. (Final results reported)

Table S 43: Rasch model details for item 3

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label	/	Upper category label	measure	st. error		
		Agree	/	Strongly Agree	3.92	0.05	3.96	0.99
		Neutral	/	Agree	0.79	0.05	0.82	1.00
		Disagree	/	Neutral	-1.72	0.11	-1.59	1.01
		Strongly Disagree	/	Disagree	-2.99	0.27	-3.20	1.16

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C => M	Coherence M => C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	5.06	4.03	∞	46%	70%	1.02	1.01	0.6802	648
1	Agree	2.37	0.86	4.03	79%	64%	0.98	1.01	0.3882	1486
0	Neutral	-0.41	-1.48	0.86	49%	55%	1.01	1.01	0.6412	679
-1	Disagree	-2.40	-3.43	-1.48	14%	31%	1.01	0.99	1.0274	110
-2	Strongly Disagree	-4.26	−∞	-3.43	11%	67%	0.71	0.66	1.2025	18

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures empirical	Raw variance in observed data explained by measures modelled
Persons	838	2.11	1.60	0.74	0.86	0.35	0.43	40.2%	40.0%
Experiments	33	0.00	0.99	2.83	2.87	0.89	0.89	10.0%	10.0%

Data points: 2941 Log-likelihood chi square: 4905.13 df: 2068

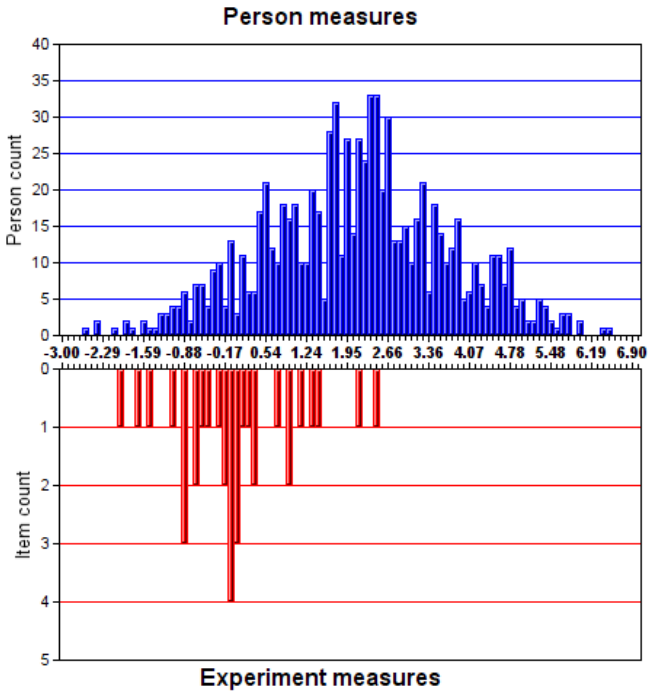


Figure S 8: Measure distributions for item 3

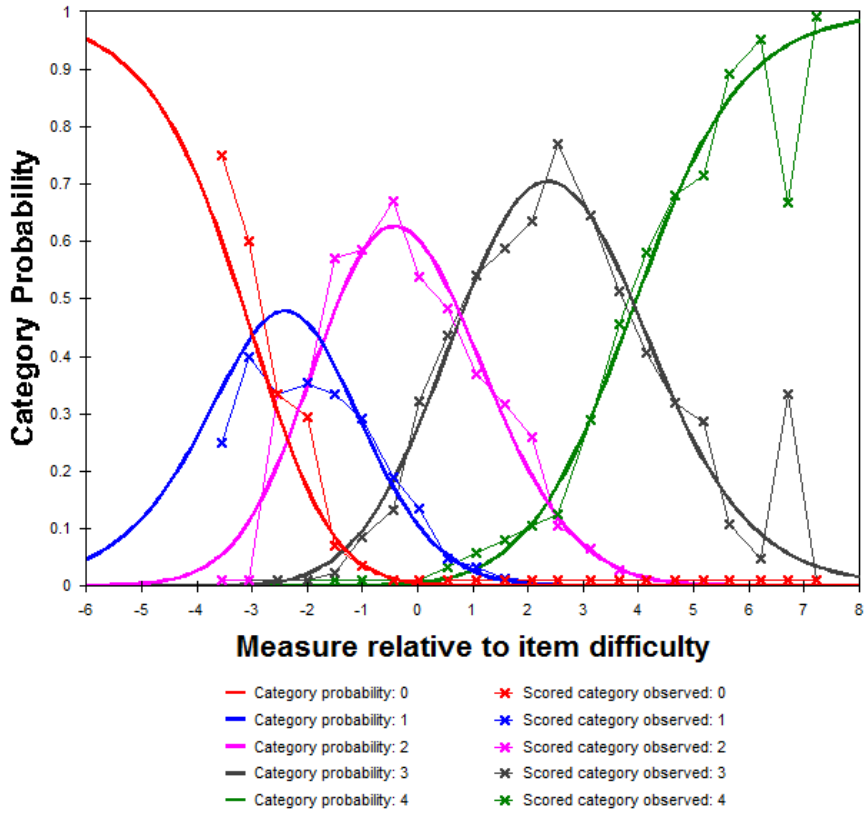


Figure S 9: Category structure for item 3

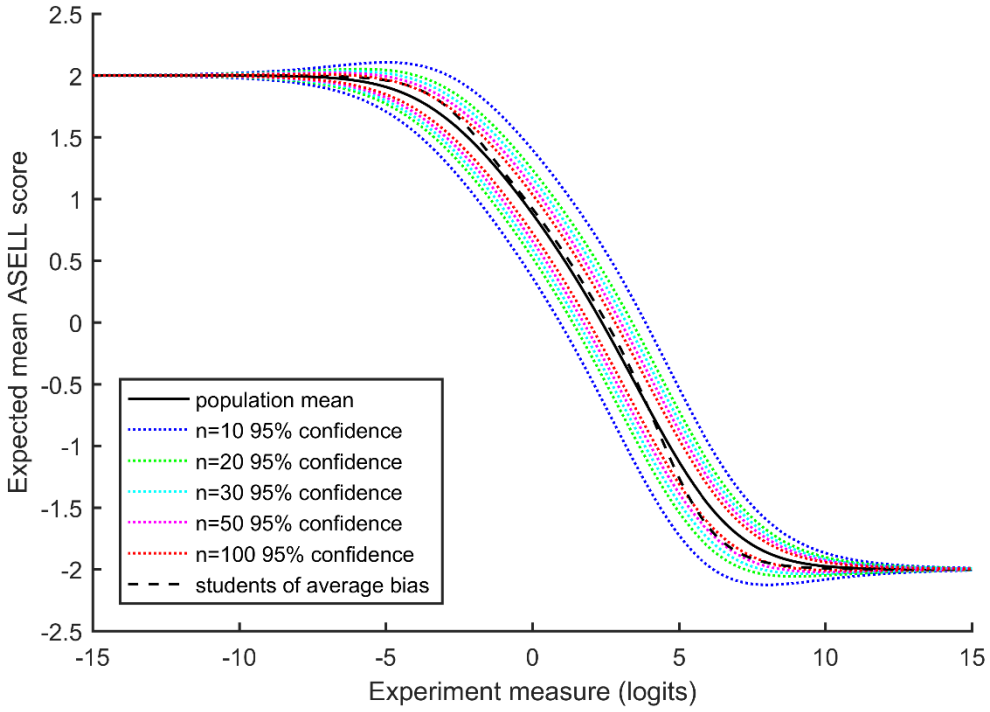


Figure S 10: Expected mean ASELL scores for item 3

7.3.7 Item 4: "It was clear to me how this laboratory exercise would be assessed"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items.
2. Connectivity issues, extreme persons and blank responses resolved: 241 persons, 1 item removed. Student cohorts were assigned separate rating scale structures.
3. Misfit issues resolved: 99 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 1 person removed (final results reported).

Table S 44: Rasch model details for item 4

CHEMISTRY IA/B	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
RESPONSE CATEGORY ASSOCIATED STATISTICS	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	3.82	0.07	3.85	0.98
	Neutral	/	Agree	0.44	0.07	0.52	1.02
	Disagree	/	Neutral	-1.57	0.15	-1.46	0.99
	Strongly Disagree	/	Disagree	-2.68	0.40	-2.92	1.12

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C => M	Coherence M => C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	4.94	3.91	∞	42%	68%	1.04	1.01	0.6741	418
1	Agree	2.15	0.63	3.91	83%	66%	0.97	1.00	0.3695	972
0	Neutral	-0.48	-1.38	0.63	45%	52%	0.98	0.98	0.6767	326
-1	Disagree	-2.20	-3.17	-1.38	8%	31%	1.05	1.05	1.1896	52
-2	Strongly Disagree	-3.98	-∞	-3.17	0%	0%	0.77	0.71	1.5054	7

FOUNDATIONS OF CHEMISTRY IA/B	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
RESPONSE CATEGORY ASSOCIATED STATISTICS	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	5.06	0.10	5.07	0.01
	Neutral	/	Agree	0.45	0.10	0.52	0.99
	Disagree	/	Neutral	-1.93	0.26	-1.86	1.01
	Strongly Disagree	/	Disagree	-3.58	1.02	-3.73	1.09

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C => M	Coherence M => C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	6.17	5.09	∞	54%	67%	0.99	0.95	0.5866	220
1	Agree	2.76	0.62	5.09	86%	74%	0.95	1.01	0.3228	649
0	Neutral	-0.69	-1.80	0.62	43%	59%	1.04	1.05	0.6696	158
-1	Disagree	-2.79	-3.93	-1.80	12%	67%	1.00	0.94	1.1922	17
-2	Strongly Disagree	-4.8	-∞	-3.93	0%	0%	0.73	0.63	1.5583	1

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	786	2.40	1.63	0.52	0.66	0.21	0.30	42.5%	42.2%
Experiments	32	0.00	0.60	1.12	1.20	0.56	0.59	2.9%	2.9%

Data points:	2820	Log-likelihood chi square:	4316.53	df:	1997

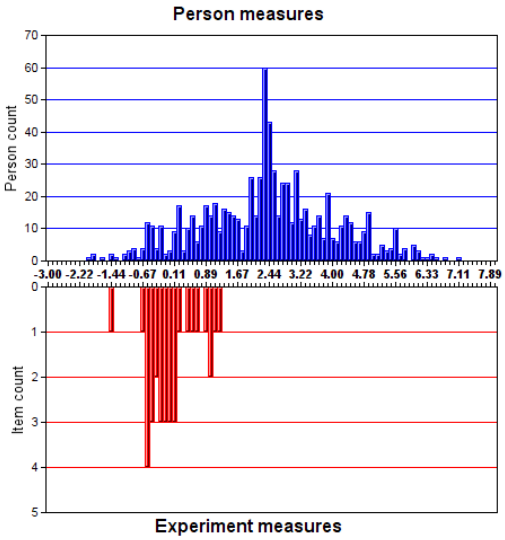


Figure S 11: Measure distributions for item 4

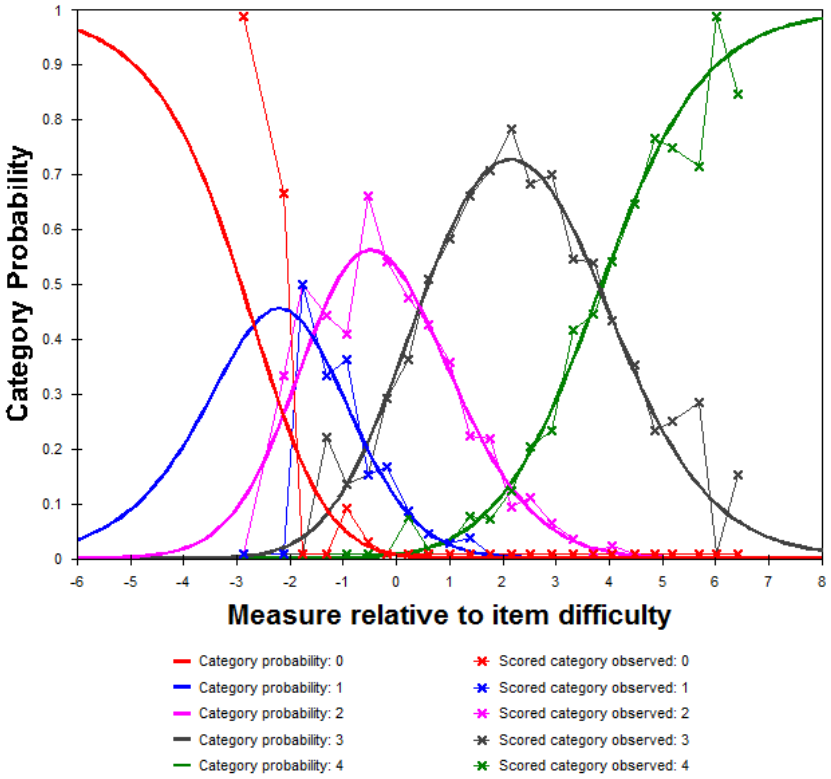


Figure S 12: Category structure for item 4 - Chemistry IA/B students

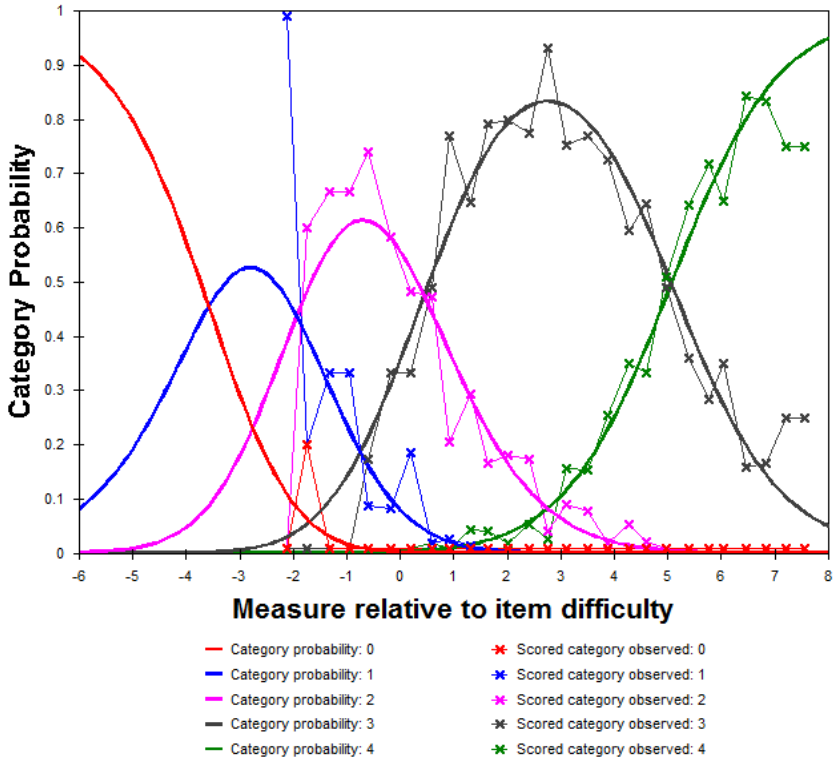


Figure S 13: Category structure for item 4 - Foundations of Chemistry IA/B students

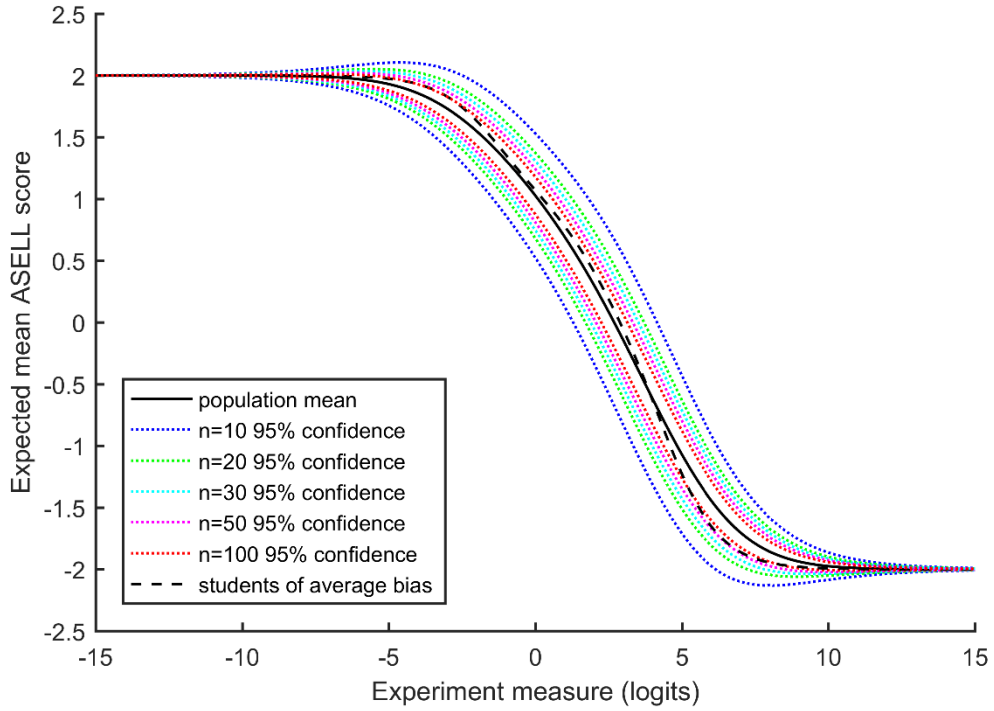


Figure S 14: Expected mean ASELL scores for item 4 – Chemistry IA/B students

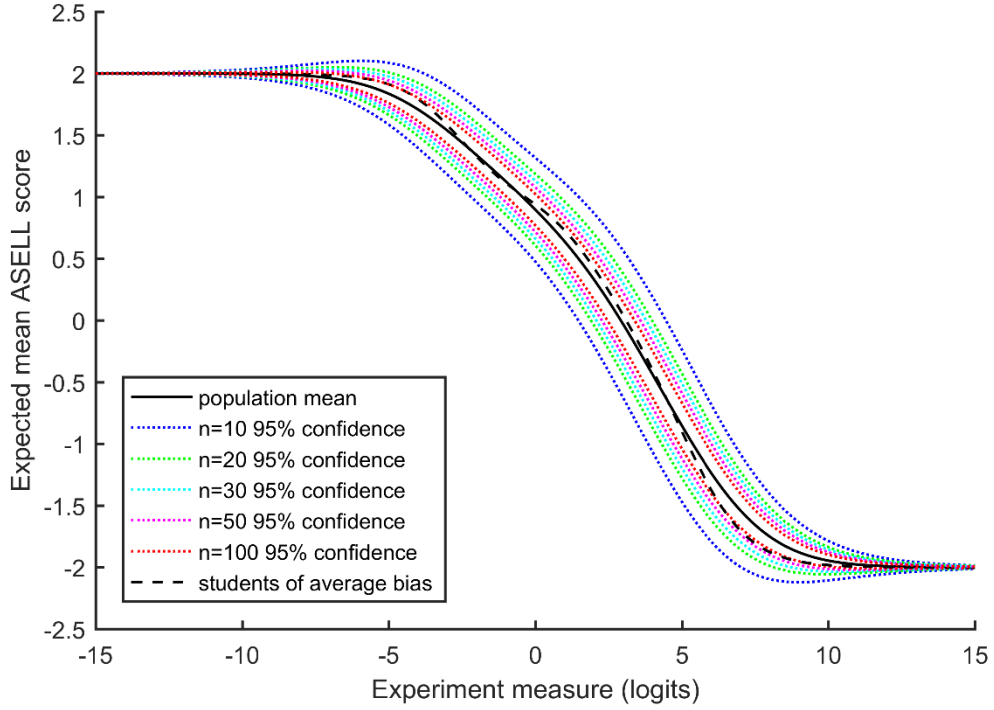


Figure S 15: Expected mean ASELL scores for item 4 - Foundations of Chemistry IA/B students

7.3.8 Item 5: "It was clear to me what I was expected to learn from completing this experiment"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items.
2. Connectivity issues, extreme persons and blank responses resolved: 205 persons removed. Student cohorts were assigned separate rating scale structures.
3. Misfit issues resolved: 94 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 13 persons removed (final results reported).

Table S 45: Rasch model details for item 5

CHEMISTRY IA/B RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	3.60	0.06	3.63	0.92
	Neutral	/	Agree	0.07	0.07	0.20	1.08
	Disagree	/	Neutral	-1.60	0.15	-1.40	0.96
	Strongly Disagree	/	Disagree	-2.07	0.35	-2.45	1.15

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C⇒M	Coherence M⇒C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	4.72	3.68	∞	44%	68%	1.09	1.04	0.6631	473
1	Agree	1.87	0.36	3.68	83%	67%	0.95	0.99	0.3681	1015
0	Neutral	-0.59	-1.30	0.36	45%	51%	0.90	0.88	0.6717	274
-1	Disagree	-1.95	-2.77	-1.30	0%	0%	1.12	1.10	1.3312	44
-2	Strongly Disagree	-3.49	−∞	-2.77	0%	0%	0.80	0.71	1.6939	9

FOUNDATIONS OF CHEMISTRY IA/B RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	4.79	0.10	4.80	1.02
	Neutral	/	Agree	0.00	0.11	0.10	1.10
	Disagree	/	Neutral	-2.14	0.28	-1.89	0.96
	Strongly Disagree	/	Disagree	-2.66	0.73	-3.02	1.02

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C⇒M	Coherence M⇒C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	5.90	4.82	∞	53%	67%	0.99	0.97	0.5999	210
1	Agree	2.41	0.24	4.82	86%	75%	0.99	0.98	0.3113	642
0	Neutral	-0.92	-1.76	0.24	33%	47%	1.06	1.04	0.7144	121
-1	Disagree	-2.48	-3.33	-1.76	8%	50%	0.78	0.74	1.0856	13
-2	Strongly Disagree	-4.06	−∞	-3.33	0%	0%	0.82	0.76	1.7300	2

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	815	2.23	1.60	0.36	0.52	0.11	0.21	39.9%	39.7%
Experiments	33	0.00	0.67	1.32	1.46	0.63	0.68	3.3%	3.3%

Data points:	2803	Log-likelihood chi square:	4214.61	df:	1950

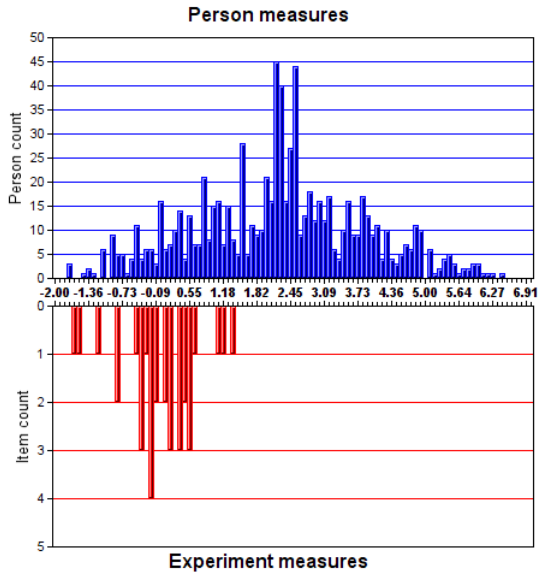


Figure S 16: Measure distributions for item 5

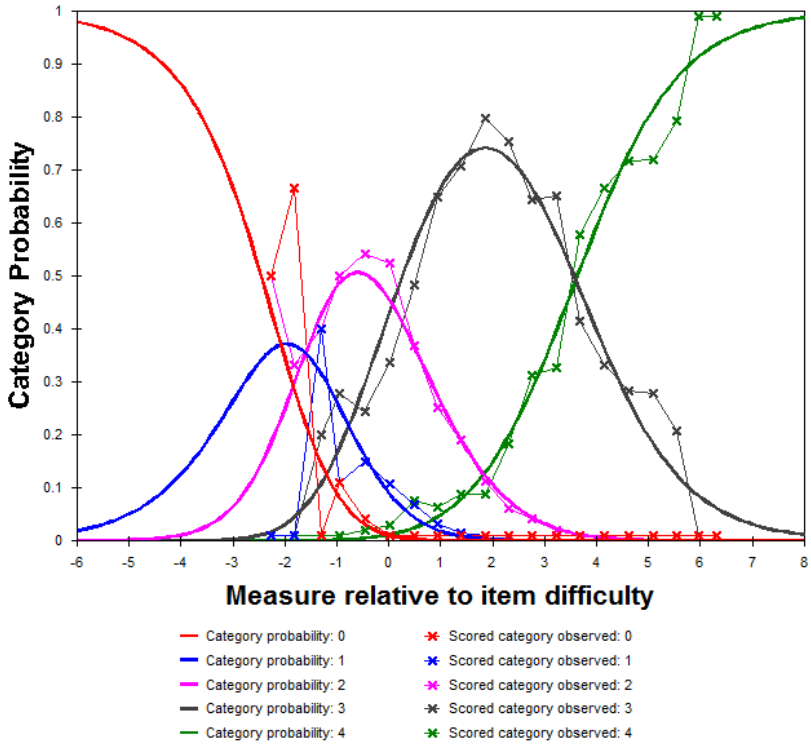


Figure S 17: Category structure for item 5 - Chemistry IA/B students

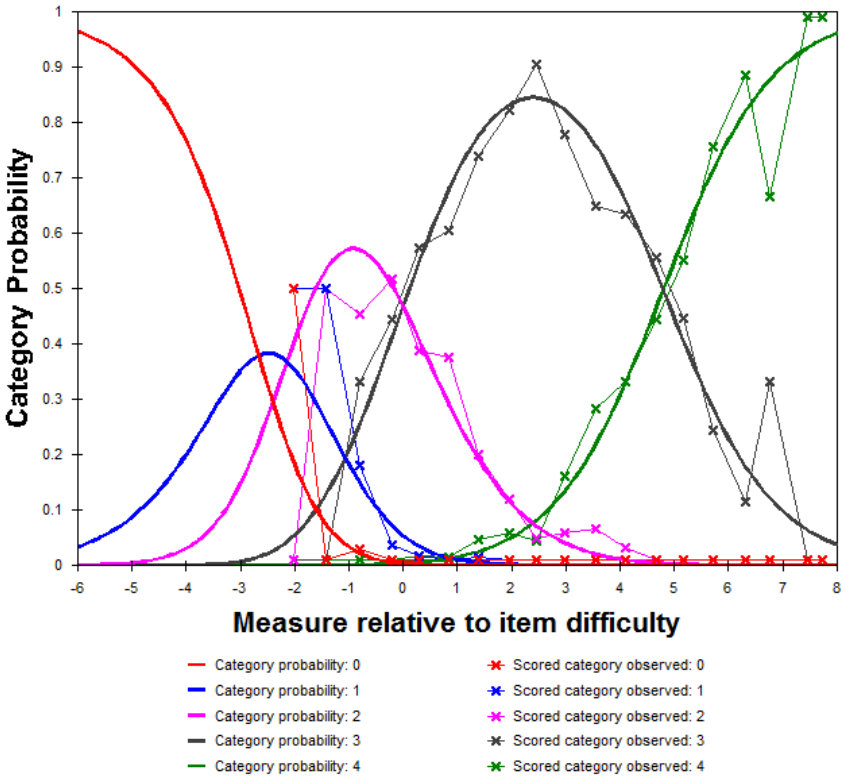


Figure S 18: Category structure for item 5 - Foundations of Chemistry IA/B students

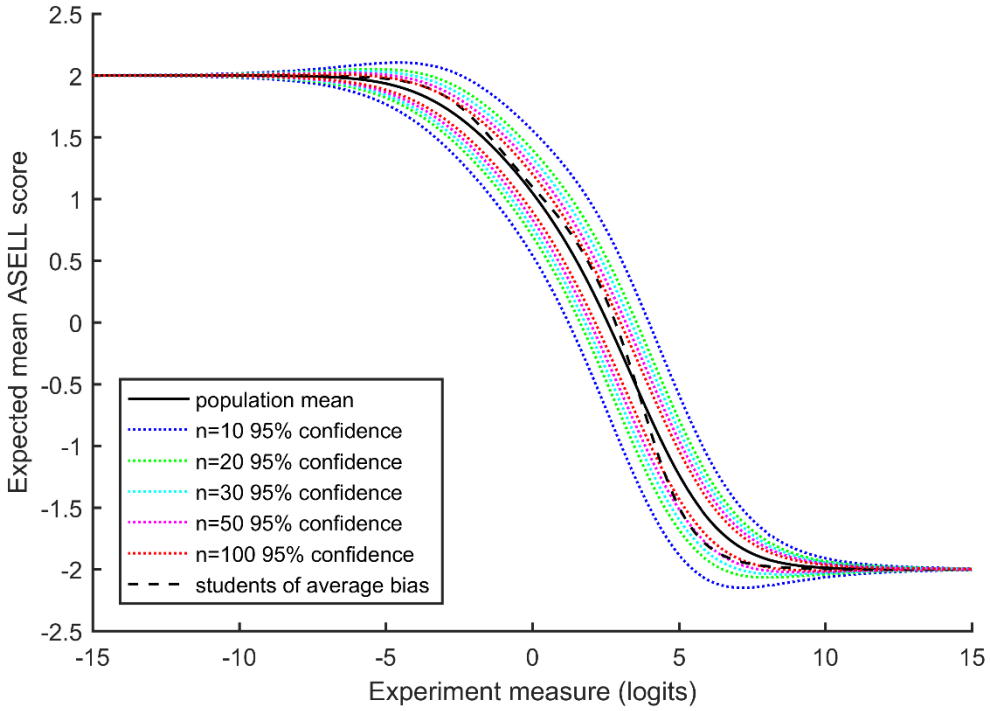


Figure S 19: Expected mean ASELL scores for item 5 - Chemistry IA/B students

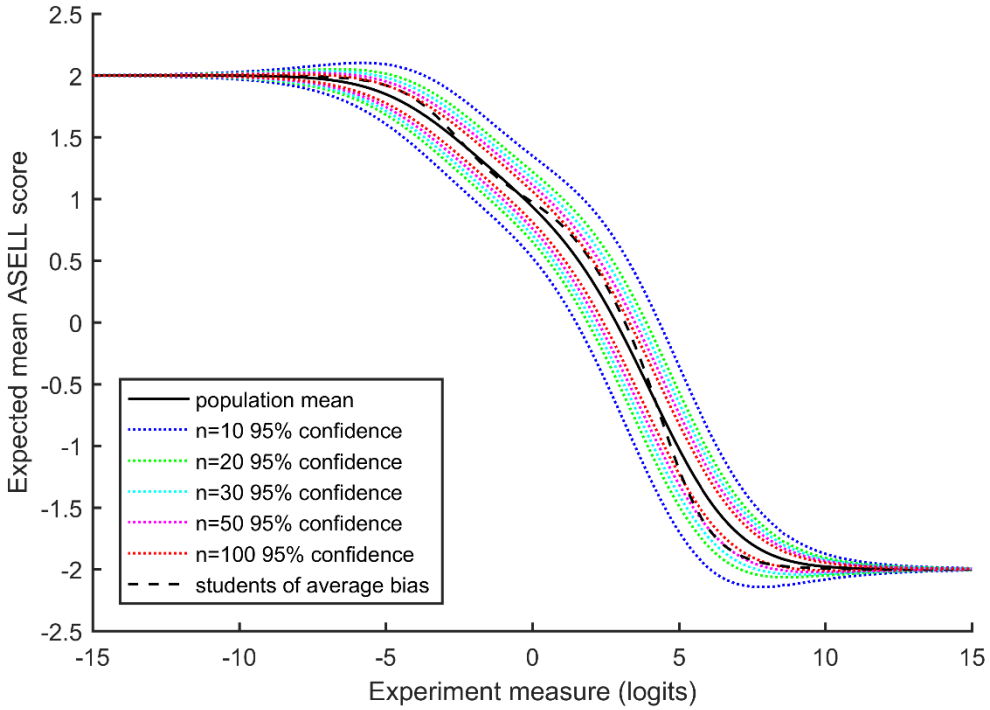


Figure S 20: Expected mean ASELL scores for item 5 - Foundations of Chemistry IA/B students

7.3.9 Item 6: "Completing this experiment has increased my understanding of chemistry"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 202 persons removed
3. Misfit issues resolved: 109 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 4 persons removed (final results reported).

Table S 46: Rasch model details for item 6

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label		Upper category label	measure	st. error		
		Agree	/	Strongly Agree	3.72	0.05	3.75	0.95
		Neutral	/	Agree	0.3	0.06	0.37	1.06
		Disagree	/	Neutral	-1.91	0.14	-1.59	0.98
		Strongly Disagree	/	Disagree	-2.11	0.31	-2.55	1.17

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C⇒M	Coherence M⇒C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	4.84	3.80	∞	49%	68%	1.08	1.05	0.6626	725
1	Agree	2.03	0.47	3.80	83%	66%	0.96	0.97	0.3559	1587
0	Neutral	-0.63	-1.42	0.47	39%	55%	0.94	0.92	0.6919	483
-1	Disagree	-2.09	-2.89	-1.42	7%	40%	1.08	1.09	1.2083	54
-2	Strongly Disagree	-3.58	−∞	-2.89	23%	100%	0.82	0.73	1.5798	13

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	812	2.16	1.49	0.54	0.65	0.22	0.30	40.3%	40.1%
Experiments	33	0.00	0.69	1.77	1.80	0.76	0.76	3.1%	3.1%

Data points: 2862 Log-likelihood chi square: 4573.53 df: 2015

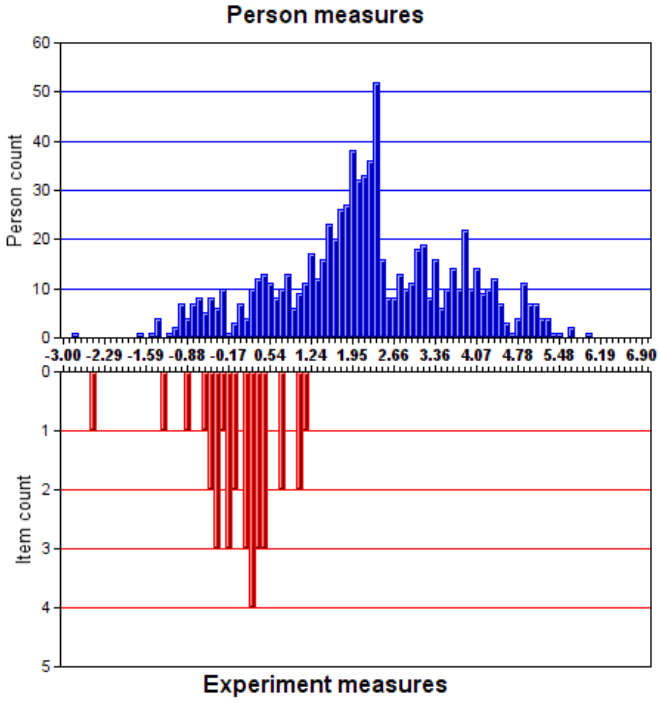


Figure S 21: Measure distributions for item 6

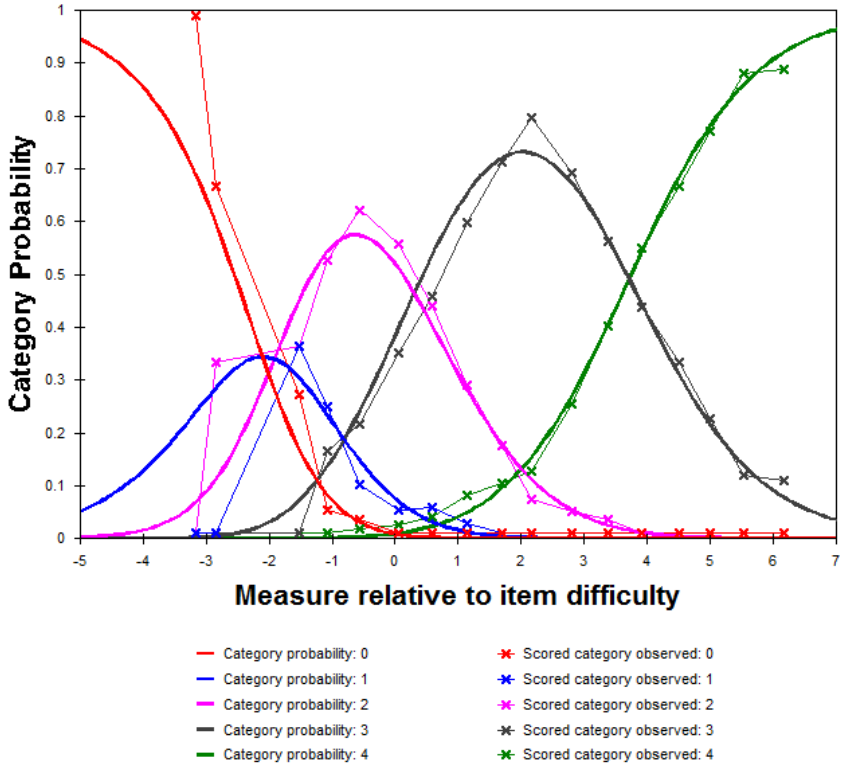


Figure S 22: Category structure for item 6

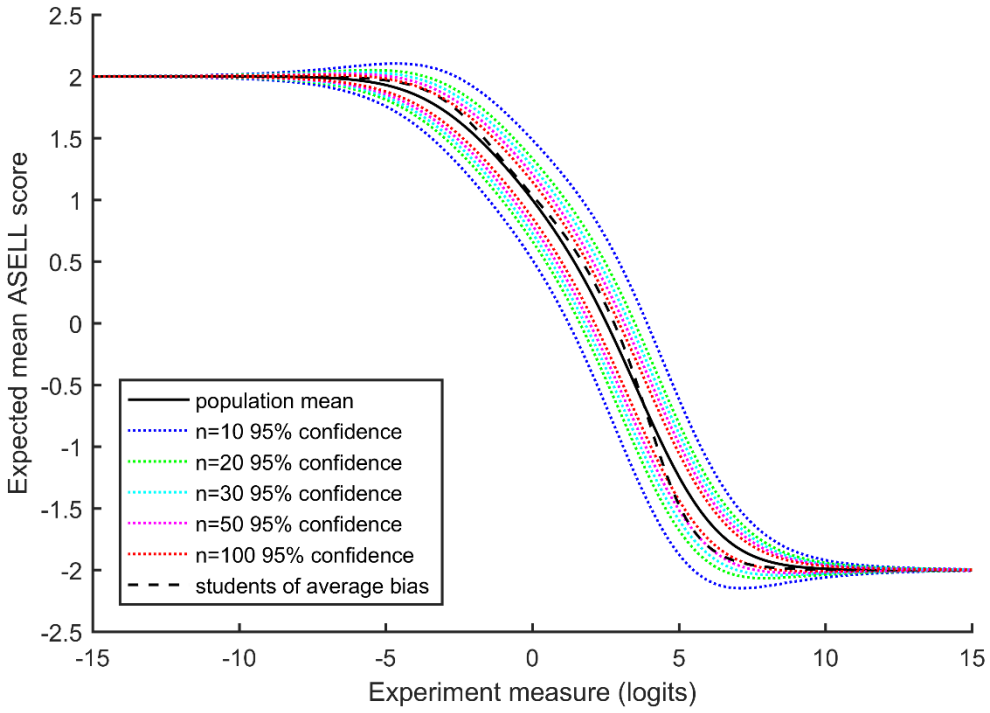


Figure S 23: Expected mean ASELL scores for item 6

7.3.10 Item 7: "Sufficient background information, of an appropriate standard, is provided in the introduction"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 214 persons removed
3. Misfit issues resolved: 112 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 16 persons removed (final results reported).

Table S 47: Rasch model details for item 7

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label		Upper category label	measure	st. error		
		Agree	/	Strongly Agree	3.79	0.05	3.82	0.96
		Neutral	/	Agree	0.52	0.06	0.65	1.04
		Disagree	/	Neutral	-1.05	0.12	-1.12	1.02
		Strongly Disagree	/	Disagree	-3.25	0.44	-3.35	0.99
ASELL score	Category Label	Category measure	Range from	Range to	Coherence C⇒M	Coherence M⇒C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	4.91	3.89	∞	53%	70%	1.04	1.03	0.6273	806
1	Agree	2.19	0.78	3.89	80%	65%	0.96	0.98	0.3783	1503
0	Neutral	-0.24	-1.18	0.78	41%	49%	0.98	0.97	0.7144	449
-1	Disagree	-2.21	-3.49	-1.18	8%	50%	0.96	0.91	1.1786	91
-2	Strongly Disagree	-4.43	-∞	-3.49	17%	50%	1.00	1.01	1.6363	6

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	785	2.50	1.46	0.59	0.73	0.26	0.34	38.3%	38.0%
Experiments	33	0.00	1.47	3.46	3.49	0.92	0.92	7.4%	7.4%

Data points: 2855 Log-likelihood chi square: 4646.49 df: 2035

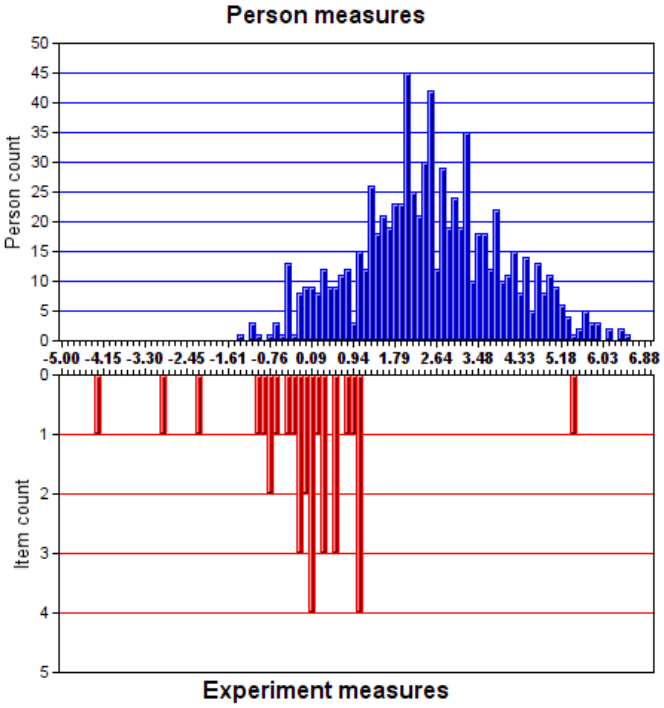


Figure S 24: Measure distributions for item 7

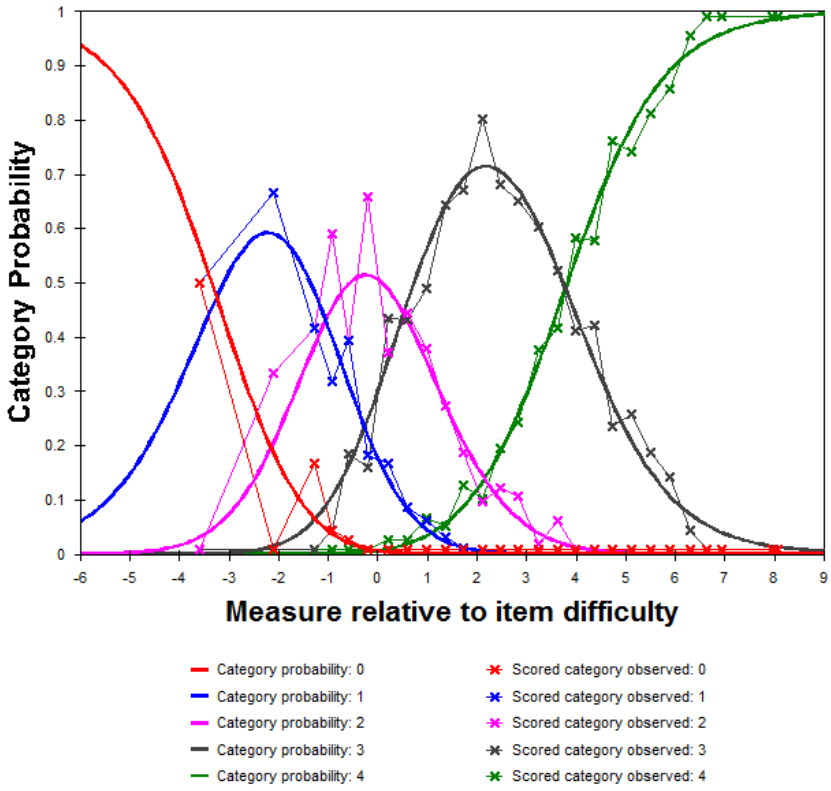


Figure S 25: Category structure for item 7

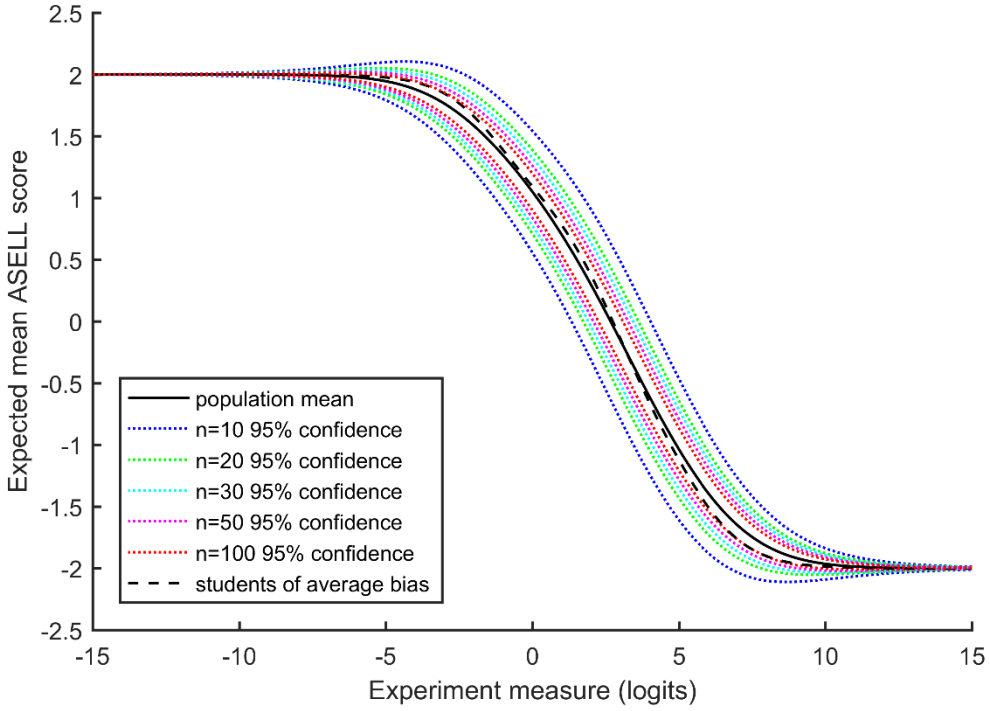


Figure S 26: Expected mean ASELL scores for item 7

7.3.11 Item 8: "The demonstrators offered effective supervision and guidance"

NOTE: Demonstrators are not experiment specific – students often (but not always) had the same demonstrator for every experiment. What the results for this question mean is open to interpretation. Potentially, the experiment measures reflect the difficulty of rating any given demonstrator in general positively for that experiment. However, the variable frequency of responses from different demonstrator groups may mean measures are biased to reflect the quality of the demonstrators who provided more survey responses in that experiment. A different type of analysis is likely required to analyse this question properly *(It is probably impossible to gather the data required for this).*

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 456 persons removed (mostly extreme positives)
3. Misfit issues resolved: 67 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 15 persons removed. (final data set results displayed)

Table S 48: Rasch model details for item 8

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label		Upper category label	measure	st. error		
		Agree	/	Strongly Agree	3.37	0.05	3.41	0.98
		Neutral	/	Agree	0.16	0.09	0.28	0.99
		Disagree	/	Neutral	-1.50	0.26	-1.31	1.12
		Strongly Disagree	/	Disagree	-2.04	0.74	-2.40	1.31

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C => M	Coherence M => C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	4.50	3.48	∞	73%	72%	1.02	1.00	0.4733	1095
1	Agree	1.80	0.42	3.48	68%	63%	1.01	1.02	0.4426	1061
0	Neutral	-0.51	-1.22	0.42	12%	32%	1.02	1.01	0.9291	161
-1	Disagree	-1.88	-2.72	-1.22	0%	0%	0.75	0.64	1.2888	15
-2	Strongly Disagree	-3.44	-∞	-2.72	0%	0%	0.44	0.30	1.3617	2

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	589	2.80	1.31	0.31	0.48	0.09	0.18	32.8%	32.5%
Experiments	33	0.00	0.97	1.65	1.75	0.73	0.75	4.4%	4.3%

Data points: 2334 Log-likelihood chi square: 3373.50 df: 1710

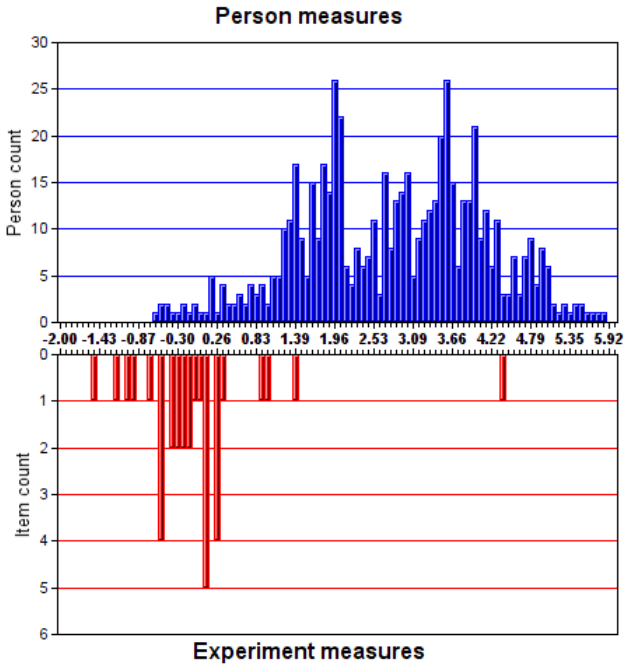


Figure S 27: Measure distributions for item 8

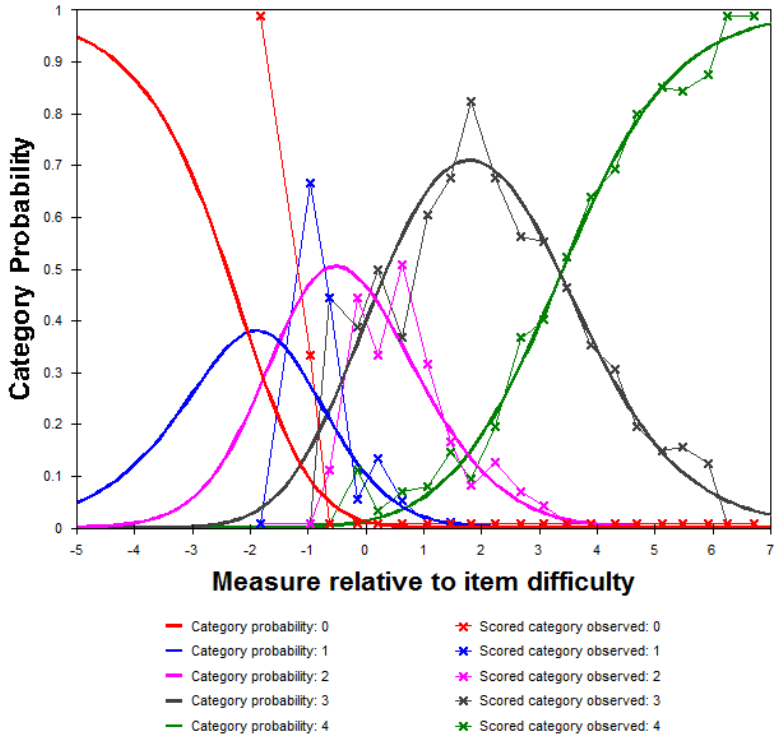


Figure S 28: Category structure for item 8

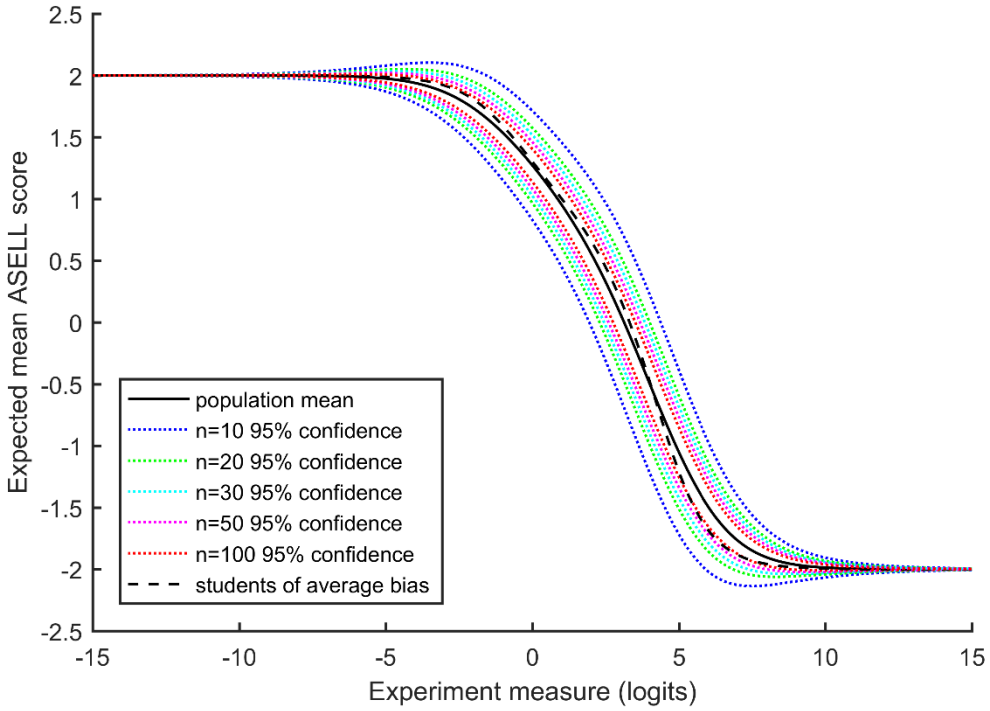


Figure S 29: Expected mean ASELL scores for item 8

7.3.12 Item 9: "The experimental procedure was clearly explained in the lab manual or notes"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 262 persons removed
3. Misfit issues resolved: 113 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Remaining data split into two subsets: one containing Foundations of Chemistry experiments, the other containing Chemistry IA experiments. To ensure connectivity, one previously removed misfitting person was added back into analysis. Resulting data had connectivity issues, but with the two cohorts still connected. Connectivity issues were resolved: 8 persons, 1 item removed (final results reported).

Table S 49: Rasch model details for item 9

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label	/	Upper category label	measure	st. error		
		Agree	/	Strongly Agree	2.90	0.05	2.95	1.06
		Neutral	/	Agree	-0.05	0.06	0.20	1.09
		Disagree	/	Neutral	-0.92	0.11	-0.95	0.98
		Strongly Disagree	/	Disagree	-1.94	0.24	-2.21	0.99

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C => M	Coherence M => C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	4.04	3.04	∞	50%	70%	1.01	1.00	0.6411	854
1	Agree	1.51	0.40	3.04	78%	61%	1.01	1.05	0.4001	1426
0	Neutral	-0.35	-0.97	0.40	30%	36%	1.04	1.04	0.8150	372
-1	Disagree	-1.61	-2.48	-0.97	18%	61%	0.91	0.86	1.2597	97
-2	Strongly Disagree	-3.25	−∞	-2.48	5%	100%	0.89	0.79	1.6746	20

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	745	1.91	1.19	0.20	0.44	0.04	0.16	34.7%	34.4%
Experiments	32	0.00	0.76	1.84	1.86	0.77	0.77	6.9%	6.8%

Data points: 2769 Log-likelihood chi square: 4838.56 df: 1990

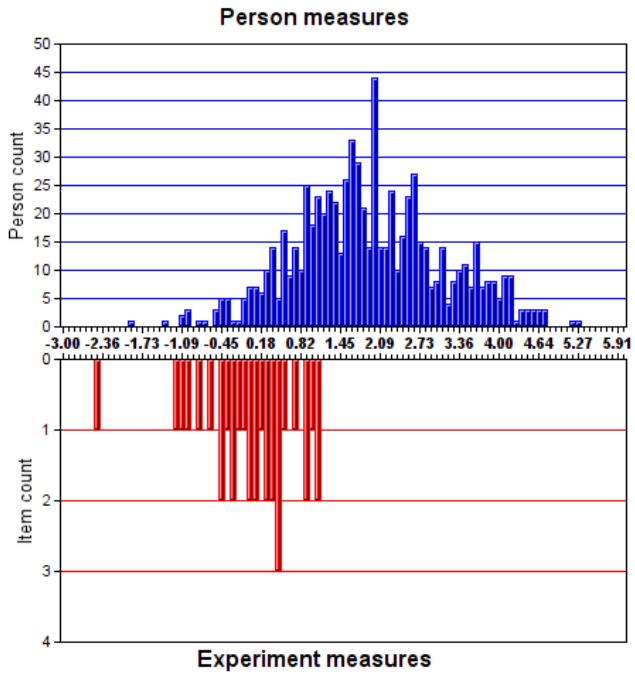


Figure S 30: Measure distributions for item 9

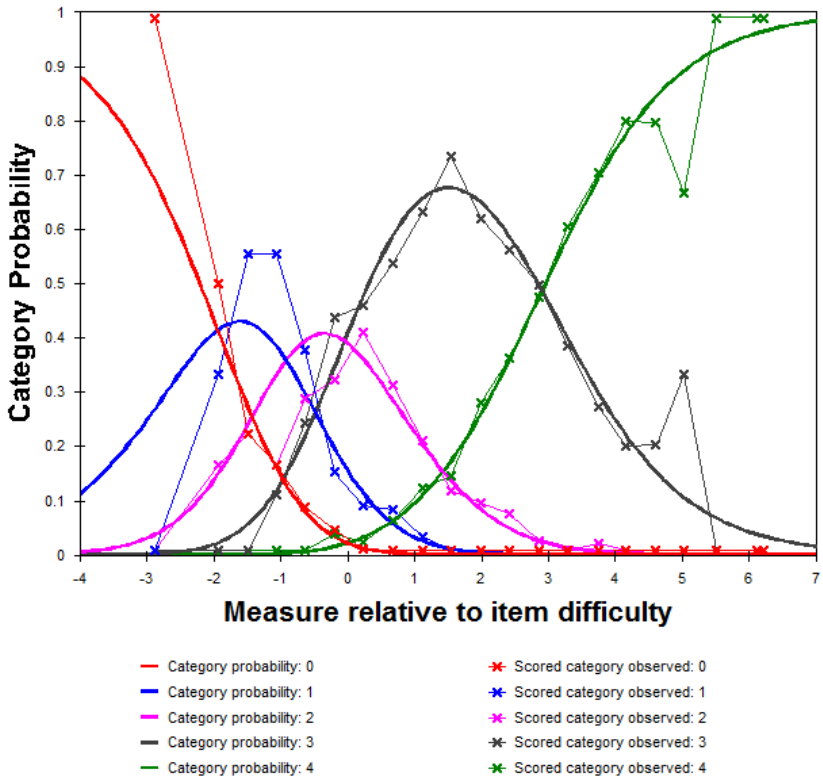


Figure S 31: Category structure for item 9

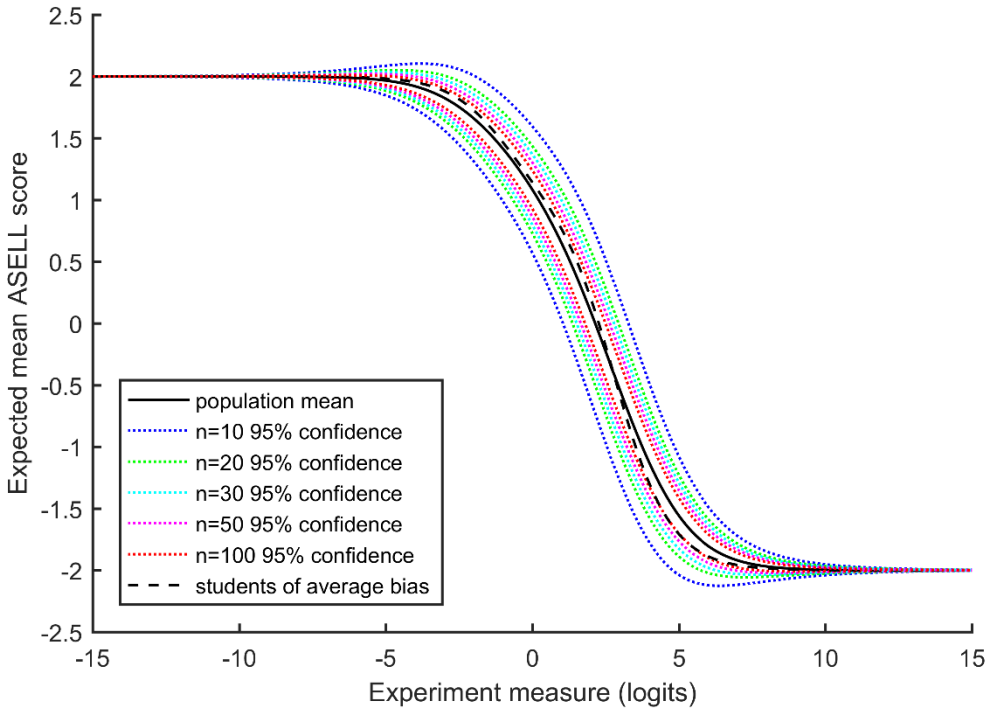


Figure S 32: Expected mean ASELL scores for item 9

7.3.13 Item 10: "I can see the relevance of this experiment to my chemistry studies"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 217 persons removed
3. Misfit issues resolved: 88 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 4 persons removed (final results reported)

Table S 50: Rasch model details for item 10

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label	/	Upper category label	measure	st. error		
		Agree	/	Strongly Agree	4.50	0.05	4.52	1.00
		Neutral	/	Agree	0.53	0.06	0.57	0.99
		Disagree	/	Neutral	-2.17	0.16	-1.92	1.05
		Strongly Disagree	/	Disagree	-2.86	0.40	-3.18	1.02

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C⇒M	Coherence M⇒C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Strongly Agree	5.62	4.56	∞	52%	69%	1.01	0.98	0.5983	744
1	Agree	2.52	0.64	4.56	84%	70%	0.97	1.00	0.3448	1655
0	Neutral	-0.72	-1.75	0.64	43%	57%	1.02	1.02	0.6927	427
-1	Disagree	-2.56	-3.47	-1.75	11%	50%	0.89	0.83	1.0334	44
-2	Strongly Disagree	-4.22	−∞	-3.47	0%	0%	0.98	0.94	1.6798	7

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	818	2.53	1.70	0.61	0.73	0.27	0.35	43.2%	42.9%
Experiments	33	0.00	0.66	1.56	1.58	0.71	0.71	4.1%	4.1%

Data points: 2877 Log-likelihood chi square: 4145.71 df: 2024

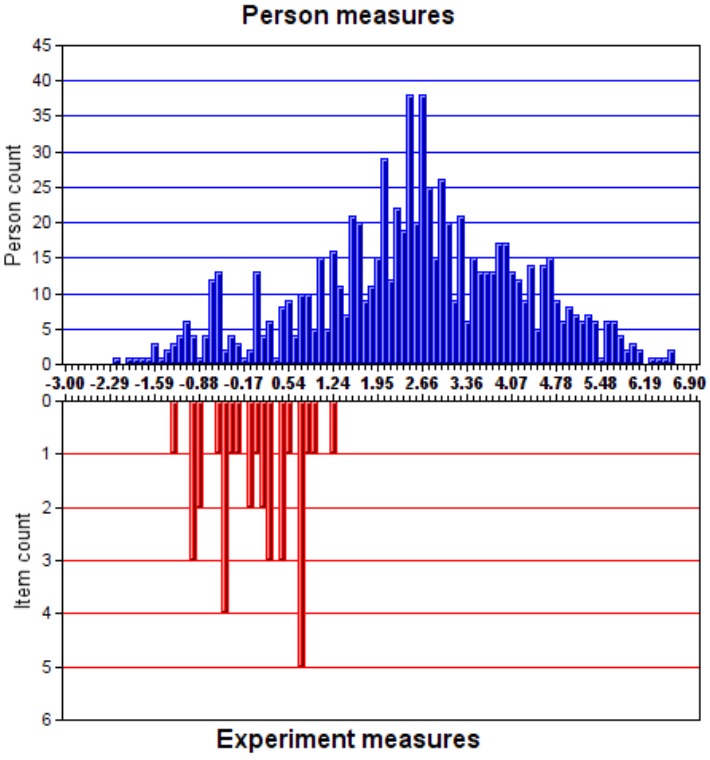


Figure S 33: Measure distributions for item 10

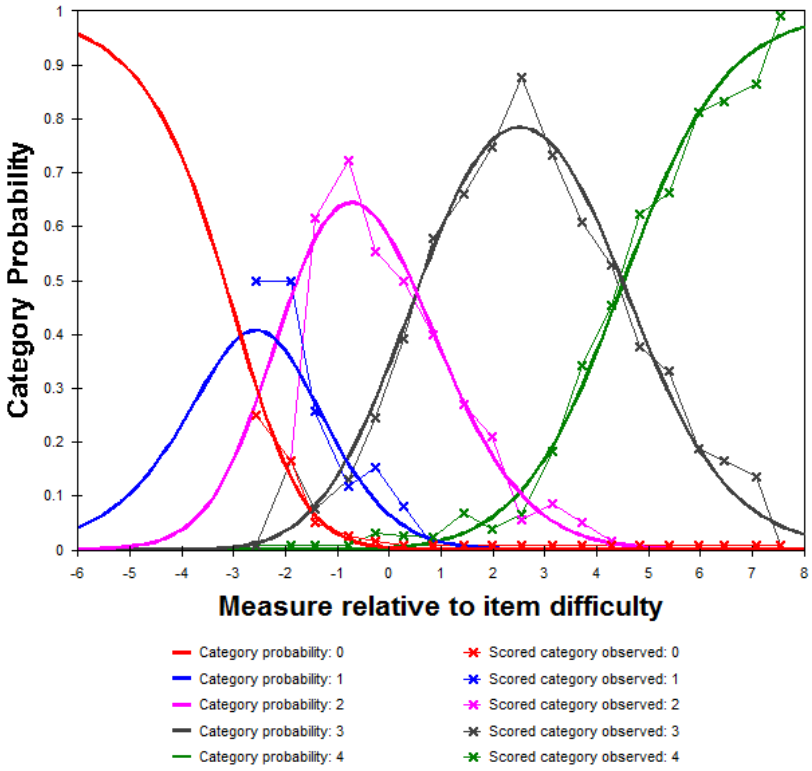


Figure S 34: Category structure for item 10

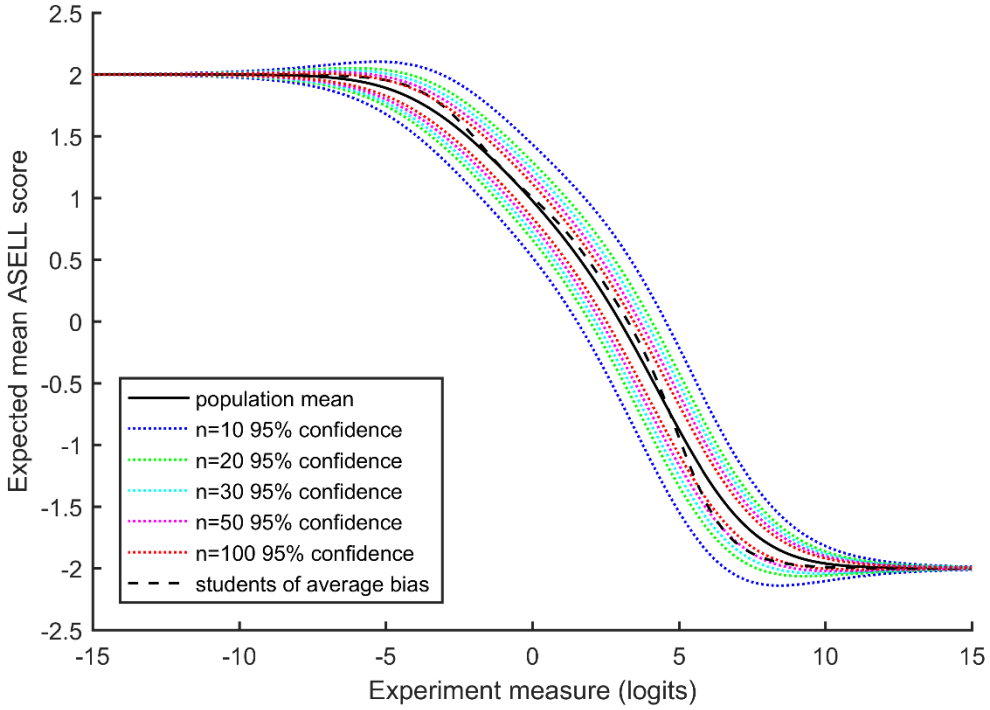


Figure S 35: Expected mean ASELL scores for item 10

7.3.14 Item 11: "Working in a team to complete this experiment was beneficial"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 339 persons removed. Student cohorts were assigned separate rating scale structures.
3. Misfit issues resolved: 149 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 16 persons and 1 item removed (final results reported).

Table S 51: Rasch model details for item 11

CHEMISTRY IA/B RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	2.66	0.07	2.85	0.91
	Neutral	/	Agree	1.27	0.08	1.09	1.14
	Disagree	/	Neutral	-3.27	0.20	-1.85	0.97
	Strongly Disagree	/	Disagree	-0.66	0.26	-2.11	0.79

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C=>M	Coherence M=>C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	3.91	3.07	∞	66%	73%	1.04	1.03	0.5625	609
1	Agree	1.97	0.91	3.07	57%	44%	1.04	1.15	0.5247	488
0	Neutral	-0.53	-1.44	0.91	60%	70%	0.88	0.82	0.6307	413
-1	Disagree	-2.00	-2.61	-1.44	0%	0%	1.18	1.15	1.1434	15
-2	Strongly Disagree	-3.10	-∞	-2.61	14%	60%	1.13	1.18	1.5514	22

FOUNDATIONS OF CHEMISTRY IA/B RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	2.95	0.08	3.04	0.91
	Neutral	/	Agree	0.65	0.10	0.64	1.12
	Disagree	/	Neutral	-1.85	0.23	-1.42	0.92
	Strongly Disagree	/	Disagree	-1.75	0.41	-2.28	0.89

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C=>M	Coherence M=>C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	4.12	3.17	∞	52%	67%	1.08	1.05	0.6413	304
1	Agree	1.83	0.64	3.17	75%	56%	0.94	1.04	0.4051	396
0	Neutral	-0.42	-1.22	0.64	40%	61%	0.86	0.84	0.7297	161
-1	Disagree	-1.88	-2.64	-1.22	26%	42%	1.25	1.30	1.2748	19
-2	Strongly Disagree	-3.30	-∞	-2.64	13%	100%	1.09	1.11	1.5286	8

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	623	1.96	1.24	0.52	0.69	0.21	0.32	27.2%	27.1%
Experiments	32	0.00	1.08	2.75	2.78	0.88	0.89	25.5%	25.5%

Data points: 2435 Log-likelihood chi square: 4080.24 df: 1775

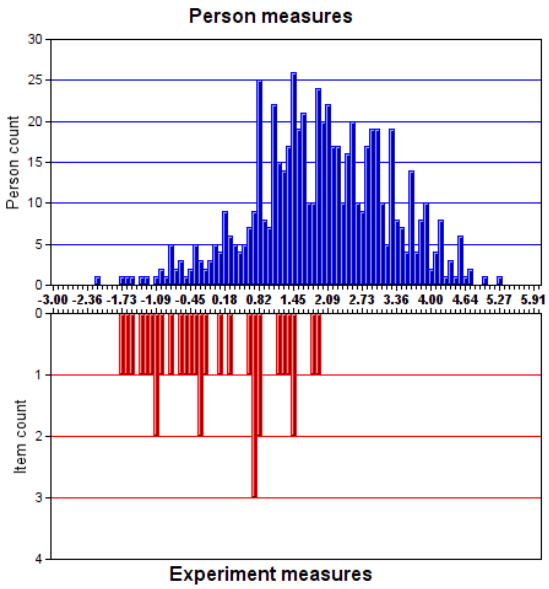


Figure S 36: Measure distributions for item 11

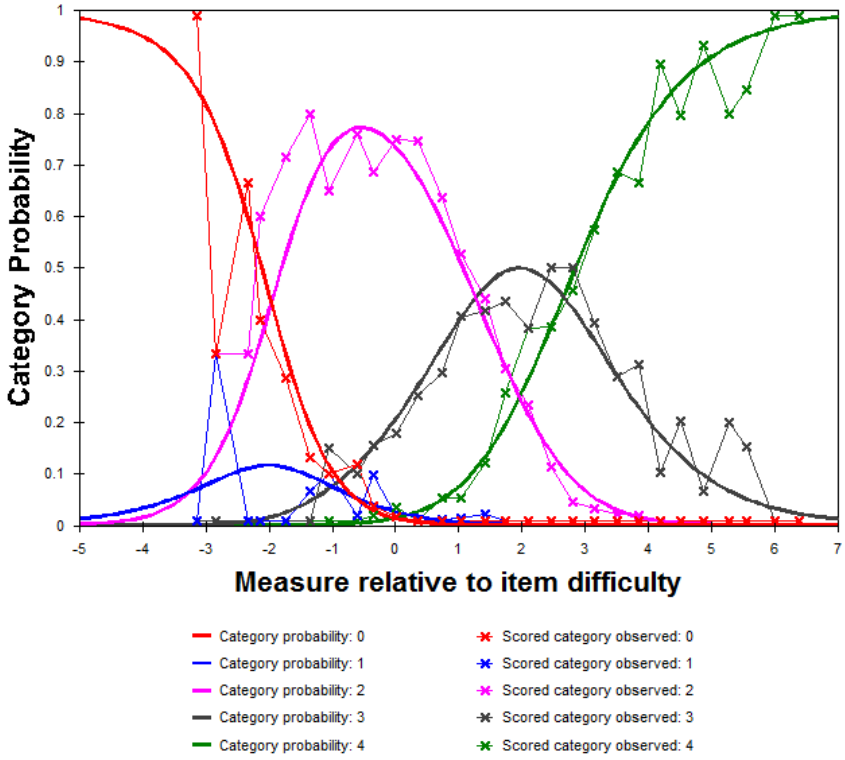


Figure S 37: Category structure for item 11 - Chemistry IA/B students

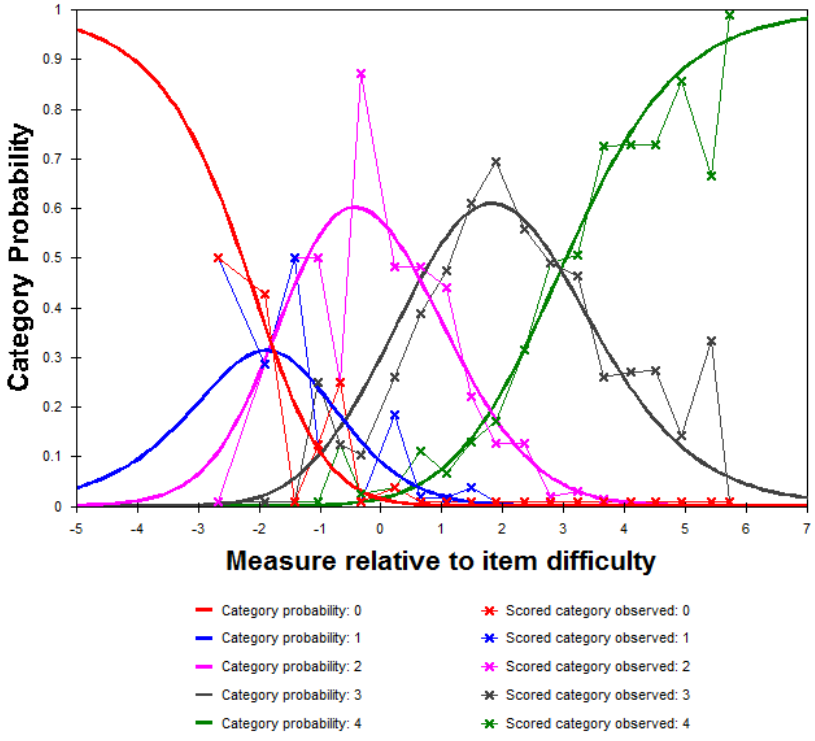


Figure S 38: Category structure for item 11 - Foundations of Chemistry IA/B students

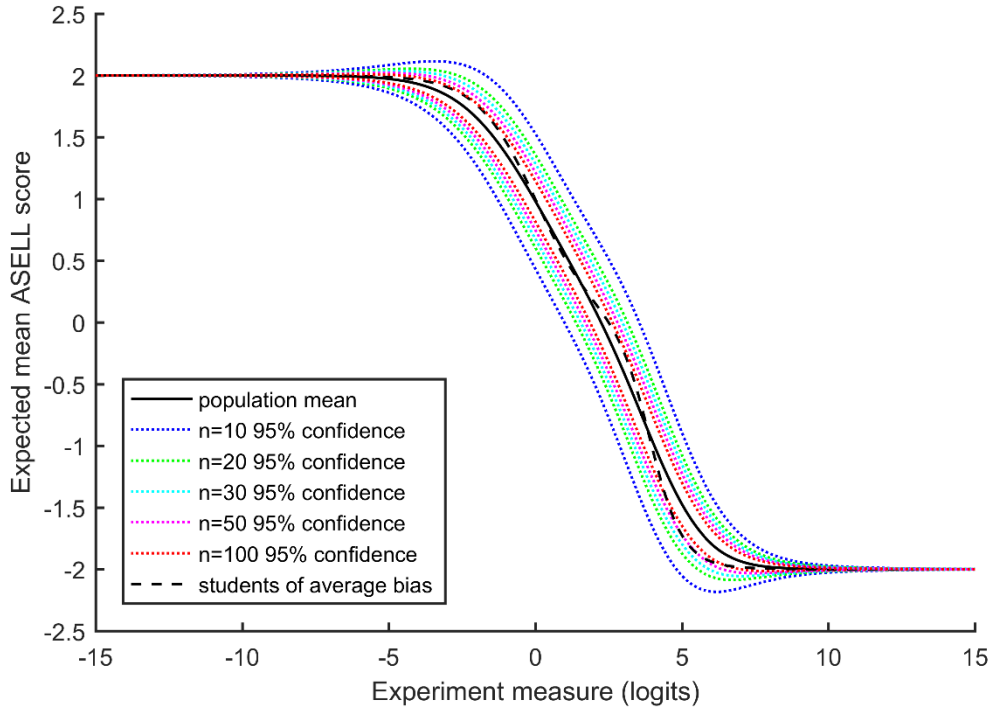


Figure S 39: Expected mean ASELL scores for item 11 - Chemistry IA/B students

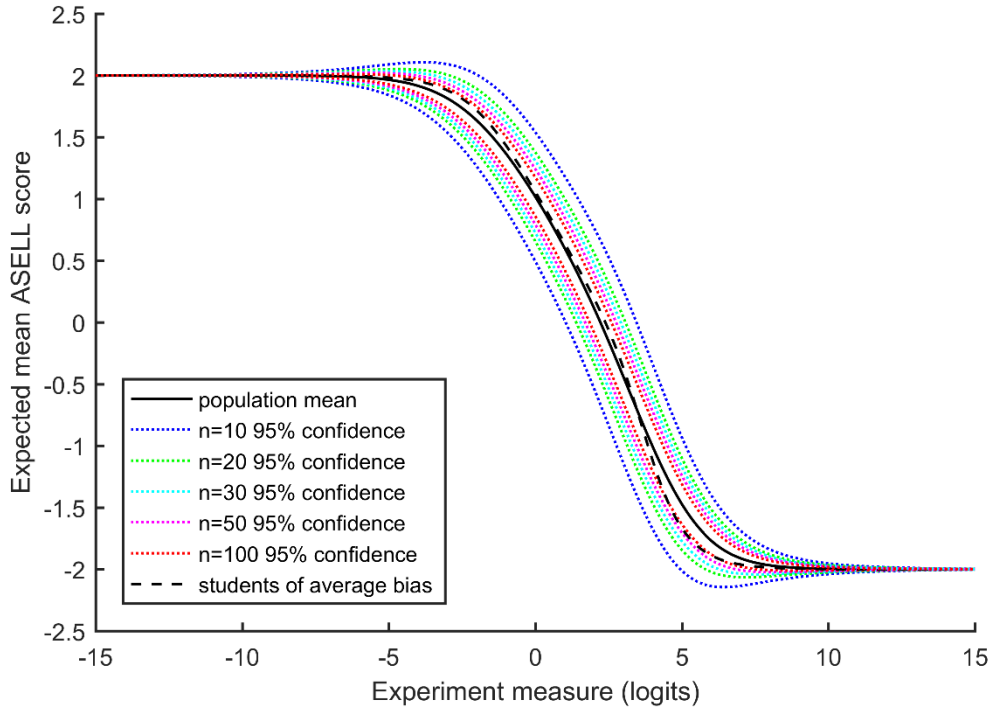


Figure S 40: Expected mean ASELL scores for item 11 - Foundations of Chemistry IA/B students

7.3.15 Item 12: "The experiment provided me with the opportunity to take responsibility for my own learning"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved:188 persons removed. Student cohorts were assigned separate rating scale structures.
3. Misfit issues resolved: 147 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 25 persons removed (final results reported).

Table S 52: Rasch model details for item 12

CHEMISTRY IA/B RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	4.84	0.07	4.87	1.01
	Neutral	/	Agree	1.64	0.07	1.61	1.02
	Disagree	/	Neutral	-3.19	0.28	-2.74	0.95
	Strongly Disagree	/	Disagree	-3.29	0.64	-3.74	1.20

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C=>M	Coherence M=>C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	5.97	4.94	∞	44%	70%	1.00	0.99	0.6677	371
1	Agree	3.24	1.56	4.94	85%	64%	0.96	0.94	0.3475	864
0	Neutral	-0.71	-2.42	1.56	50%	73%	1.00	1.02	0.6319	408
-1	Disagree	-3.25	-4.09	-2.42	7%	17%	1.23	1.09	1.0878	14
-2	Strongly Disagree	-4.78	−∞	-4.09	0%	0%	0.81	0.71	1.3060	3

FOUNDATIONS OF CHEMISTRY IA/B RESPONSE CATEGORY ASSOCIATED STATISTICS	Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
	Lower category label		Upper category label	measure	st. error		
	Agree	/	Strongly Agree	6.30	0.11	6.31	0.93
	Neutral	/	Agree	1.95	0.10	1.94	1.06
	Disagree	/	Neutral	-4.60	0.48	-3.82	0.93
	Strongly Disagree	/	Disagree	-3.65	1.06	-4.43	0.81

ASELL score	Category Label	Category measure	Range from	Range to	Coherence C=>M	Coherence M=>C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Strongly Agree	7.41	6.34	∞	44%	64%	1.08	1.06	0.6710	162
1	Agree	4.12	1.92	6.34	83%	72%	1.00	0.99	0.3391	567
0	Neutral	-1.24	-3.41	1.92	68%	76%	0.92	0.91	0.4872	288
-1	Disagree	-4.13	-4.85	-3.41	0%	0%	1.28	1.07	1.0840	4
-2	Strongly Disagree	-5.44	−∞	-4.85	0%	0%	1.25	0.94	1.8169	1

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	767	3.41	1.99	1.19	1.30	0.36	0.42	45.1%	45.0%
Experiments	33	0.00	0.76	1.38	1.42	0.66	0.67	4.3%	4.3%

Data points:	2682	Log-likelihood chi square:	3799.33	df:	1877

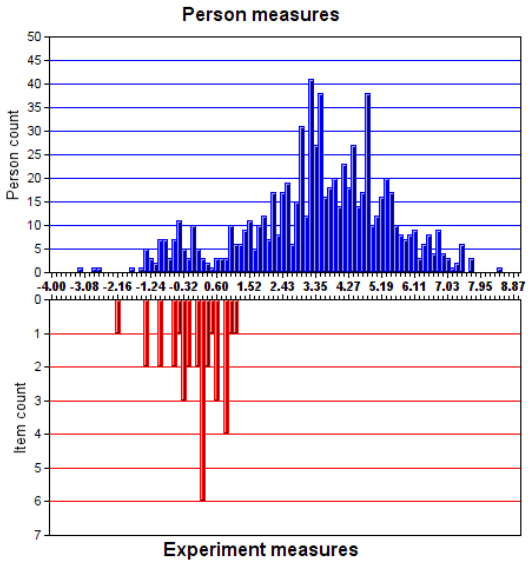


Figure S 41: Measure distributions for item 12

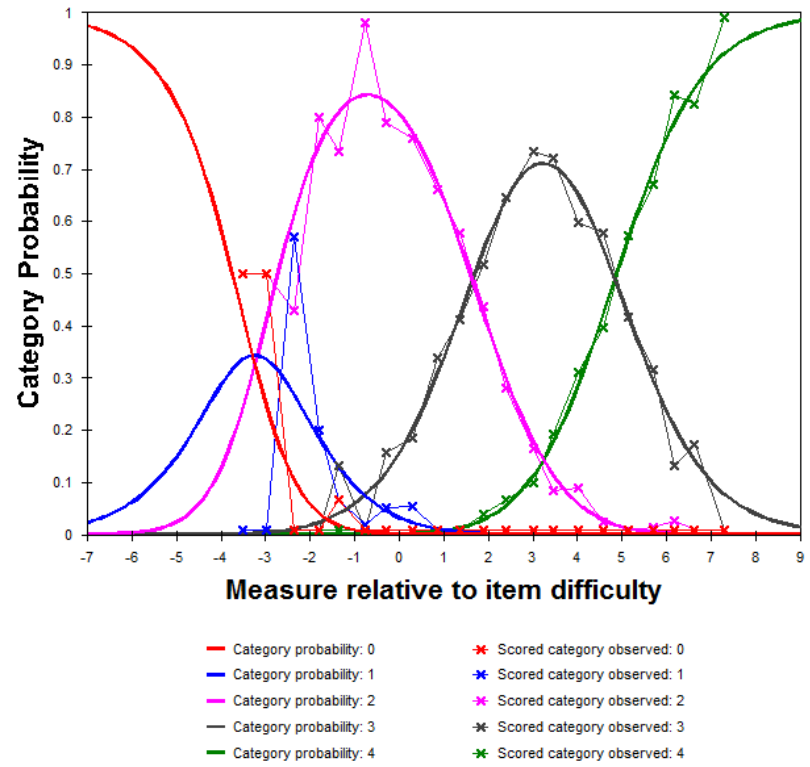


Figure S 42: Category structure for item 12 - Chemistry IA/B students

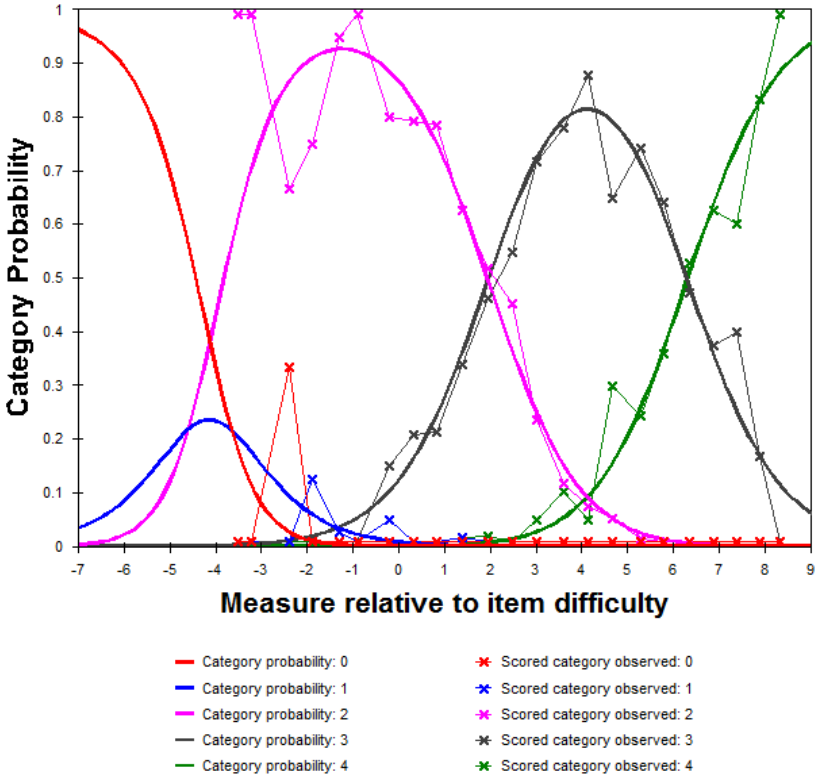


Figure S 43: Category structure for item 12 - Foundations of Chemistry IA/B students

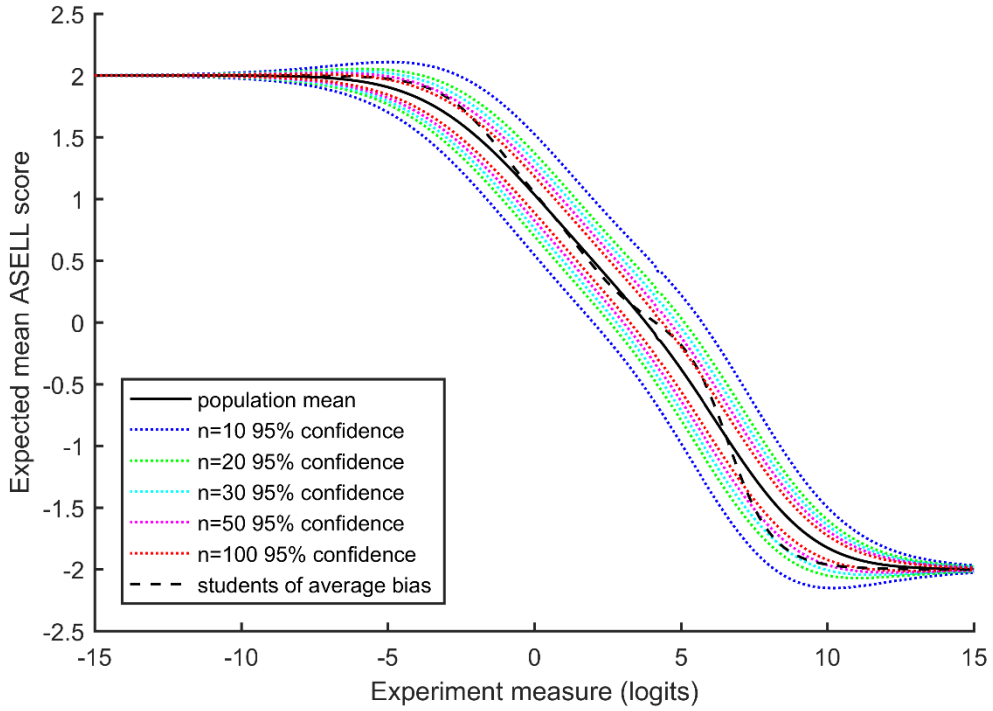


Figure S 44: Expected mean ASELL scores for item 12 - Chemistry IA/B students

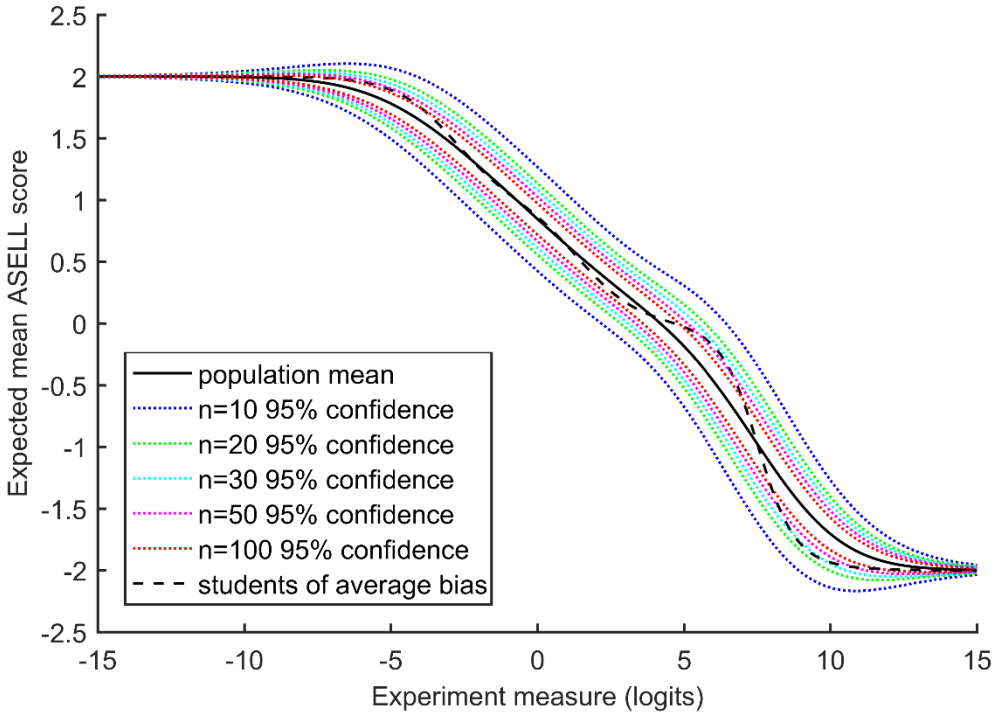


Figure S 45: Expected mean ASELL scores for item 12 - Foundations of Chemistry IA/B students

7.3.16 Item 13: "I found the time available to complete this experiment was"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 124 persons and 2 items removed
3. Further connectivity issues and blank responses resolved: 103 persons and 5 items removed
4. Further connectivity issues and blank responses resolved: 10 persons and 1 item removed
5. Misfit issues resolved: 115 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
6. Further connectivity issues resolved: 36 persons removed (final results presented).

Table S 53: Rasch model details for item 13

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label	/	Upper category label	measure	st. error		
		Too Much	/	Way Too Much	7.82	0.19	7.85	1.05
		About Right	/	Too Much	4.26	0.07	4.23	1.04
		Not Enough	/	About Right	-4.96	0.11	-4.86	0.94
		Nowhere Near Enough	/	Not Enough	-7.12	0.33	-7.22	1.14
ASELL score	Category Label	Category measure	Range from	Range to	Coherence C=>M	Coherence M=>C	Fit Statistics Infit	Fit Statistics Outfit	RMSR	counts in sampled data
2	Way Too Much	8.94	7.90	∞	15%	50%	0.85	0.82	0.8439	40
1	Too Much	6.04	4.18	7.9	57%	65%	0.96	0.88	0.5478	396
0	About Right	-0.35	-4.72	4.18	95%	91%	0.97	0.98	0.2008	2388
-1	Not Enough	-6.04	-7.36	-4.72	37%	58%	1.04	0.92	0.7213	112
-2	Nowhere Near Enough	-8.3	−∞	-7.36	9%	100%	0.75	0.51	1.0435	11

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures: empirical	Raw variance in observed data explained by measures: modelled
Persons	739	0.03	2.29	0.00	0.00	0.00	0.00	38.6%	37.1%
Experiments	25	0.00	2.08	4.09	4.14	0.94	0.94	14.7%	14.1%

Data points: 2947 Log-likelihood chi square: 2085.72 df: 2181

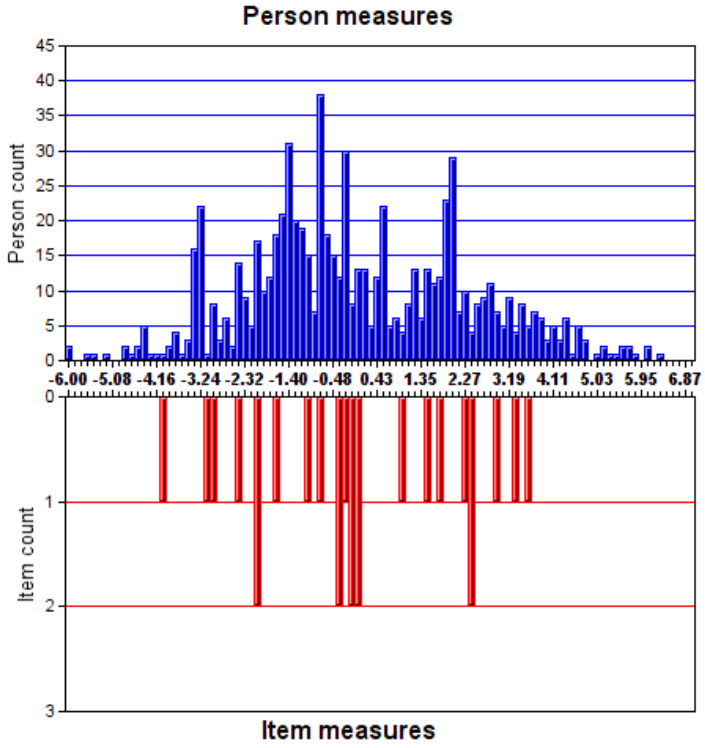


Figure S 46: Measure distributions for item 13

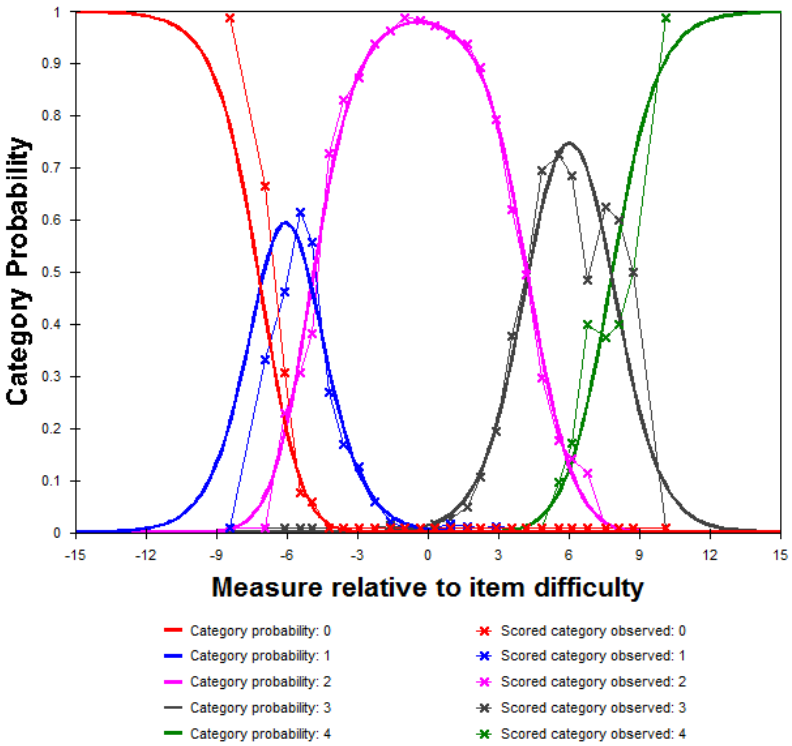


Figure S 47: Category structure for item 13

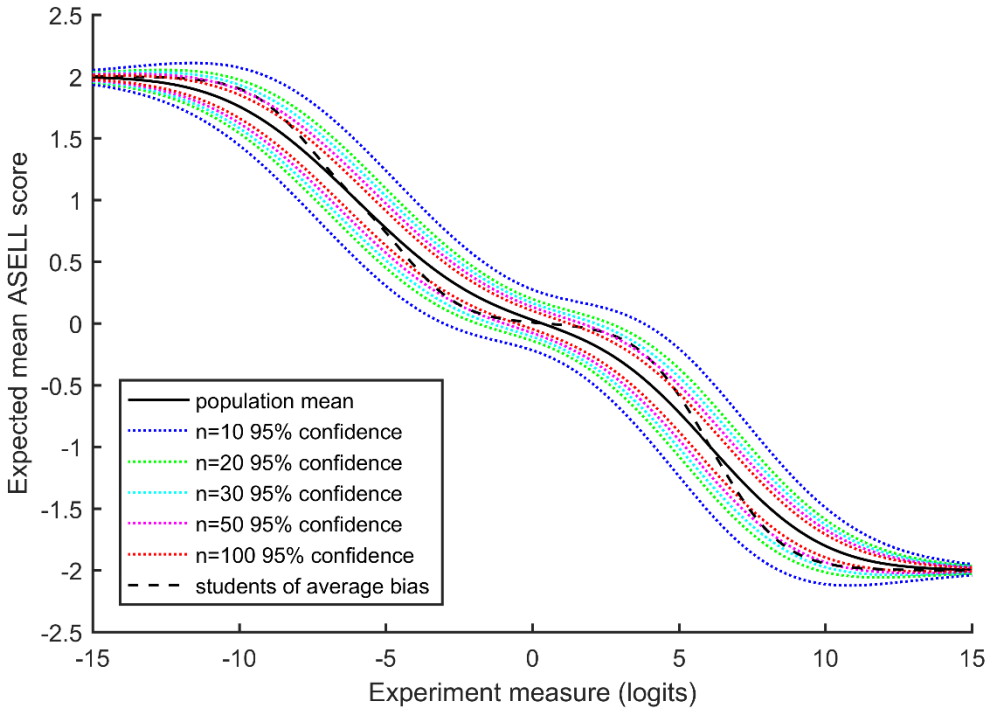


Figure S 48: Expected mean ASELL scores for item 13

7.3.17 Item 14: "Overall, as a learning experience, I would rate this experiment as"

Data preparation/ analyses run

1. **Initial data:** 1127 persons, 33 items
2. Connectivity issues, extreme persons and blank responses resolved: 153 persons removed
3. Misfit issues resolved: 88 persons with z-scores for infit or outfit $|z| \geq 2$ were removed.
4. Further connectivity issues resolved. 51 persons and 1 item removed (final results presented).

Table S 54: Rasch model details for item 14

RESPONSE CATEGORY ASSOCIATED STATISTICS		Category threshold between			Andrich threshold		Thurstone threshold	Estimated discrimination
		Lower category label		Upper category label	measure	st. error		
		Good	/	Excellent	5.72	0.06	5.72	0.98
		Average	/	Good	0.13	0.06	0.17	1.02
		Poor	/	Average	-3.06	0.18	-2.55	0.99
		Very Poor	/	Poor	-2.79	0.43	-3.36	1.25
ASELL score	Category Label	Category measure	Range from	Range to	Coherence C=>M	Coherence M=>C	Fit Statistics Infit	Fit Statistics Outfit	Fit Statistics RMSR	counts in sampled data
2	Excellent	6.82	5.73	∞	46%	68%	1.04	0.98	0.6269	457
1	Good	2.93	0.23	5.73	89%	78%	0.97	0.98	0.2973	1981
0	Average	-1.28	-2.27	0.23	48%	60%	0.97	0.94	0.5933	474
-1	Poor	-2.97	-3.74	-2.27	7%	50%	1.09	1.11	1.2291	29
-2	Very Poor	-4.38	-∞	-3.74	0%	0%	0.75	0.59	1.6048	6

BROAD SCALE STATISTICS	Number used for estimates	Measures mean	Measures st. dev	separation observed	separation model	reliability observed	reliability model	Raw variance in observed data explained by measures empirical	Raw variance in observed data explained by measures modelled
Persons	835	2.69	2.02	0.39	0.51	0.13	0.21	45.2%	44.9%
Experiments	33	0.00	1.02	2.01	2.04	0.80	0.81	1.9%	1.8%

Data points: 2947 Log-likelihood chi square: 3501.83 df: 2078

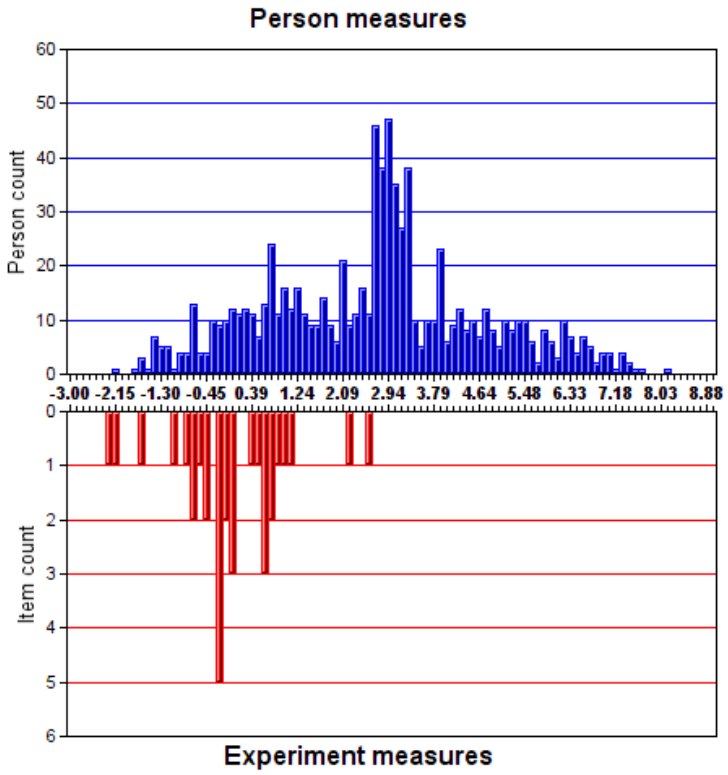


Figure S 49: measure distributions for item 14

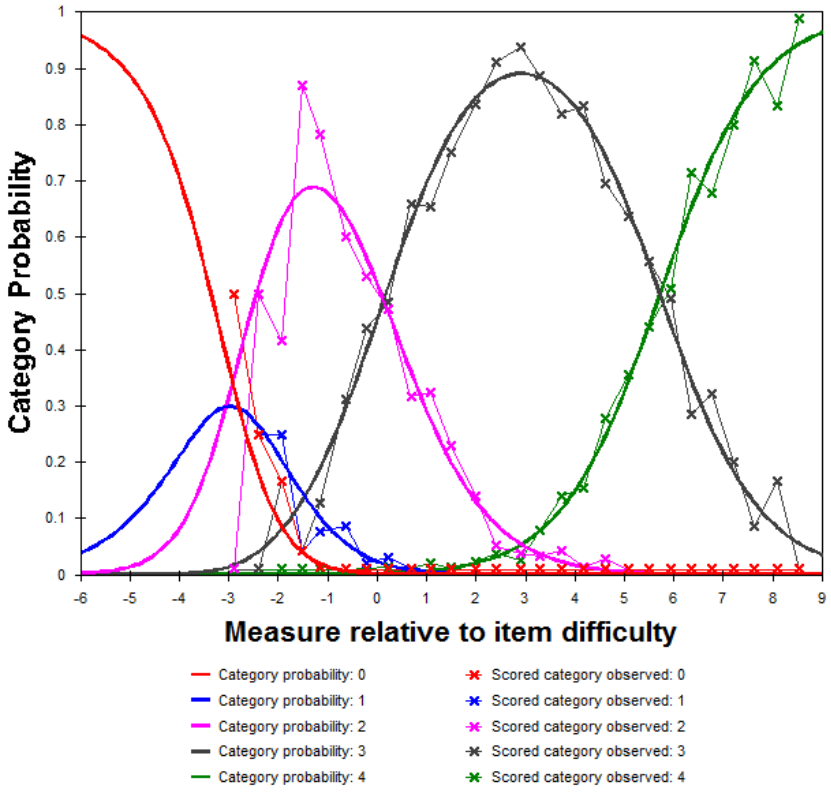


Figure S 50: Category structure for item 14

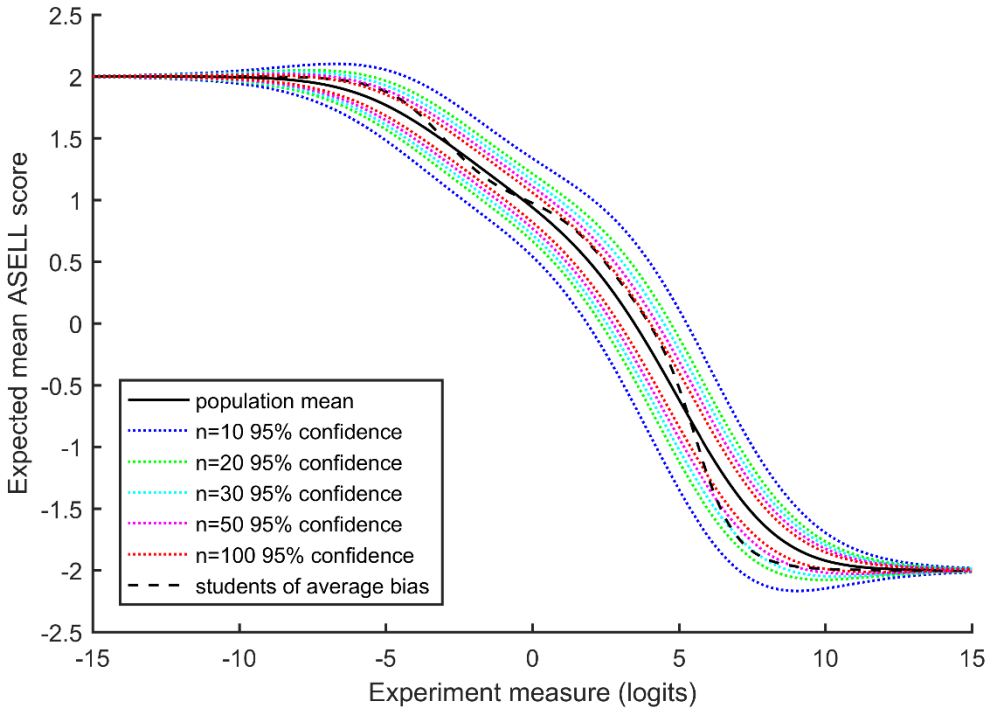


Figure S 51: Expected mean ASELL scores for item 14

7.4 Supporting information for section 4.1

7.4.1 Rasch model derivations

Included in this section are simple derivations of each Rasch model used in this thesis, notably contrasted in section 4.1. Equations labelled with white text within a black box (for example A) correspond to Rasch model formulations used in the main body of research, as presented in Table 18 of section 4.1.2.2.

Symbols used in the model derivations following differ to the facet symbols used in the main discussion. This is because facets are constructed in the main body discussion based on how their element numbers are assigned, rather than their precise mathematical origins. In some cases, a variety of different facet terms presented here vary in the identical way, hence are assigned facet element numbers identically and thus have identical measures estimated, all else being equal. It is appropriate therefore to use the same facet symbol in the main body research in these cases, despite possible differences in their basic formulation within mathematical derivations. A summary of the variety of symbols used to refer to different types of facet is provided below.

Table S 55: Symbols used to represent different facets in various Rasch model formulations

	Symbol used in main body discussion	Facet description	Facets which vary in this manner, as noted in these model derivations
STUDENT INDEPENDENT (see Table 16)	**Q**	Values vary between questions only, otherwise constant	δ_q
	E	Values vary between contexts (experiments) only, otherwise constant	δ_c
	δ	Values vary between questions and contexts (experiments) only, otherwise constant	$\delta_{q,c}$, $\delta'_{q,c}$, $\delta^*_{q,c}$, $\delta^{**}_{q,c}$
STUDENT DEPENDENT (see Table 17)	**β**	Values vary between students only, otherwise constant	β_s
	β_Q	Values vary between students and questions only, otherwise constant	$\beta_{s,q}$, $\beta'_{s,q}$, $\beta^*_{s,q}$, $\beta^{**}_{s,q}$
	β_E	Values vary between students and contexts (experiments) only, otherwise constant	$\beta_{s,c}$, $\beta'_{s,c}$, $\beta^*_{s,c}$, $\beta^{**}_{s,c}$

Symbols used in the main body discussion imply the way facet element numbers are assigned. A greater number of symbols is needed to express the different facets which vary similarly, but with different mathematical justifications. Assignment of facet element numbers for each is shown in Table 16 and Table 17, section 4.1.2.2.

The following derivations express the probability P that the observed response X will occur in category k of rating scale structure g (i.e. in category $x_{g,k}$), under survey circumstances described by vector $\vec{D} = [s \quad q \quad c]$. Here, s, q and c index the student responding, the question they are responding to and the context in which the student is asked that question respectively.

Consider the probability of response in category k relative to the probability of response in the previous category, using this as a relative measure of the tendency to observe positive response in a given circumstance:

$$\frac{P\left(X = x_{g,k} | \vec{D} = [s \quad q \quad c]\right)}{P\left(X = x_{g,k-1} | \vec{D} = [s \quad q \quad c]\right)} = \pi_{g,k}(\vec{D}) = \pi_{g,k}(s, q, c)$$

The tool used in order to obtain a measurement should not influence the measure value obtained. It is therefore required that the measure of one circumstance described by $\vec{D}_a = [s_a \quad q_a \quad c_a]$ relative to another circumstance described by $\vec{D}_b = [s_b \quad q_b \quad c_b]$ remains equivalent regardless of which rating scale structure or rating scale category is used. Using some hypothetical "reference" rating scale structure (g_0) and category (k_0), this requirement implies that for any rating scale structure (g) and response category (k):

$$\frac{\pi_{g,k}(s_a, q_a, c_a)}{\pi_{g,k}(s_b, q_b, c_b)} = \frac{\pi_{g_0,k_0}(s_a, q_a, c_a)}{\pi_{g_0,k_0}(s_b, q_b, c_b)}$$

and hence,

$$\pi_{g,k}(s_a, q_a, c_a) = \pi_{g,k}(s_b, q_b, c_b) \times \frac{\pi_{g_0,k_0}(s_b, q_b, c_b)}{\pi_{g_0,k_0}(s_b, q_b, c_b)}$$

Defining some arbitrary "reference" circumstance $\vec{\Phi} = [\phi_s \quad \phi_q \quad \phi_c]$ with which other circumstances may be contrasted for the purposes of measurement, then substituting this reference circumstance in place of $\vec{D}_b = [s_b \quad q_b \quad c_b]$ in the equation above obtains:

$$\pi_{g,k}(s, q, c) = \pi_{g,k}(\vec{\Phi}) \times \frac{\pi_{g_0,k_0}(s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}$$

which, after taking the natural logarithm,

$$ln\left[\pi_{g,k}(s, q, c)\right] = ln\left[\pi_{g,k}(\vec{\Phi})\right] + ln\left[\frac{\pi_{g_0,k_0}(s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right]$$

yields the generalised Rasch model first presented as Equation 1 in the introductory material (replacing $\pi_{g,k}(s, q, c)$ with the full probability ratio described at the beginning of this discussion).

$$ln\left[\frac{P\left(X = x_{g,k} | \vec{D} = [s \quad q \quad c]\right)}{P\left(X = x_{g,k-1} | \vec{D} = [s \quad q \quad c]\right)}\right] = \varphi_{s,q,c} - \tau_{g,k} \qquad \textbf{1}$$

Here the latent trait measure underpinning response in circumstance $\vec{D} = [s \quad q \quad c]$ is given by:

$$\varphi_{s,q,c} = ln\left[\frac{\pi_{g_0,k_0}(s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right]$$

and the rating scale structure is defined by parameters:

$$\tau_{g,k} = ln\left[\pi_{g,k}(\vec{\Phi})^{-1}\right]$$

The latent trait parameter may be reformulated at this point by introducing various constraints, restricting the way measure values change as descriptions of the circumstance change. Requiring that the measure for one student relative to a hypothetical "reference student" (ϕ_s) is independent of question and context (i.e. requiring "specific objectivity" of the student measure) necessitates:

$$\frac{\pi_{g_0,k_0}(s,q,c)}{\pi_{g_0,k_0}(\phi_s,q,c)} = \frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\phi_s,\phi_q,\phi_c)}$$

which therefore implies:

$$\pi_{g_0,k_0}(s,q,c) = \pi_{g_0,k_0}(\phi_s,q,c) \times \frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\phi_s,\phi_q,\phi_c)}$$

and hence, substituting back into the expression for the latent trait measure:

$$\varphi_{s,q,c} = ln\left[\frac{\pi_{g_0,k_0}(s,q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] = ln\left[\frac{\pi_{g_0,k_0}(\phi_s,q,c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right]$$

Or, introducing simplified variable labels:

$$\varphi_{s,q,c} = \beta_s - \delta_{q,c} \quad ; \quad \beta_s = ln\left[\frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] \quad , \quad \delta_{q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(\phi_s,q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1}\right] \qquad \textbf{A}$$

which is a two-facet Rasch model containing one facet specific to student only (β_s) and another specific to both question and context ($\delta_{q,c}$). Similar procedures can be used to obtain other two-facet Rasch models, either by requiring specific objectivity with respect to question asked:

$$\varphi_{s,q,c} = \beta_{s,c} - \delta_q \quad ; \quad \beta_{s,c} = ln\left[\frac{\pi_{g_0,k_0}(s,\phi_q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] \quad , \quad \delta_q = ln\left[\left(\frac{\pi_{g_0,k_0}(\phi_s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1}\right] \qquad \textbf{B}$$

or survey context:

$$\varphi_{s,q,c} = \beta_{s,q} - \delta_c \quad ; \quad \beta_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] \quad , \quad \delta_c = ln\left[\left(\frac{\pi_{g_0,k_0}(\phi_s,\phi_q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1}\right] \qquad \textbf{C}$$

These all produce two-facet Rasch models, with facets varying in different ways. Facets derived in the latter two formulations include a context-specific student facet ($\beta_{s,c}$), a question facet (δ_q), a question-specific student facet ($\beta_{s,q}$) and a context-specific facet (δ_c). Any of these formulations can be further used to derive a simple three-facet Rasch model by introducing further specific objectivity requirements. For example, specific objectivity with respect to the survey question posed requires that:

$$\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\phi_s,q,\phi_c)} = \frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\phi_s,\phi_q,\phi_c)}$$

i.e.

$$\pi_{g_0,k_0}(s,q,\phi_c) = \pi_{g_0,k_0}(\phi_s,q,\phi_c) \times \frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}$$

and therefore, the $\beta_{s,q}$ facet can be reformulated as follows:

$$\beta_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] = ln\left[\frac{\pi_{g_0,k_0}(s,\phi_c,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(\phi_s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] = \beta_s - \delta_q$$

with similar reformulations for the $\beta_{s,c}$ or $\delta_{q,c}$ facets. The net result is that all three of the prior latent trait parameter expressions may be reformulated as:

$$\varphi_{s,q,c} = \beta_s - \delta_q - \delta_c \qquad \textbf{D}$$

which is a three-facet Rasch model containing a student facet, question facet and context facet. Thus, four different expressions for the latent trait parameter have been derived. A fifth expression may be derived by considering the sum of the first three latent trait parameter expressions discovered previously. The expression below requires all specific objectivity restraints thus far introduced:

$$\varphi_{s,q,c} + \varphi_{s,q,c} + \varphi_{s,q,c} = (\beta_s - \delta_{q,c}) + (\beta_{s,c} - \delta_q) + (\beta_{s,q} - \delta_c)$$

or more simply,

$$3\varphi_{s,q,c} = (\beta_{s,q} + \beta_{s,c} - \delta_{q,c}) + (\beta_s - \delta_q - \delta_c)$$

The simple three-facet model expression presented previously can easily be subtracted on both sides of this equation to yield:

$$2\varphi_{s,q,c} = \beta_{s,q} + \beta_{s,c} - \delta_{q,c}$$

And therefore:

$$\varphi_{s,q,c} = \frac{1}{2}\beta_{s,q} + \frac{1}{2}\beta_{s,c} - \frac{1}{2}\delta_{q,c}$$

This expression contains all three facet terms which are jointly specific to more than one component of the circumstance description. The coefficient of ½ outside of each facet term can conveniently be removed by re-labelling variables. A complex three-facet Rasch model can then be expressed:

$$\varphi_{s,q,c} = \beta'_{s,q} + \beta'_{s,c} - \delta'_{q,c} \qquad \textbf{E}$$

where the facet measures then reflect the following:

$$\beta'_{s,q} = \frac{1}{2}\beta_{s,q} = ln\left[\left(\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right]$$

$$\beta'_{s,c} = \frac{1}{2}\beta_{s,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s,\phi_q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right]$$

$$\delta_{q,c} = \frac{1}{2}\delta_{q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(\phi_s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1/2}\right]$$

rather than the usual relationships as presented in previous discussion. Restrictions can be imposed on the parameters of either this complex three facet model or the previously described simple three-facet model to obtain other Rasch model formulations.

For example, assume a scenario such that the question asked does not alter the latent trait parameter. That is, for all s, q and c:

$$\frac{\pi_{g_0,k_0}(s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})} = \frac{\pi_{g_0,k_0}(s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}$$

This restriction reduces the previously described δ_q facet to zero by the following:

$$\delta_q = ln\left[\left(\frac{\pi_{g_0,k_0}(\phi_s, q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1}\right] = ln\left[\left(\frac{\pi_{g_0,k_0}(\vec{\Phi})}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1}\right] = 0$$

Therefore, under this presumption, the previously presented simple three-facet model is reduced to a simple two-facet model:

$$\varphi_{s,q,c} = \beta_s - \delta_c \qquad \textbf{F}$$

The identical presumption also restricts the complex three-facet model to the identical formulation. Beginning with the complex three-facet model (written in expanded form below):

$$\varphi_{s,q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s, q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] + ln\left[\left(\frac{\pi_{g_0,k_0}(s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] - ln\left[\left(\frac{\pi_{g_0,k_0}(\phi_s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{-1/2}\right]$$

which simplifies to:

$$\varphi_{s,q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s, q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(\phi_s, q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right]$$

Applying the restriction described previously yields:

$$\varphi_{s,q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s, \phi_q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(\phi_s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right]$$

Again applying previously stated specific objectivity requirements, this expands to give:

$$\varphi_{s,q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s, \phi_q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(s, \phi_q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(\phi_s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(\phi_s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right]$$

which simplifies to Equation F, as was to be shown:

$$\varphi_{s,q,c} = ln\left[\frac{\pi_{g_0,k_0}(s, \phi_q, \phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \frac{\pi_{g_0,k_0}(\phi_s, \phi_q, c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right] = \beta_s - \delta_c$$

Presuming instead that latent trait measure is independent of survey context, procedures analogous to those above may be employed to give another alternate model:

$$\varphi_{s,q,c} = \beta_s - \delta_q \qquad \textbf{G}$$

The complex three-facet Rasch model can also be restricted in other ways, however. Any one of the three complex facets can be presumed zero. That is, it may be assumed that either of the following are true:

$$\beta'_{s,q} = ln\left[\left(\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] = ln\left[\left(\frac{\pi_{g_0,k_0}(\vec{\Phi})}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] = 0 \quad \forall s,q$$

$$\beta'_{s,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s,\phi_q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] = ln\left[\left(\frac{\pi_{g_0,k_0}(\vec{\Phi})}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] = 0 \quad \forall s,c$$

$$\delta'_{q,c} = ln\left[\left(\frac{\pi_{g_0,k_0}(s,\phi_q,c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] = ln\left[\left(\frac{\pi_{g_0,k_0}(\vec{\Phi})}{\pi_{g_0,k_0}(\vec{\Phi})}\right)^{1/2}\right] = 0 \quad \forall q,c$$

which respectively give the three models shown below:

$$\varphi_{s,q,c} = \beta'_{s,c} - \delta'_{q,c} \qquad \textbf{H}$$

$$\varphi_{s,q,c} = \beta'_{s,q} - \delta'_{q,c} \qquad \textbf{I}$$

$$\varphi_{s,q,c} = \beta'_{s,q} + \beta'_{s,c}$$

These three models are all formulated such that all terms are dependent on one “overarching” component of the survey circumstance description (the context, question or student respectively). Note that the third does not include any student independent term, and so has not been labelled. These models represent cases where the data are described using the same two facets, but the measures of those facets are not comparable for different cases of the “overarching” variable. For example, Model H above has a student facet and a question facet, which remain comparable for some specific survey context. However, once the survey context changes, the student facet and question facet may take on different values. A lack of connectivity is therefore necessitated: measures are not comparable across different survey contexts. Similar scenarios exist for the other two models of the three presented above. This necessary disconnectivity does not occur for the other models presented thus far.

Restricting specific facet values to zero in a similar manner to the above can lead to formulations of Rasch models where the latent trait measure is given by a singular facet: any one of the facets presented in models described up to this point. A notable case of this is the

scenario in which one facet in any of models A, B or C is restricted to equal zero, leaving only the facet specific to two components of the circumstance description. That is, the latent trait variable φ is modelled to equal $\beta_{s,c}$, $\beta_{s,q}$ or $-\delta_{q,c}$ only. Take for example:

$$\varphi_{s,q,c} = \beta_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right]$$

This may be reformulated by multiplying the term within the square parentheses with a convenient ratio equal to one (shown below within large curved parentheses):

$$\varphi_{s,q,c} = \beta_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})} \times \left(\frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}\right)\right]$$

Rearranging the fractions within, this becomes:

$$\varphi_{s,q,c} = \beta_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(s,\phi_q,\phi_c)} \times \frac{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}{\pi_{g_0,k_0}(\vec{\Phi})}\right]$$

which may be written as a new model:

$$\varphi_{s,q,c} = \beta_{s,q} = \beta_s + \beta^*_{s,q} \quad ; \quad \beta^*_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(s,\phi_q,\phi_c)}\right]$$

Where β_s is defined as previously, and the newly introduced facet label $\beta^*_{s,q}$ is another question-specific student term, with a value defined relative to a student-specific reference point $\pi_{g_0,k_0}(s,\phi_q,\phi_c)$ rather than being defined relative to the universal reference point $\pi_{g_0,k_0}(\vec{\Phi})$ used for all other facet definitions thus far. Notice, however, that no assumptions needed to be made to obtain this formulation from the usual question-specific student facet: this expression is an equivalent, alternate form. It appears that the usual question-specific student facet, as defined relative to the universal reference point, is equivalent to modelling a facet which is student specific only, then adding a facet term which expresses a question specific component specific to that student. It can similarly be shown that the $\beta_{s,q}$ term may alternately be deconstructed as follows:

$$\varphi_{s,q,c} = \beta_{s,q} = \beta^{**}_{s,q} - \delta_q \quad ; \quad \beta^{**}_{s,q} = ln\left[\frac{\pi_{g_0,k_0}(s,q,\phi_c)}{\pi_{g_0,k_0}(\phi_s,q,\phi_c)}\right] \qquad \textbf{J}$$

which is an expression containing the usual question specific facet, plus yet another question-specific student facet, this time defined relative to a question specific reference point $\pi_{g_0,k_0}(\phi_q,q,\phi_c)$. Again, however, this model was obtained as a simple rearrangement of the single $\beta_{s,q}$ facet, and therefore if data were fit to this model it would yield no different information whatsoever. Model J is in this way *redundant*: the same information is achieved by utilising a simpler model, where the jointly specific facet (specific jointly to student and question) has not been split. This redundancy occurs for any case where a facet jointly specific to two components of the circumstance description (eg. question-specific student term, context-specific student term or question-specific context term) is coupled with a facet singularly specific to one of those same components (student, question or context). The same information would be obtained without modelling the singularly specific facet; it is made

redundant by inclusion of the jointly specific facet. For example, a model of solely the $\beta_{s,c}$ facet (as previously defined) is equivalent to the following:

$$\varphi_{s,q,c} = \beta_{s,c} = \beta_{s,c}^{**} - \delta_c \quad ; \quad \beta_{s,q}^{**} = ln\left[\frac{\pi_{g_0,k_0}(s,\phi_q,c)}{\pi_{g_0,k_0}(\phi_s,\phi_q,c)}\right] \qquad \textbf{K}$$

Because of these considerations of redundancy, no further Rasch model formulations exist where circumstances are differentiated only based on three components: student, question and context. Models J and K are still included in the main discussion, however, to complete the array of student dependent and student independent facet combinations shown in Table 18 of section 4.1.2.2.

7.4.2 Data tables

Original experiment titles presented here denote the student cohort the experiment was presented to (C indicates Chemistry IA/B, F indicated Foundations of Chemistry IA/B and u denotes unknown or mixed cohort), the year in which it was presented (for example ’10 is the notation for 2010) and the title of the experiment.

Table S 56: Original and equated experiment numbers and surveys gathered

ORIGINAL VALUES: Experiments conducted either in different semesters, conducted by different cohorts of students (Chemistry IA/B vs Foundations of Chemistry IA/B) or with different designs entirely are assigned separate sets of measures		EQUATED VALUES: Identically structured experiments are assigned identical sets of measures, regardless of the semester or year in which they were presented and regardless of the student cohort		Surveys received		
#	Experiment title	#	Experiment title (Title is only listed on the first occasion its facet element number appears)	initial count	after extremes removed	Used for equating
1	u-'10-Expt.5	5	Quantitative techniques	52	52	0
2	C-'11-Ex.1 - Biological buffers	1	Biological Buffers	136	134	0
3	C-'11-Ex.2 - Thermochemistry	2	Thermochemistry	104	104	0
4	C-'11-Ex.3 - Vapour pressure	3	Vapour Pressure	84	83	0
5	C-'11-Ex.4 - Melting points and recrystallisation	4	Melting Points and Recrystallisation	147	146	0
6	C-'11-Ex.5 - Quantitative techniques	5		97	96	0
7	C-'11-Ex.6 - Reaction kinetics	6	Reaction Kinetics	154	154	0
8	C-'11-Ex.8 - Liquid-liquid extraction and TLC	8	Liquid-Liquid Extraction and TLC	120	118	0
9	C-'11-Ex.9 - Synthesis of aspirin	9	Synthesis of Aspirin	95	93	0
10	C-'11-Ex.10 - coloured complexes of iron	10	Coloured Complexes of Iron	90	90	0
11	C-'11-Ex.11 - analysis of spinach extracts	11	Analysis of Spinach Extracts	144	141	0
12	C-'11-Ex.12 - Ion exchange chromatography	12	Ion exchange Chromatography	108	106	0
13	C-'11-Ex.13 - Copper(II) ion absorption spectrophotometry	13	Copper(II) Ion Absorption Spectrophotometry	126	123	0
14	F-'11-Ex.1 - Biological buffers	1		69	67	0
15	F-'11-Ex.4 - Melting points and recrystallysation	4		85	82	0
16	F-'11-Ex.5 - Quantitative techniques	5		64	63	0
17	F-'11-Ex.6 - Reaction kinetics	6		93	92	0
18	F-'11-Ex.8 - Liquid-liquid extraction and TLC	8		57	54	0
19	F-'11-Ex.10 - Coloured complexes of iron	10		71	70	0
20	F-'11-Ex.12 - Ion exchange chromatography	12		78	78	0
21	F-'11-Ex.13 - Copper(II) ion absorption spectrophotometry	13		83	83	0
22	F-'12-Ex.0F - Introductory experiment	7	Introductory experiment	19	19	0
23	F-'12-Ex.1F - Quantitative techniques	14	Quantitative Techniques (revised: for foundations)	102	102	0
24	F-'12-Ex.2F - Vitamin C titration	15	Determination of vitamin C concentration	73	73	0

ORIGINAL VALUES: Experiments conducted either in different semesters, conducted by different cohorts of students (Chemistry IA/B vs Foundations of Chemistry IA/B) or with different designs entirely are assigned separate sets of measures		EQUATED VALUES: Identically structured experiments are assigned identical sets of measures, regardless of the semester or year in which they were presented and regardless of the student cohort		Surveys received		
#	Experiment title	#	Experiment title (Title is only listed on the first occasion its facet element number appears)	initial count	after extremes removed	Used for equating
25	F-'12-Ex.3F - Equilibrium & Le Chatalier's principle	16	Equilibrium and Le Chatalier's Principle (revised: for foundations)	104	102	0
26	F-'12-Ex.4F - Visible absorption spectrophotometry	17	Visible Absorption Spectrophotometry	73	70	0
27	F-'12-Ex.5F - Aromas	18	Aromachemistry	103	101	0
28	F-'12-Ex.6F - Analysis of spinach extracts	19	Analysis of Spinach Extracts (revised: for foundation - in pairs)	107	103	0
29	F-'12-Ex.7F - Activity Series	20	Metal Activity Series	81	76	0
30	F-'12-Ex.8F - Thermochemistry	21	Thermochemistry (revised: for foundations)	77	77	0
31	F-'12-Ex.9F - Reaction kinetics	22	Reaction Kinetics (revised: for foundations)	74	74	0
32	F-'13-Ex 0F - Introductory experiment	23	Introductory experiment (revised: observations video)	126	126	9
33	F-'13-Ex 1F - Quantitative techniques	14		61	60	4
34	F-'13-Ex 2F - Vitamin C titration	15		57	56	7
35	F-'13-Ex 3F - Visible absorption spectrophotometry	17		137	136	10
36	F-'13-Ex 4F - Equilibrium & Le Chatalier's principle	16		83	82	6
37	F-'13-Ex 5F - Aromas	18		248	248	26
38	F-'13-Ex 6F - Analysis of spinach extracts	19		206	205	24
39	F-'13-Ex 7F - Thermochemistry	21		204	202	23
40	F-'13-Ex 8F - Activity series	20		161	156	23
41	F-'13-Ex 9F - Reaction kinetics	24	Reaction Kinetics (revised: question order and phrasing)	82	81	13
42	C-'12-Ex.1 - Biological buffers	25	Biological Buffers (revised: laptop)	80	80	0
43	C-'12-Ex.4 - Melting points and recrystallisation	4		70	70	0
44	C-'12-Ex.6 - Reaction kinetics	6		84	84	0
45	C-'12-Ex.8 - Liquid-liquid extraction and TLC	8		72	72	0
46	C-'12-Ex.9 - Synthesis of aspirin	9		36	36	0
47	C-'12-Ex.11 - Analysis of spinach extracts	11		77	77	0
48	C-'13-Ex 2 - Thermochemistry	2		227	227	23
49	C-'13-Ex 3 - Vapour pressure	26	Vapour Pressure (revised: laptop)	148	148	20
50	C-'13-Ex 5 - Quantitative techniques	5		203	201	30
51	C-'13-Ex 10 - Equilibrium & Le Chatelier's principle	10		174	172	23
52	C-'13-Ex 12 - Ion exchange chromatography	12		252	252	36

ORIGINAL VALUES: Experiments conducted either in different semesters, conducted by different cohorts of students (Chemistry IA/B vs Foundations of Chemistry IA/B) or with different designs entirely are assigned separate sets of measures		EQUATED VALUES: Identically structured experiments are assigned identical sets of measures, regardless of the semester or year in which they were presented and regardless of the student cohort		Surveys received		
#	Experiment title	#	Experiment title (Title is only listed on the first occasion its facet element number appears)	initial count	after extremes removed	Used for equating
53	C-'13-Ex 13 - Copper(II) ion absorption spectrophotometry	27	Copper(II) Ion Absorption Spectrophotometry (revised: laptop)	232	231	26
54	C-'13-Ex 11 - Analysis of spinach extracts	11		218	217	30
55	C-'13-Ex 9 - Synthesis of aspirin	9		129	129	18
56	C-'13-Ex 6 - Reaction kinetics	6		205	202	33
57	C-'13-Ex 4 - Melting points and recrystallisation	4		182	178	27
58	C-'13-Ex 1 - Biological buffers	25		170	166	21
59	C-'13-Ex 8 - Liquid-liqud extraction and TLC	8		128	128	19
60	C-'12-Ex 2 - Thermochemistry	2		140	138	0
61	C-'12-Ex 3 - Vapour pressure	26		102	102	0
62	C-'12-Ex 5 - Quantitative techniques	5		112	111	0
63	C-'12-Ex 10 - Equilibrium & Le Chatelier's principle	10		83	82	0
64	C-'12-Ex 12 - Ion exchange chromatography	12		128	127	0
65	C-'12-Ex 13 - Copper(II) ion absorption spectrophotometry	27		120	119	0
66	F-'14- Ex 0F - Introductory Experiment	28	Introductory Experiment (revised: pipetting)	143	143	7
67	F-'14- Ex 1F - Quantitative Techniques	29	Quantitative Techniques (revised: no pipetting)	177	177	10
68	F-'14- Ex 2F - Determination of Vitamin C Concentration	15		137	133	8
69	F-'14- Ex 3F - Equilibrium & Le Chatelier's Principle	16		109	107	6
70	F-'14- Ex 4F - Visible Absorption Spectrophotometry	17		177	176	7
71	C-'14-Thermochemistry	2		237	235	6
72	C-'14-Vapour Pressure	26		144	144	6
73	C-'14-Ion Exchange Chromatography	12		199	196	8
74	C-'14-Visible Absorption Spectrophotometry	27		189	189	7
75	C-'14-Quantitative Techniques	5		140	140	6
76	C-'14-Equilibrium & Le Chatelier's principle	10		121	120	8
			Totals:	9380	9287	530

Table S 57: Tests of normality for student measures gathered from different occasions (subset 1 only)

Unequated experiment number and title (sample group)		Kolmogorov-Smirnov test			Shapiro-Wilk test		
		Statistic	df	p	Statistic	df	p
1	u-'10-Expt.5	.081	52	.200*	.980	52	.519
3	C-'11-Ex.2 - Thermochemistry	.088	104	.045	.980	104	.117
5	C-'11-Ex.4 - Melting points and recrystallisation	.125	146	.000	.951	146	.000
6	C-'11-Ex.5 - Quantitative techniques	.124	96	.001	.956	96	.003
7	C-'11-Ex.6 - Reaction kinetics	.149	154	.000	.953	154	.000
8	C-'11-Ex.8 - Liquid-liquid extraction and TLC	.098	118	.008	.981	118	.098
9	C-'11-Ex.9 - Synthesis of aspirin	.159	93	.000	.952	93	.002
10	C-'11-Ex.10 - coloured complexes of iron	.180	90	.000	.914	90	.000
11	C-'11-Ex.11 - analysis of spinach extracts	.094	141	.004	.970	141	.004
12	C-'11-Ex.12 - Ion exchange chromatography	.121	106	.001	.984	106	.225
15	F-'11-Ex.4 - Melting points and recrystallysation	.137	82	.001	.941	82	.001
16	F-'11-Ex.5 - Quantitative techniques	.134	63	.007	.924	63	.001
17	F-'11-Ex.6 - Reaction kinetics	.122	92	.002	.956	92	.003
18	F-'11-Ex.8 - Liquid-liquid extraction and TLC	.133	54	.018	.971	54	.204
19	F-'11-Ex.10 - Coloured complexes of iron	.104	70	.059	.970	70	.087
20	F-'11-Ex.12 - Ion exchange chromatography	.125	78	.004	.944	78	.002
42	C-'12-Ex.1 - Biological buffers	.103	80	.036	.978	80	.175
43	C-'12-Ex.4 - Melting points and recrystallisation	.098	70	.092	.961	70	.027
44	C-'12-Ex.6 - Reaction kinetics	.141	84	.000	.959	84	.010
45	C-'12-Ex.8 - Liquid-liquid extraction and TLC	.106	72	.043	.953	72	.009
46	C-'12-Ex.9 - Synthesis of aspirin	.192	36	.002	.948	36	.093
47	C-'12-Ex.11 - Analysis of spinach extracts	.108	77	.028	.969	77	.058
48	C-'13-Ex 2 - Thermochemistry	.125	204	.000	.956	204	.000
49	C-'13-Ex 3 - Vapour pressure	.082	128	.036	.986	128	.210
50	C-'13-Ex 5 - Quantitative techniques	.087	171	.003	.959	171	.000
51	C-'13-Ex 10 - Equilibrium & Le Chatelier's principle	.091	149	.004	.976	149	.011
52	C-'13-Ex 12 - Ion exchange chromatography	.097	216	.000	.978	216	.002
53	C-'13-Ex 13 - Copper(II) ion absorption spectrophotometry	.143	205	.000	.944	205	.000
54	C-'13-Ex 11 - Analysis of spinach extracts	.096	187	.000	.977	187	.004
55	C-'13-Ex 9 - Synthesis of aspirin	.133	111	.000	.947	111	.000
56	C-'13-Ex 6 - Reaction kinetics	.080	169	.011	.976	169	.005
57	C-'13-Ex 4 - Melting points and recrystallisation	.126	151	.000	.952	151	.000
58	C-'13-Ex 1 - Biological buffers	.078	145	.031	.987	145	.194
59	C-'13-Ex 8 - Liquid-liqud extraction and TLC	.088	109	.038	.975	109	.036
60	C-'12-Ex 2 - Thermochemistry	.116	138	.000	.974	138	.011
61	C-'12-Ex 3 - Vapour pressure	.117	102	.002	.953	102	.001
62	C-'12-Ex 5 - Quantitative techniques	.099	111	.009	.961	111	.002
63	C-'12-Ex 10 - Equilibrium & Le Chatelier's principle	.169	82	.000	.918	82	.000
64	C-'12-Ex 12 - Ion exchange chromatography	.110	127	.001	.936	127	.000
65	C-'12-Ex 13 - Copper(II) ion absorption spectrophotometry	.137	119	.000	.946	119	.000
71	C-'14-Thermochemistry	.071	229	.007	.981	229	.004
72	C-'14-Vapour Pressure	.118	138	.000	.971	138	.005
73	C-'14-Ion Exchange Chromatography	.106	188	.000	.968	188	.000
74	C-'14-Visible Absorption Spectrophotometry	.130	182	.000	.941	182	.000
75	C-'14-Quantitative Techniques	.091	134	.008	.969	134	.004
76	C-'14-Equilibrium & Le Chatelier's principle	.079	112	.086	.986	112	.281

* This is a lower bound of the true significance.

Table S 58: Experiment quality measures (δ) estimated using the final equated model

#	Experiment title	Survey item (question) number and topic 1 data interpretation	2 laboratory skills	3 interest	4 clear assessment	5 expected learning	6 increased understanding	7 background in introduction	8 demonstrators	9 procedure in manual	10 relevance to chemistry studies	11 teamwork	12 responsibility for own learning	13 time availability	14 overall learning experience	measurement subset
2	Thermochemistry	0.24	0.77	1.10	-0.14	-0.52	-0.26	0.11	-0.92	-0.54	-0.87	-0.81	-0.25	0.01	-0.03	1
4	Melting Points and Recrystallisation	0.11	-0.70	-0.22	-0.37	-0.31	-0.07	-0.15	-0.69	-0.33	0.37	1.18	-0.36	1.54	-0.35	1
5	Quantitative techniques	-0.27	-1.15	0.34	-0.32	-0.45	0.51	-0.34	-0.96	-0.73	-0.19	1.34	-0.89	2.31	-0.25	1
6	Reaction Kinetics	0.14	0.20	0.56	-0.01	-0.16	0.14	0.15	-0.54	0.22	-0.35	-0.73	-0.13	1.68	0.31	1
8	Liquid-Liquid Extraction and TLC	0.49	-0.74	0.21	0.00	0.13	0.06	0.15	-0.69	-0.11	0.26	0.96	-0.57	2.05	0.30	1
9	Synthesis of Aspirin	0.31	-0.76	-0.38	-0.43	-0.01	0.14	-0.54	-0.80	-0.65	0.16	1.20	-0.49	3.29	-0.21	1
10	Coloured Complexes of Iron	0.56	-0.62	0.24	0.13	0.40	0.23	0.07	-0.68	-0.27	0.11	-1.02	-0.14	3.07	0.20	1
11	Analysis of Spinach Extracts	0.21	-0.85	-0.41	-0.27	-0.07	0.08	-0.34	-0.78	0.36	0.21	1.38	-0.51	0.36	-0.45	1
12	Ion exchange Chromatography	0.32	-0.63	-0.05	0.12	0.02	-0.54	0.37	-0.82	0.03	0.06	-1.20	-0.54	1.33	-0.35	1
25	Biological Buffers (revised: laptop)	-0.42	0.05	0.95	0.43	0.26	0.23	0.51	-0.79	0.91	-0.35	-0.90	0.29	1.52	0.34	1
26	Vapour Pressure (revised: laptop)	-0.47	-0.58	0.61	0.41	0.07	-0.28	-0.03	-1.07	0.56	0.08	-1.03	-0.19	2.22	0.24	1
27	Copper(II) Ion Absorption Spectrophotometry (revised: laptop)	-0.27	-0.35	0.06	-0.21	-0.34	0.18	-0.42	-0.98	-0.45	-0.11	-0.74	-0.16	1.13	-0.40	1
1	Biological Buffers	0.15	0.25	0.97	0.15	0.22	0.29	0.25	-0.63	-0.08	-0.51	-0.86	-0.40	1.82	0.71	2
3	Vapour Pressure	0.64	0.38	1.42	0.72	0.69	0.63	0.06	-0.92	1.06	0.01	-1.50	0.09	3.41	1.44	3
13	Copper(II) Ion Absorption Spectrophotometry	-0.07	-0.34	0.55	0.01	0.00	0.29	-0.03	-0.78	-0.33	0.30	-1.07	-0.41	1.16	0.08	4
7	Introductory experiment	0.44	2.34	1.40	-0.50	-0.12	0.62	-0.18	-1.03	-0.23	-0.68	1.08	-0.09	-2.40	0.53	5
14	Quantitative Techniques (revised: for foundations)	-0.19	-1.38	0.47	0.34	-0.26	0.26	0.00	-0.49	-0.01	-0.40	0.68	-0.70	2.97	-0.03	6
15	Determination of Vitamin C concentration	-0.31	-0.99	0.00	-0.12	-0.05	0.12	-0.01	-0.53	-0.33	-0.25	0.70	-0.35	3.19	0.00	6
16	Equilibrium and Le Chatalier's Principle (revised: for foundations)	0.22	-0.92	-0.17	0.13	0.33	0.18	0.16	-0.53	-0.05	-0.13	-0.72	-0.26	2.92	0.26	6
17	Visible Absorption Spectrophotometry	-0.26	-0.54	0.14	-0.22	-0.16	0.01	-0.25	-0.66	-0.20	0.22	-0.63	-0.02	0.23	-0.41	6
18	Aromachemistry	0.59	1.53	0.08	-0.44	-0.40	-0.10	0.20	-0.71	-0.68	-0.46	-0.58	-0.07	-0.42	-0.09	6
19	Analysis of Spinach Extracts (revised: for foundation - in pairs)	-0.25	-1.05	-0.47	-0.15	-0.14	-0.28	0.00	-1.11	0.41	-0.02	-0.78	-0.33	1.26	-0.67	6
20	Metal Activity Series	-0.10	-0.29	0.07	-0.37	-0.49	-0.48	-0.39	-0.75	-0.49	-0.63	-0.71	-0.18	1.65	-0.01	6
21	Thermochemistry (revised: for foundations)	-0.17	0.60	0.73	-0.22	-0.18	-0.26	0.05	-0.90	-0.35	-0.41	-0.75	-0.44	1.19	0.21	6
23	Introductory experiment (revised: observations video)	0.36	2.04	1.92	-0.02	-0.63	0.08	0.40	-0.80	0.32	-0.68	1.06	-0.29	-0.58	0.02	6
24	Reaction Kinetics (revised: question order and phrasing)	-0.84	-0.44	-0.08	-0.75	-0.71	-0.79	-0.20	-0.92	-0.07	-0.97	-1.28	-0.02	1.62	0.01	6
28	Introductory Experiment (revised: pipetting)	-0.35	-0.76	1.21	-0.34	-0.41	0.89	-0.29	-1.35	-0.43	-0.19	1.09	-0.33	-0.92	-0.38	6
29	Quantitative Techniques (revised: no pipetting)	-0.32	-1.25	-0.14	-0.01	0.03	-0.12	0.03	-0.77	-0.01	-0.12	0.96	-0.38	1.40	-0.62	6
22	Reaction Kinetics (revised: for foundations)	-0.26	-0.18	0.37	0.33	-0.08	-0.02	0.79	-0.73	0.28	0.04	-0.88	-0.42	2.55	0.41	7

More positive measure values imply greater difficulty of providing positive response. Measures are not comparable between different subsets.

7.5 Supporting information for section 4.2

Experiment numbers referenced in the table following refer to those of the equated best explanatory model of the ASLE data. The δ values obtained without taking gender into account has previously been reported for each of these experiments, for each question (survey item), as detailed in Table S 56.

Table S 59: DIF between genders for the equated model

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
2	1	0.01	0.12	0.36	0.11	-0.36	0.17	-2.13	419	0.0341
2	2	0.54	0.11	0.83	0.10	-0.28	0.15	-1.89	422	0.0593
2	3	1.14	0.09	1.04	0.09	0.10	0.13	0.75	422	0.4525
2	4	-0.18	0.12	-0.05	0.11	-0.13	0.16	-0.82	422	0.4140
2	5	-0.56	0.12	-0.47	0.11	-0.09	0.17	-0.52	421	0.6052
2	6	-0.24	0.12	-0.37	0.11	0.13	0.16	0.78	423	0.4346
2	7	-0.04	0.11	0.23	0.10	-0.27	0.15	-1.84	420	0.0668
2	8	-0.96	0.13	-0.98	0.12	0.02	0.18	0.08	421	0.9338
2	9	-0.62	0.12	-0.59	0.11	-0.04	0.16	-0.22	422	0.8248
2	10	-0.65	0.12	-0.89	0.12	0.24	0.17	1.39	423	0.1662
2	11	-0.65	0.12	-0.82	0.11	0.17	0.16	1.02	420	0.3065
2	12	0.08	0.12	-0.23	0.12	0.31	0.16	1.90	420	0.0586
2	13	0.16	0.14	0.03	0.14	0.12	0.20	0.61	422	0.5406
2	14	-0.24	0.13	-0.19	0.12	-0.05	0.18	-0.29	420	0.7749
4	1	0.32	0.17	-0.03	0.18	0.35	0.25	1.39	196	0.1654
4	2	-0.81	0.19	-0.75	0.19	-0.07	0.26	-0.25	198	0.7995
4	3	-0.56	0.17	-0.30	0.16	-0.26	0.24	-1.10	198	0.2735
4	4	-0.04	0.16	-0.50	0.17	0.45	0.24	1.89	199	0.0599
4	5	-0.22	0.17	-0.33	0.17	0.11	0.24	0.47	199	0.6424
4	6	-0.37	0.17	-0.14	0.16	-0.23	0.24	-0.97	199	0.3330
4	7	-0.24	0.16	-0.13	0.16	-0.12	0.23	-0.51	198	0.6113
4	8	-0.93	0.19	-0.97	0.19	0.04	0.27	0.15	197	0.8774
4	9	-0.07	0.16	-0.37	0.16	0.30	0.23	1.31	197	0.1904
4	10	-0.05	0.16	0.47	0.15	-0.53	0.22	-2.36	198	0.0192
4	11	1.50	0.14	1.27	0.15	0.24	0.20	1.15	148	0.2504
4	12	-0.02	0.17	0.25	0.17	-0.27	0.24	-1.12	197	0.2642
4	13	1.78	0.24	1.26	0.26	0.52	0.35	1.47	197	0.1420
4	14	-0.54	0.20	-0.47	0.19	-0.07	0.27	-0.25	197	0.8021
5	1	-0.41	0.14	-0.34	0.15	-0.07	0.21	-0.36	313	0.7207
5	2	-1.24	0.15	-1.31	0.16	0.07	0.22	0.33	313	0.7452
5	3	0.54	0.11	0.19	0.13	0.35	0.17	2.08	311	0.0381
5	4	-0.24	0.13	-0.33	0.14	0.09	0.19	0.48	312	0.6337
5	5	-0.49	0.13	-0.40	0.14	-0.09	0.19	-0.48	313	0.6308
5	6	0.66	0.11	0.42	0.13	0.24	0.17	1.41	312	0.1599
5	7	-0.45	0.13	-0.20	0.13	-0.24	0.18	-1.32	313	0.1873
5	8	-1.02	0.15	-1.02	0.16	0.01	0.22	0.03	312	0.9751
5	9	-0.71	0.13	-0.73	0.14	0.02	0.20	0.08	311	0.9340
5	10	-0.53	0.13	-0.21	0.14	-0.32	0.19	-1.66	312	0.0970
5	11	1.60	0.10	1.52	0.11	0.08	0.15	0.54	249	0.5918
5	12	-0.66	0.14	-0.70	0.15	0.05	0.20	0.23	309	0.8144
5	13	1.99	0.20	2.35	0.21	-0.36	0.29	-1.25	310	0.2108
5	14	-0.50	0.15	-0.57	0.16	0.07	0.22	0.33	311	0.7410
6	1	0.36	0.16	0.50	0.15	-0.14	0.21	-0.68	247	0.5003
6	2	0.28	0.15	0.51	0.14	-0.22	0.21	-1.09	247	0.2777

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
6	3	0.83	0.13	0.31	0.13	0.52	0.19	2.78	245	0.0059
6	4	-0.06	0.15	-0.05	0.14	-0.02	0.21	-0.09	246	0.9319
6	5	-0.25	0.15	-0.31	0.15	0.06	0.22	0.27	246	0.7850
6	6	0.12	0.15	0.20	0.14	-0.08	0.20	-0.38	245	0.7065
6	7	-0.17	0.15	0.08	0.14	-0.24	0.20	-1.22	246	0.2253
6	8	-1.15	0.19	-0.45	0.15	-0.70	0.24	-2.95	242	0.0035
6	9	-0.02	0.14	0.31	0.13	-0.33	0.19	-1.74	246	0.0829
6	10	-0.57	0.16	-0.47	0.15	-0.09	0.22	-0.42	246	0.6713
6	11	-0.59	0.16	-0.63	0.15	0.05	0.22	0.22	241	0.8282
6	12	0.19	0.15	0.38	0.15	-0.19	0.21	-0.90	246	0.3679
6	13	1.93	0.22	1.40	0.23	0.53	0.32	1.66	244	0.0991
6	14	0.19	0.16	0.05	0.16	0.13	0.23	0.58	245	0.5603
8	1	0.35	0.20	0.77	0.19	-0.42	0.28	-1.53	144	0.1294
8	2	-0.69	0.21	-0.98	0.22	0.29	0.31	0.92	145	0.3569
8	3	0.28	0.18	0.06	0.18	0.22	0.26	0.87	145	0.3882
8	4	0.00	0.19	0.00	0.19	0.00	0.27	-0.01	144	0.9921
8	5	0.13	0.19	0.42	0.18	-0.29	0.26	-1.10	145	0.2713
8	6	-0.21	0.20	0.24	0.18	-0.45	0.27	-1.67	145	0.0972
8	7	0.19	0.18	0.26	0.18	-0.07	0.25	-0.27	144	0.7904
8	8	-0.84	0.22	-0.92	0.22	0.08	0.31	0.25	145	0.8017
8	9	-0.11	0.18	-0.31	0.19	0.19	0.27	0.73	145	0.4674
8	10	0.26	0.18	0.23	0.19	0.04	0.26	0.14	144	0.8925
8	11	1.13	0.17	1.28	0.17	-0.14	0.24	-0.59	111	0.5560
8	12	-0.25	0.20	-0.45	0.21	0.19	0.29	0.67	145	0.5037
8	13	2.08	0.28	1.75	0.29	0.33	0.40	0.81	145	0.4171
8	14	0.30	0.21	-0.32	0.22	0.63	0.30	2.07	145	0.0404
9	1	0.43	0.22	0.33	0.19	0.11	0.29	0.37	130	0.7107
9	2	-0.69	0.24	-0.98	0.21	0.29	0.32	0.90	130	0.3721
9	3	-0.43	0.22	-0.71	0.19	0.28	0.29	0.97	130	0.3351
9	4	-0.61	0.23	-0.35	0.19	-0.25	0.30	-0.85	128	0.3957
9	5	-0.39	0.23	0.16	0.17	-0.55	0.29	-1.95	127	0.0539
9	6	0.15	0.21	0.15	0.17	0.00	0.27	-0.01	129	0.9905
9	7	-0.83	0.23	-0.57	0.19	-0.25	0.30	-0.85	128	0.3993
9	8	-1.01	0.25	-0.87	0.20	-0.15	0.32	-0.45	127	0.6500
9	9	-0.28	0.21	-0.62	0.19	0.34	0.28	1.21	130	0.2301
9	10	0.13	0.21	0.08	0.18	0.05	0.27	0.19	127	0.8480
9	11	1.27	0.18	1.33	0.16	-0.05	0.24	-0.23	108	0.8196
9	12	-0.34	0.23	0.05	0.19	-0.38	0.30	-1.29	128	0.2002
9	13	3.67	0.25	3.27	0.22	0.40	0.33	1.21	124	0.2281
9	14	-0.44	0.26	-0.45	0.21	0.02	0.33	0.06	124	0.9561
10	1	0.55	0.14	0.55	0.14	0.00	0.20	0.01	270	0.9960
10	2	-0.54	0.15	-0.90	0.16	0.36	0.22	1.62	269	0.1056
10	3	0.15	0.13	0.09	0.13	0.06	0.19	0.31	270	0.7574
10	4	0.32	0.13	0.00	0.14	0.32	0.19	1.65	267	0.1007
10	5	0.40	0.13	0.37	0.13	0.03	0.19	0.16	269	0.8709
10	6	0.30	0.13	0.21	0.14	0.09	0.19	0.46	269	0.6464
10	7	0.25	0.13	0.04	0.13	0.21	0.19	1.12	268	0.2617
10	8	-0.92	0.16	-0.54	0.15	-0.38	0.22	-1.74	268	0.0827
10	9	-0.24	0.14	-0.09	0.13	-0.15	0.19	-0.79	269	0.4276
10	10	-0.12	0.14	-0.08	0.14	-0.04	0.20	-0.20	269	0.8436
10	11	-1.03	0.16	-0.86	0.15	-0.17	0.22	-0.78	268	0.4334
10	12	0.07	0.14	0.28	0.14	-0.21	0.20	-1.04	269	0.2973
10	13	3.58	0.17	2.80	0.19	0.77	0.25	3.06	265	0.0025
10	14	0.00	0.15	0.01	0.16	-0.02	0.22	-0.08	268	0.9340

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
11	1	0.30	0.16	0.51	0.16	-0.21	0.22	-0.96	223	0.3377
11	2	-0.67	0.17	-0.87	0.18	0.19	0.24	0.78	223	0.4361
11	3	-0.23	0.15	-0.67	0.16	0.43	0.22	1.97	223	0.0495
11	4	-0.23	0.15	-0.46	0.17	0.23	0.22	1.02	222	0.3101
11	5	-0.12	0.15	-0.18	0.16	0.07	0.22	0.31	224	0.7571
11	6	0.10	0.15	-0.01	0.15	0.12	0.21	0.54	223	0.5889
11	7	-0.43	0.15	-0.41	0.16	-0.01	0.22	-0.05	222	0.9603
11	8	-0.79	0.16	-0.94	0.18	0.15	0.24	0.62	222	0.5358
11	9	0.01	0.14	0.24	0.14	-0.23	0.20	-1.14	222	0.2545
11	10	0.14	0.14	0.25	0.15	-0.11	0.21	-0.54	223	0.5864
11	11	1.32	0.13	1.40	0.14	-0.08	0.19	-0.41	172	0.6807
11	12	-0.11	0.16	-0.05	0.16	-0.06	0.22	-0.27	222	0.7862
11	13	0.29	0.20	0.72	0.22	-0.42	0.29	-1.44	222	0.1510
11	14	-0.55	0.17	-0.52	0.18	-0.03	0.25	-0.10	222	0.9165
12	1	0.50	0.11	0.30	0.11	0.20	0.16	1.27	411	0.2053
12	2	-0.76	0.13	-0.35	0.11	-0.41	0.17	-2.37	410	0.0183
12	3	0.12	0.11	-0.15	0.10	0.28	0.15	1.86	413	0.0634
12	4	0.04	0.11	0.26	0.10	-0.22	0.15	-1.48	412	0.1391
12	5	-0.12	0.11	0.02	0.10	-0.14	0.15	-0.90	410	0.3687
12	6	-0.63	0.12	-0.57	0.11	-0.06	0.17	-0.36	411	0.7164
12	7	0.29	0.10	0.36	0.09	-0.07	0.14	-0.49	410	0.6223
12	8	-0.89	0.12	-0.59	0.11	-0.30	0.16	-1.84	410	0.0659
12	9	-0.08	0.11	0.13	0.09	-0.21	0.14	-1.47	410	0.1421
12	10	0.03	0.11	-0.35	0.11	0.38	0.16	2.43	412	0.0154
12	11	-0.92	0.12	-1.30	0.12	0.39	0.17	2.30	412	0.0218
12	12	-0.37	0.12	-0.63	0.12	0.27	0.17	1.61	409	0.1086
12	13	1.36	0.18	1.33	0.17	0.03	0.25	0.10	412	0.9178
12	14	-0.35	0.13	-0.49	0.13	0.14	0.18	0.77	412	0.4430
14	1	-0.28	0.38	-0.26	0.37	-0.02	0.53	-0.03	44	0.9736
14	2	-1.06	0.38	-1.53	0.41	0.47	0.56	0.84	44	0.4055
14	3	0.49	0.30	0.87	0.29	-0.38	0.42	-0.91	44	0.3667
14	4	-0.27	0.34	0.24	0.33	-0.51	0.47	-1.07	43	0.2884
14	5	-0.40	0.35	-0.93	0.38	0.52	0.51	1.02	44	0.3121
14	6	-0.03	0.33	0.39	0.31	-0.42	0.45	-0.94	44	0.3547
14	7	-0.30	0.33	0.23	0.31	-0.53	0.45	-1.17	43	0.2490
14	8	-0.31	0.32	-0.38	0.33	0.07	0.46	0.15	44	0.8792
14	9	0.17	0.30	0.07	0.31	0.10	0.43	0.23	44	0.8215
14	10	-0.64	0.35	-0.64	0.36	0.00	0.50	0.00	44	0.9981
14	11	1.36	0.27	0.96	0.29	0.40	0.40	1.01	40	0.3182
14	12	-0.50	0.36	-0.99	0.38	0.49	0.52	0.95	44	0.3464
14	13	3.21	0.41	2.46	0.48	0.74	0.64	1.17	44	0.2486
14	14	-0.25	0.39	0.04	0.37	-0.29	0.53	-0.55	44	0.5867
15	1	-0.64	0.22	-0.26	0.21	-0.38	0.30	-1.24	147	0.2162
15	2	-1.12	0.22	-1.02	0.21	-0.11	0.31	-0.35	147	0.7274
15	3	-0.23	0.19	0.02	0.18	-0.25	0.26	-0.97	147	0.3357
15	4	-0.05	0.19	0.02	0.19	-0.07	0.27	-0.26	147	0.7927
15	5	0.15	0.18	0.00	0.19	0.15	0.26	0.58	147	0.5660
15	6	0.03	0.19	-0.01	0.19	0.04	0.26	0.16	147	0.8755
15	7	-0.34	0.19	-0.03	0.18	-0.31	0.26	-1.17	147	0.2440
15	8	-0.53	0.20	-0.35	0.19	-0.18	0.27	-0.66	146	0.5082
15	9	-0.18	0.18	-0.19	0.18	0.01	0.26	0.03	147	0.9767
15	10	-0.22	0.19	-0.25	0.19	0.03	0.27	0.10	147	0.9237
15	11	0.87	0.17	0.75	0.17	0.12	0.24	0.51	129	0.6135
15	12	-0.12	0.20	-0.21	0.20	0.09	0.28	0.32	143	0.7519

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
15	13	3.21	0.24	3.14	0.24	0.07	0.34	0.22	146	0.8282
15	14	0.09	0.21	-0.33	0.21	0.43	0.30	1.42	147	0.1564
16	1	0.04	0.21	0.13	0.19	-0.09	0.28	-0.31	154	0.7550
16	2	-0.67	0.22	-1.11	0.21	0.44	0.30	1.46	155	0.1460
16	3	-0.51	0.20	-0.13	0.18	-0.37	0.27	-1.40	152	0.1626
16	4	-0.19	0.20	0.26	0.17	-0.45	0.26	-1.70	152	0.0906
16	5	-0.12	0.20	0.49	0.16	-0.61	0.26	-2.39	153	0.0180
16	6	-0.02	0.20	0.15	0.17	-0.17	0.26	-0.65	153	0.5165
16	7	0.27	0.18	0.14	0.17	0.13	0.25	0.52	154	0.6060
16	8	-0.10	0.18	-0.15	0.17	0.06	0.25	0.24	155	0.8109
16	9	-0.14	0.19	-0.11	0.17	-0.02	0.25	-0.09	154	0.9311
16	10	-0.20	0.20	-0.19	0.18	-0.01	0.27	-0.02	155	0.9849
16	11	-0.08	0.18	-0.72	0.19	0.64	0.26	2.44	154	0.0156
16	12	-0.05	0.20	-0.20	0.19	0.15	0.28	0.54	153	0.5899
16	13	3.31	0.23	2.65	0.24	0.66	0.33	1.99	154	0.0482
16	14	0.17	0.21	0.01	0.20	0.16	0.29	0.55	154	0.5835
17	1	-0.47	0.17	-0.22	0.15	-0.26	0.22	-1.15	264	0.2518
17	2	-0.70	0.16	-0.49	0.15	-0.21	0.22	-0.95	266	0.3416
17	3	-0.10	0.14	0.20	0.12	-0.30	0.19	-1.61	264	0.1082
17	4	-0.15	0.15	-0.29	0.14	0.14	0.20	0.69	268	0.4894
17	5	-0.08	0.14	-0.12	0.13	0.04	0.20	0.19	267	0.8516
17	6	-0.06	0.14	0.03	0.13	-0.09	0.19	-0.44	267	0.6596
17	7	-0.14	0.14	-0.41	0.13	0.27	0.19	1.40	267	0.1614
17	8	-0.81	0.16	-0.45	0.13	-0.36	0.21	-1.75	264	0.0816
17	9	0.02	0.13	-0.19	0.13	0.21	0.19	1.11	268	0.2675
17	10	0.14	0.14	0.20	0.13	-0.05	0.19	-0.29	265	0.7753
17	11	-0.37	0.14	-0.80	0.14	0.43	0.19	2.22	269	0.0271
17	12	0.03	0.15	0.00	0.14	0.03	0.20	0.13	267	0.8969
17	13	0.30	0.20	0.34	0.18	-0.04	0.27	-0.15	266	0.8784
17	14	-0.36	0.17	-0.60	0.16	0.23	0.23	1.01	267	0.3152
18	1	0.42	0.15	0.84	0.15	-0.42	0.21	-1.95	230	0.0525
18	2	1.33	0.12	1.72	0.13	-0.39	0.18	-2.21	232	0.0279
18	3	0.16	0.13	-0.09	0.15	0.25	0.20	1.24	228	0.2174
18	4	-0.29	0.15	-0.55	0.17	0.26	0.22	1.16	229	0.2484
18	5	-0.19	0.15	-0.72	0.17	0.53	0.22	2.35	229	0.0198
18	6	-0.28	0.15	-0.06	0.16	-0.22	0.21	-1.03	232	0.3021
18	7	0.07	0.13	0.24	0.14	-0.17	0.20	-0.85	230	0.3955
18	8	-0.56	0.15	-0.83	0.18	0.28	0.23	1.21	229	0.2288
18	9	-0.42	0.14	-0.95	0.17	0.53	0.23	2.36	229	0.0189
18	10	-0.39	0.15	-0.71	0.17	0.32	0.22	1.41	231	0.1596
18	11	-0.64	0.14	-0.35	0.15	-0.29	0.21	-1.38	232	0.1683
18	12	0.07	0.15	0.07	0.16	0.00	0.22	0.02	231	0.9878
18	13	-0.41	0.17	-0.38	0.18	-0.03	0.25	-0.12	229	0.9059
18	14	-0.13	0.17	-0.29	0.18	0.16	0.24	0.66	231	0.5124
19	1	-0.23	0.18	-0.23	0.17	0.00	0.25	-0.02	206	0.9858
19	2	-0.84	0.18	-1.06	0.18	0.23	0.26	0.88	207	0.3818
19	3	-0.48	0.16	-0.41	0.16	-0.07	0.23	-0.30	207	0.7637
19	4	-0.23	0.16	-0.21	0.16	-0.03	0.23	-0.12	207	0.9064
19	5	-0.09	0.16	-0.39	0.16	0.30	0.23	1.32	207	0.1891
19	6	-0.31	0.17	-0.28	0.16	-0.03	0.23	-0.13	206	0.8948
19	7	-0.07	0.15	0.02	0.15	-0.09	0.21	-0.42	205	0.6780
19	8	-1.15	0.19	-1.23	0.19	0.07	0.26	0.28	207	0.7790
19	9	0.06	0.15	0.62	0.13	-0.56	0.19	-2.86	205	0.0047
19	10	0.05	0.16	-0.05	0.15	0.10	0.22	0.45	207	0.6562

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
19	11	-0.62	0.16	-0.83	0.16	0.21	0.23	0.91	203	0.3619
19	12	-0.24	0.17	0.02	0.16	-0.26	0.23	-1.11	206	0.2681
19	13	1.34	0.25	1.01	0.23	0.33	0.34	0.97	203	0.3340
19	14	-0.51	0.19	-0.99	0.19	0.49	0.27	1.82	205	0.0709
20	1	-0.13	0.20	0.08	0.19	-0.20	0.28	-0.73	158	0.4663
20	2	-0.28	0.20	-0.37	0.19	0.09	0.28	0.32	159	0.7521
20	3	-0.08	0.18	0.18	0.16	-0.26	0.25	-1.08	158	0.2839
20	4	-0.63	0.20	-0.33	0.18	-0.30	0.27	-1.10	158	0.2715
20	5	-0.65	0.20	-0.26	0.18	-0.39	0.27	-1.41	158	0.1598
20	6	-0.50	0.20	-0.34	0.18	-0.16	0.27	-0.59	158	0.5591
20	7	-0.57	0.20	-0.44	0.18	-0.13	0.26	-0.50	158	0.6188
20	8	-0.95	0.22	-0.76	0.19	-0.19	0.29	-0.66	157	0.5106
20	9	-0.33	0.19	-0.71	0.19	0.37	0.26	1.42	159	0.1573
20	10	-0.73	0.21	-0.57	0.19	-0.16	0.28	-0.56	158	0.5773
20	11	-0.43	0.19	-0.82	0.18	0.39	0.26	1.49	159	0.1388
20	12	0.25	0.19	-0.29	0.19	0.54	0.27	2.02	158	0.0454
20	13	1.66	0.26	1.39	0.27	0.27	0.38	0.72	157	0.4731
20	14	-0.10	0.21	0.02	0.19	-0.11	0.28	-0.40	158	0.6873
21	1	-0.09	0.18	-0.20	0.17	0.11	0.25	0.45	196	0.6551
21	2	0.62	0.16	0.65	0.14	-0.02	0.21	-0.10	196	0.9198
21	3	0.67	0.15	0.85	0.13	-0.19	0.20	-0.95	196	0.3453
21	4	-0.54	0.18	-0.16	0.15	-0.39	0.23	-1.67	195	0.0972
21	5	-0.24	0.17	-0.26	0.15	0.02	0.23	0.09	196	0.9293
21	6	-0.30	0.17	-0.23	0.15	-0.07	0.23	-0.29	196	0.7702
21	7	-0.14	0.16	0.23	0.14	-0.37	0.21	-1.75	195	0.0814
21	8	-0.91	0.18	-1.17	0.17	0.26	0.25	1.04	197	0.2983
21	9	-0.47	0.17	-0.31	0.14	-0.16	0.22	-0.72	195	0.4722
21	10	-0.24	0.17	-0.52	0.16	0.28	0.23	1.22	197	0.2230
21	11	-0.55	0.17	-0.78	0.15	0.23	0.23	1.02	197	0.3112
21	12	-0.40	0.18	-0.20	0.16	-0.20	0.24	-0.81	195	0.4171
21	13	1.60	0.25	0.73	0.24	0.87	0.34	2.56	196	0.0112
21	14	0.14	0.18	0.25	0.16	-0.11	0.24	-0.46	194	0.6432
22	1	-0.67	0.65	0.01	0.51	-0.68	0.82	-0.83	14	0.4230
22	2	0.06	0.56	-0.29	0.52	0.34	0.76	0.45	14	0.6575
22	3	0.62	0.50	0.41	0.47	0.21	0.68	0.30	14	0.7660
22	4	0.35	0.52	0.20	0.48	0.15	0.71	0.22	14	0.8328
22	5	-1.75	0.73	0.22	0.47	-1.97	0.87	-2.26	13	0.0415
22	6	0.16	0.52	0.46	0.46	-0.30	0.70	-0.43	14	0.6764
22	7	0.84	0.48	0.41	0.46	0.44	0.67	0.66	14	0.5227
22	8	-0.20	0.50	-0.76	0.48	0.57	0.69	0.82	14	0.4259
22	9	0.14	0.50	0.20	0.46	-0.06	0.68	-0.09	14	0.9327
22	10	0.13	0.53	-0.48	0.50	0.62	0.73	0.84	14	0.4141
22	11	-0.29	0.55	-1.41	0.57	1.13	0.79	1.43	14	0.1759
22	12	0.04	0.60	-0.28	0.54	0.32	0.80	0.40	14	0.6955
22	13	2.65	0.66	2.46	0.61	0.19	0.90	0.21	14	0.8401
22	14	-0.35	0.63	1.02	0.48	-1.37	0.80	-1.72	14	0.1080
23	1	0.69	0.26	0.51	0.23	0.18	0.35	0.50	87	0.6193
23	2	1.75	0.21	2.30	0.17	-0.55	0.27	-1.99	85	0.0495
23	3	2.13	0.20	1.95	0.17	0.18	0.26	0.69	87	0.4910
23	4	-0.09	0.26	-0.24	0.23	0.15	0.35	0.42	87	0.6721
23	5	-0.57	0.28	-0.99	0.25	0.42	0.38	1.10	88	0.2738
23	6	0.36	0.25	-0.10	0.23	0.46	0.34	1.37	88	0.1732
23	7	0.16	0.24	0.42	0.20	-0.26	0.32	-0.84	86	0.4045
23	8	-0.70	0.29	-0.89	0.27	0.19	0.39	0.50	88	0.6215

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
23	9	0.06	0.24	0.30	0.20	-0.24	0.32	-0.74	86	0.4601
23	10	-0.72	0.28	-0.55	0.24	-0.17	0.37	-0.46	86	0.6485
23	11	1.32	0.21	0.93	0.18	0.39	0.27	1.42	86	0.1606
23	12	-0.33	0.27	-0.18	0.23	-0.15	0.36	-0.42	86	0.6764
23	13	-1.12	0.26	-0.40	0.23	-0.72	0.35	-2.05	88	0.0434
23	14	0.55	0.28	-0.19	0.25	0.74	0.38	1.97	88	0.0521
24	1	-0.94	0.34	-0.58	0.30	-0.36	0.46	-0.78	65	0.4406
24	2	-0.73	0.32	-0.21	0.28	-0.52	0.43	-1.21	65	0.2322
24	3	-0.02	0.27	-0.03	0.25	0.01	0.37	0.02	64	0.9840
24	4	-1.00	0.32	-0.78	0.29	-0.22	0.43	-0.50	65	0.6155
24	5	-0.51	0.31	-0.94	0.30	0.43	0.43	1.01	64	0.3163
24	6	-0.76	0.31	-0.91	0.30	0.15	0.43	0.34	65	0.7315
24	7	-0.02	0.27	-0.36	0.26	0.34	0.37	0.90	64	0.3691
24	8	-0.85	0.31	-1.07	0.30	0.22	0.43	0.51	65	0.6084
24	9	-0.56	0.29	0.21	0.23	-0.77	0.37	-2.06	65	0.0438
24	10	-1.09	0.32	-0.94	0.29	-0.15	0.44	-0.35	65	0.7276
24	11	-0.97	0.30	-1.81	0.34	0.84	0.46	1.84	64	0.0706
24	12	-0.18	0.31	0.57	0.27	-0.76	0.41	-1.84	62	0.0701
24	13	1.31	0.40	1.29	0.45	0.02	0.60	0.04	64	0.9670
24	14	0.10	0.31	0.20	0.28	-0.10	0.42	-0.24	65	0.8082
25	1	-0.39	0.19	-0.43	0.20	0.04	0.27	0.15	180	0.8831
25	2	0.08	0.17	0.12	0.18	-0.04	0.25	-0.17	182	0.8682
25	3	0.88	0.15	1.14	0.15	-0.26	0.21	-1.25	182	0.2112
25	4	0.37	0.16	0.57	0.16	-0.20	0.23	-0.86	182	0.3890
25	5	0.01	0.17	0.34	0.17	-0.33	0.24	-1.40	183	0.1630
25	6	0.30	0.16	0.04	0.17	0.26	0.24	1.09	181	0.2776
25	7	0.21	0.16	0.60	0.16	-0.39	0.22	-1.76	182	0.0796
25	8	-0.63	0.19	-0.98	0.20	0.36	0.28	1.29	180	0.1974
25	9	0.79	0.14	0.78	0.15	0.00	0.21	0.02	181	0.9845
25	10	-0.50	0.18	-0.30	0.18	-0.20	0.25	-0.78	181	0.4341
25	11	-0.52	0.17	-1.23	0.21	0.71	0.27	2.63	176	0.0092
25	12	0.27	0.17	0.72	0.17	-0.45	0.24	-1.86	179	0.0641
25	13	1.95	0.24	1.42	0.27	0.53	0.36	1.48	179	0.1401
25	14	0.58	0.17	0.15	0.19	0.43	0.26	1.67	180	0.0966
26	1	-0.50	0.16	-0.28	0.15	-0.22	0.22	-1.02	264	0.3070
26	2	-0.58	0.16	-0.57	0.15	-0.01	0.21	-0.04	265	0.9670
26	3	0.71	0.12	0.37	0.12	0.33	0.17	1.95	267	0.0528
26	4	0.24	0.13	0.60	0.12	-0.36	0.18	-1.99	266	0.0473
26	5	0.07	0.14	0.18	0.13	-0.12	0.19	-0.63	266	0.5312
26	6	-0.36	0.15	-0.25	0.14	-0.12	0.20	-0.58	265	0.5615
26	7	-0.11	0.13	0.09	0.12	-0.21	0.18	-1.15	266	0.2503
26	8	-1.21	0.17	-1.30	0.16	0.09	0.23	0.38	266	0.7048
26	9	0.14	0.13	0.69	0.11	-0.56	0.17	-3.31	265	0.0010
26	10	0.22	0.13	-0.09	0.13	0.31	0.19	1.66	267	0.0981
26	11	-0.65	0.14	-1.24	0.15	0.59	0.20	2.88	266	0.0043
26	12	0.19	0.15	-0.14	0.14	0.33	0.20	1.63	263	0.1040
26	13	2.32	0.21	2.29	0.19	0.03	0.29	0.11	264	0.9100
26	14	0.11	0.15	0.20	0.14	-0.08	0.21	-0.40	265	0.6863
27	1	-0.35	0.14	-0.29	0.13	-0.05	0.19	-0.29	395	0.7706
27	2	-0.33	0.13	-0.40	0.13	0.07	0.18	0.40	397	0.6901
27	3	0.24	0.11	-0.08	0.11	0.32	0.16	2.02	397	0.0435
27	4	-0.23	0.12	-0.08	0.12	-0.14	0.17	-0.85	396	0.3962
27	5	-0.26	0.12	-0.35	0.12	0.09	0.17	0.52	395	0.6019
27	6	0.08	0.12	0.23	0.11	-0.16	0.16	-0.96	393	0.3355

Experiment # (equated)	Question #	Male students		Female students		Comparison: Welch's T-test				
		δ (logits)	Standard error	δ (logits)	Standard error	Change in δ (male – female)	Joint standard error	t	d.f.	probability (p)
27	7	-0.43	0.12	-0.33	0.12	-0.11	0.17	-0.62	395	0.5328
27	8	-0.98	0.14	-0.92	0.13	-0.06	0.19	-0.29	395	0.7690
27	9	-0.44	0.12	-0.44	0.12	-0.01	0.17	-0.03	392	0.9739
27	10	-0.19	0.12	-0.09	0.12	-0.10	0.17	-0.59	394	0.5530
27	11	-0.67	0.12	-0.88	0.12	0.20	0.17	1.17	394	0.2413
27	12	-0.16	0.13	-0.12	0.12	-0.04	0.17	-0.20	393	0.8378
27	13	1.16	0.17	1.17	0.17	-0.01	0.24	-0.06	396	0.9538
27	14	-0.49	0.14	-0.48	0.13	-0.01	0.20	-0.07	392	0.9464
28	1	-0.23	0.25	-0.45	0.25	0.22	0.35	0.63	105	0.5275
28	2	-0.85	0.25	-0.85	0.25	0.00	0.35	0.00	105	0.9987
28	3	1.21	0.18	1.19	0.18	0.02	0.25	0.09	105	0.9320
28	4	-0.69	0.24	0.00	0.21	-0.68	0.32	-2.11	103	0.0371
28	5	-0.50	0.23	-0.24	0.22	-0.26	0.32	-0.80	105	0.4228
28	6	0.87	0.19	0.96	0.18	-0.09	0.27	-0.33	104	0.7434
28	7	-0.20	0.21	-0.28	0.21	0.08	0.30	0.26	105	0.7942
28	8	-1.67	0.30	-1.10	0.25	-0.57	0.39	-1.46	104	0.1482
28	9	-0.31	0.22	-0.46	0.22	0.15	0.31	0.50	105	0.6188
28	10	-0.19	0.22	-0.06	0.21	-0.13	0.31	-0.41	105	0.6823
28	11	1.31	0.18	0.89	0.18	0.42	0.25	1.68	100	0.0965
28	12	-0.32	0.24	-0.53	0.23	0.21	0.33	0.63	105	0.5275
28	13	-1.01	0.25	-1.27	0.23	0.26	0.33	0.78	103	0.4379
28	14	-0.28	0.26	-0.26	0.25	-0.01	0.36	-0.03	105	0.9733
29	1	-0.47	0.22	-0.18	0.20	-0.29	0.30	-0.98	145	0.3269
29	2	-1.39	0.23	-1.24	0.21	-0.15	0.31	-0.48	147	0.6338
29	3	-0.04	0.19	-0.20	0.18	0.16	0.26	0.61	148	0.5440
29	4	0.18	0.19	-0.07	0.18	0.25	0.26	0.96	148	0.3379
29	5	0.18	0.19	-0.05	0.18	0.23	0.26	0.89	148	0.3756
29	6	-0.15	0.20	-0.11	0.18	-0.04	0.27	-0.15	147	0.8807
29	7	0.03	0.18	0.06	0.17	-0.03	0.25	-0.14	147	0.8891
29	8	-0.80	0.21	-0.86	0.21	0.06	0.29	0.21	147	0.8349
29	9	-0.01	0.18	-0.03	0.17	0.02	0.25	0.09	145	0.9288
29	10	-0.26	0.20	0.08	0.17	-0.34	0.26	-1.28	147	0.2032
29	11	0.97	0.17	0.86	0.16	0.12	0.23	0.50	129	0.6150
29	12	-0.26	0.21	-0.40	0.19	0.14	0.28	0.49	147	0.6239
29	13	1.46	0.29	1.22	0.27	0.24	0.40	0.61	147	0.5432
29	14	-0.71	0.23	-0.49	0.21	-0.21	0.31	-0.69	146	0.4932

7.6 Supporting information for sections 4.3 and 4.4

7.6.1 Correlations used for equating prior to factor analysis

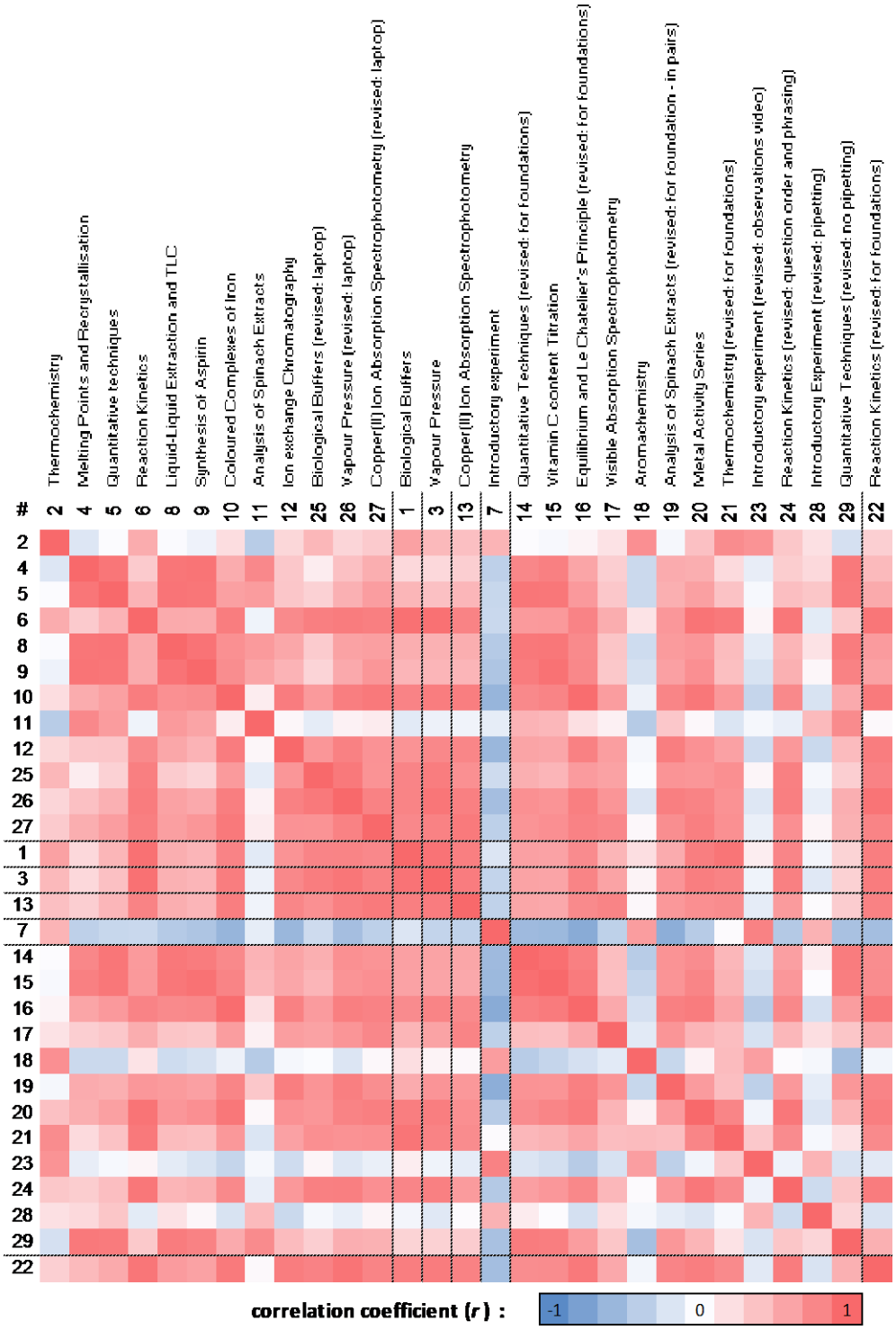


Figure S 52: Correlations between initial δ_{PCM} measures for different experiments

Dotted lines are used to separate isolated subsets of data. Experiments 10 and 16 were selected to be artificially equated based on this analysis, owing to their strong correlation and equivalence in design.

Similarity between measures for Experiment 10 in subset 1 (Coloured complexes of iron) and Experiment 16 in subset 6 (Equilibrium and Le Chatelier's principle (revised: for foundations) are additionally shown in Figure S 52 and Figure S 53 below.

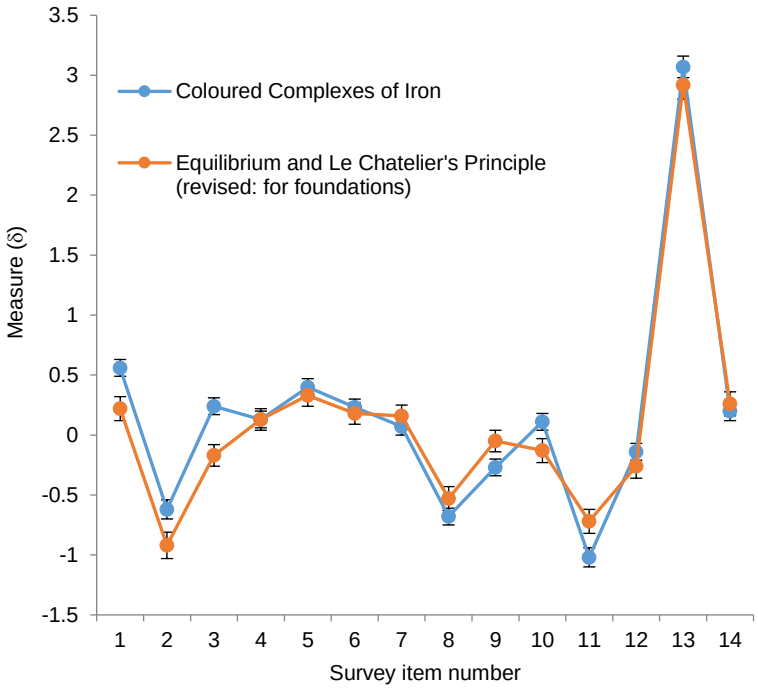


Figure S 53: Similarity of measures for experiments selected to be artificially equated

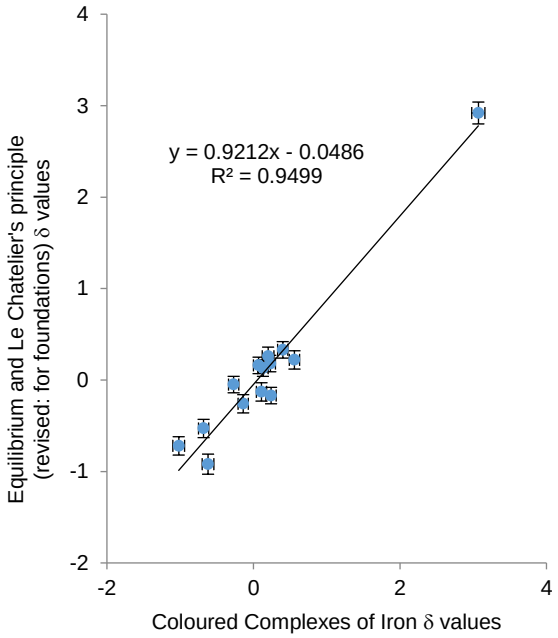


Figure S 54: Linear relationship between measures for experiments selected to be artificially equated

7.6.2 Estimating the final LLTM within Facets software

A large number of facets must be defined in order to estimate an LLTM within the Facets software:

- A **survey item facet** used as a "dummy facet" to easily define different rating scale structures for each survey item
- The **occasion-specific student bias measures** (β_E) must be included
- The **question-specific experiment measure** (δ) facet normally used for a Partial Credit Model is then defined and set as a "dummy facet" for the purposes of stipulating the LLTM structure: labels assigned to each element of this facet will be used to define the Q matrix.
- The η and μ values (**factor measures and item locations** respectively) are contained within one single "basic parameters" facet, each different measure with its own facet element number. It is these facet element numbers which are listed in the labels for each δ facet element, specifying which parameters to add together to calculate that specific δ value.
- The model must be defined to contain an extensive series of additional facets which the program can use to add additional "basic parameters" as instructed to do so by the Q matrix (stipulated by the dummy δ facet labels). Every additional time any basic parameter measure must be added (or subtracted) one more time, this requires an additional facet. In the code outlined here, the total number of basic parameter facets included is 100: the sum of the maximum values for each column in the Q matrix.

```
FACETS=104
;1 - items (dummy, used only to define rating scale structures)
;2 – Occasion-specific student bias measures
;3 - Question-specific experiment measures (dummy, codes for LLTM)
;4… - Factor measures and item locations
ENTERED-IN-DATA=1,2,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4, ...
```

Note that in the final line of code shown above, the number "4" is repeated 100 times. This is because facet number 4 is the facet containing the factor measures and item locations, which will be summated to approximate the δ measures. There are 100 times a basic parameter (either μ or η) may be stipulated to be either added or subtracted once more within the labels of the 3rd facet (to be coded as described below).

Specifications are included to define all facets as positively oriented (more positive measure implies positive response is more likely), to arrange data in the final output and to specify which of the facets is "free" or non-centred. The coding below has defined facet 2: the student measures, to be the non-centred facet. That is, the measures for this facet will not sum to zero. Measures for all other facets will be defined to sum to zero. The NULL specification has been used to allow later stipulation that no basic parameter of the LLTM be added at the relevant point of the summation.

```
NEGATIVE=0
ARRANGE=N
NONCENTER=2
NULL=0000
```

The MODEL specification must be used in such a way that it defines whether basic parameters of the model are added or subtracted. That is, it is used to define the sign of the Q matrix

coefficients. This is achieved by adding a negative sign where appropriate in front of the relevant basic parameter facets. A different model must be stipulated for each survey item, as each has a different series of Q matrix coefficients. Fourteen MODEL= statements will thus be needed for the ASLE survey LLTM. Below, “Q1Scale” and “Q14Scale” have been used as labels for the particular rating scale structures associated with survey questions 1 and 14. The “R4” indicates a rating scale structure with 4 Rasch-Andrich thresholds. Note the addition of one more “?” for each time any basic parameter may be added or subtracted one more time.

```
MODEL=1,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,
?,?,?,?,?,?,?,?,?,?,?,?,?,?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,?,?,?,?,?,?,?,?,?
,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,Q1scale
.
.     (other model statements for other survey items)
.
MODEL=14,?,?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,-?,?,?,?,?,?,?,?,?,?,?,
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,Q14s
cale
*
rating scale=Q1scale,R4
.
.   (specifications defining the rating scale structure for the other survey
items)
.
rating scale=Q14scale,R4
*
```

Facets other than the δ facet and basic parameter facet may be labelled as desired.

```
LABELS=
1= Survey item,D
1, data interpretation
.
.     (labels for other survey items)
.
14, overall learning experience
*
2= Person measures (occasion-specific and equated)
1, Ex 1 - P 1
2, Ex 1 - P 2
3, Ex 1 - P 3
.
.     (labels for other person measures)
.
9379, Ex 73 - P 5917
9380, Ex 71 - P 5520
9381-9462, Equated across occasions
*
```

The linear combination of basic parameters (several η values plus a μ value) which approximate each δ measure can be stipulated as labels for the δ facet elements. There is one δ facet element for each survey item, for each experiment. The label for each element of the dummy δ facet is a string of 100 four- digit numbers, each number being the element number associated with the next basic parameter to be added or subtracted. Whether that basic parameter is added or subtracted is stipulated within the MODEL= statement. For example, digits 0001 at the end of the label for the first facet element (below) stipulate the addition of

basic parameter 0001: the item location (μ) for survey item 1. The next δ value's label has a code of 0002 at the end, to add basic parameter 0002. The full string of text adds (or subtracts) each basic parameter the correct number of times to yield the δ value in accordance with the LLTM Q-matrix. Basic parameter 0101 is subtracted fifteen times within the specifications for the second δ value, for example, because of a -15 Q-matrix weighting of factor 01 on to item 2 (δ element 2 being item 2 for experiment 1).

```
3, Question specific experiment quality measures (LLTM codes),D
1,0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000030100000000000000000000000000000000000000000000000000000000000040104
0104010401040104010401000000000000501000000000000000000000000000000000000000000000
0601060100000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000001
2,0101010101010101010101010101010101010101010101010101010101010201000000000000
0000000000000000030100000000000000000000000000000000000000000000000000000000000040104
0104010401040104010401040104010401040105010501050105010501050105010000000000000000
0601000000000000000000000000000000000000000000000000000701070100000000000000000000
0000000000000000000000000000000000000000000000000000090100001001100110010000000000
0000000000000002
.
.     (other codes for basic parameter element numbers)
.
406,0129012900000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0006290629000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000014
*
4, LLTM Basic parameters
0001, Survey item 1 location
0002, Survey item 2 location
.
.     (other item location parameters)
.
0014, Survey item 14 location
0101, f1 Ex 1  "factor 1" - Biological Buffers
0102, f1 Ex 2  "factor 1" - Thermochemistry
0103, f1 Ex 3  "factor 1" - Vapour Pressure
0104, f1 Ex 4  "factor 1" - Melting Points and Recrystallisation
.
.     (other experiment specific factor measures)
.
1127, f11 Ex 27  "factor 11" - Copper(II) Ion Absorption Spectrophotometry
(revised: laptop)
1128, f11 Ex 28  "factor 11" - Introductory Experiment (revised: pipetting)
1129, f11 Ex 29  "factor 11" - Quantitative Techniques (revised: no pipetting)
*
```

Above, the first two digits of the LLTM basic parameter element numbers correspond to the factor number, whilst the second two digits correspond to the experiment number. For example, 0103 is the factor 1 measure for experiment 3 (Vapour pressure). This notation is simply out of convenience. The DValues= specification can then be entered to instruct the program that the additional basic parameter facets included initially in the ENTERED-IN-DATA specification have element numbers stipulated by particular locations in the string of digits for each label in facet 3 (the dummy δ facet).

```
DVALUES=
4,3,1,4
5,3,5,4
6,3,9,4
7,3,13,4
8,3,17,4
.
.      (other commands for reading labels of the basic parameter facet)
.
102,3,393,4
103,3,397,4
104,3,401,4
*
```

The specifications above stipulate the facet number (of the total 104), then the facet number whose label is to be referenced (facet 3, the dummy δ facet), the first digit of the total string of digits in the label to read when retrieving the relevant basic parameter for that facet, then the number of digits to read from that point. The stipulation **4,3,1,4** therefore instructs the program to add a basic parameter (specified in the ENTERED-IN-DATA specification initially) as the **4**th facet by reading the label of facet **3**, beginning with the **1**st digit and continuing to read **4** digits, then using this number as the element number of the basic parameters facet to retrieve.

In the study discussed within this thesis, data was entered into the specification file as one line of code per data point. Only the first 3 facets need to be specified, as all basic parameters are entered using the DValues specifications. The four digits listed specify the survey item number, the number of the relevant occasion-specific student bias measure, the δ facet number corresponding to the experiment and survey item for which the observed response was gathered, then the observed response (0 through to 4 for the lowest to highest response category respectively)

```
DATA=
1,1,57,3
2,1,58,4
3,1,59,4
4,1,60,4
5,1,61,2
.
.      (one line for each data point)
.
11,9380,25,4
12,9380,26,4
13,9380,27,3
14,9380,28,3
```

7.6.3 Full matrices comprising the final LLTM

The ASLE survey data can be described by Equation 41 reproduced below, where H is a matrix of parameters estimated from the data (Figure S 57) and Q is a matrix of coefficients weighting each of these factors' contributions to each ASLE survey item (Figure S 56). The product of these matrices yields a matrix of item-specific experiment quality measures (Figure S 58) able to predict student responses to ASLE surveys in conjunction with Equations 43 and 1.

$$\left[\delta_{LLTM\,i,m}\right] = \boldsymbol{Q} \times \boldsymbol{H} \qquad \textbf{41}$$

The most accurate (unrounded) estimate for the Q-matrix is:

one column for each factor | for inclusion of μ parameters

one row for each survey item

0.022	0.007	0.079	0.812	-0.098	0.19	0.016	0.01	0.011	-0.007	-0.006	-0.002	1	0	0	0	0	0	0	0	0	0	0	0	0
-0.695	0.054	-0.07	0.46	0.333	-0.052	0.073	-0.01	-0.032	0.162	0.013	0.023	0	1	0	0	0	0	0	0	0	0	0	0	0
-0.72	0.36	0.333	0.069	0.07	-0.183	0.238	0.054	0.147	0.042	-0.07	0.013	0	0	1	0	0	0	0	0	0	0	0	0	0
0.134	0.841	0.133	0.008	-0.034	0.099	0.229	-0.045	0.179	-0.051	0.018	0.013	0	0	0	1	0	0	0	0	0	0	0	0	0
0.686	0.497	0.074	0.135	0.093	0.184	0.187	0.078	0.142	-0.064	0.03	-0.027	0	0	0	0	1	0	0	0	0	0	0	0	0
0.058	0.058	0.814	0.032	-0.122	-0.075	0.051	-0.046	0.01	-0.002	0.002	0	0	0	0	0	0	1	0	0	0	0	0	0	0
-0.268	0.695	-0.09	0.301	0.158	0.136	0.123	-0.099	-0.008	0.15	-0.032	-0.033	0	0	0	0	0	0	1	0	0	0	0	0	0
0.115	0.104	-0.044	0.232	-0.022	0.685	0.134	0.016	0.002	0.002	-0.013	-0.001	0	0	0	0	0	0	0	1	0	0	0	0	0
0.072	0.755	-0.055	-0.143	0.209	-0.016	-0.037	0.063	-0.149	-0.021	-0.002	0.011	0	0	0	0	0	0	0	0	1	0	0	0	0
0.775	0.096	0.268	0.153	-0.13	-0.012	-0.165	0.018	-0.004	0.113	-0.051	0.035	0	0	0	0	0	0	0	0	0	1	0	0	0
0.073	-0.178	0.561	0.089	-0.476	0.102	-0.206	0.29	-0.007	-0.01	-0.014	0	0	0	0	0	0	0	0	0	0	0	1	0	0
-0.161	0.175	-0.189	-0.06	0.768	-0.025	0.06	0.023	-0.003	0.001	-0.007	-0.002	0	0	0	0	0	0	0	0	0	0	0	1	0
0.504	0.043	-0.069	-0.264	-0.254	0.458	0.404	-0.078	0.009	-0.022	0.16	0.001	0	0	0	0	0	0	0	0	0	0	0	1	0
-0.265	0.294	0.013	0.089	0.199	0.308	0.668	-0.013	0.012	0.007	-0.016	-0.002	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure S 55: Preliminary Q-matrix of non-integer values

whilst the matrix of integer values below is approximately a scalar multiple of the above (scalar value k = 24.4, see Equation 40), yet suitable for use with the *Facets* software.

one column for each factor | factor 12, can be excluded | for inclusion of μ parameters

one row for each survey item

0	0	1	7	-1	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
-15	1	-1	10	7	-1	2	0	-1	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
-10	5	5	1	1	-3	3	1	2	1	-1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1	6	1	0	0	1	2	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	3	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1	1	7	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
-2	5	-1	2	1	1	1	-1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	1	0	1	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
1	8	-1	-1	2	0	0	1	-2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
7	1	2	1	-1	0	-1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2	-4	14	2	-11	2	-5	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
-1	1	-1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
14	1	-2	-8	-7	13	12	-2	0	-1	5	0	0	0	0	0	0	0	0	0	0	0	0	1	0
-2	2	0	1	1	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure S 56: Integer value Q-matrix used for the final LLTM

It is this Q-matrix (Figure S 56) which was used in the main study. Note that factor 12 coefficients are all zero, meaning this factor could be excluded.

The matrix H in Equation 41 contains all parameters estimated for the LLTM, and is shown in full below. Note that the item locations (μ) are fixed to be identical for each experiment (column) and that factor measures for factor 12 all appear as zero, since this factor did not contribute (and could therefore be excluded) when using the matrix of integer values. The factor measures for each experiment contained within this matrix are more conveniently presented as figures within section 7.6.4, including error margins for each.

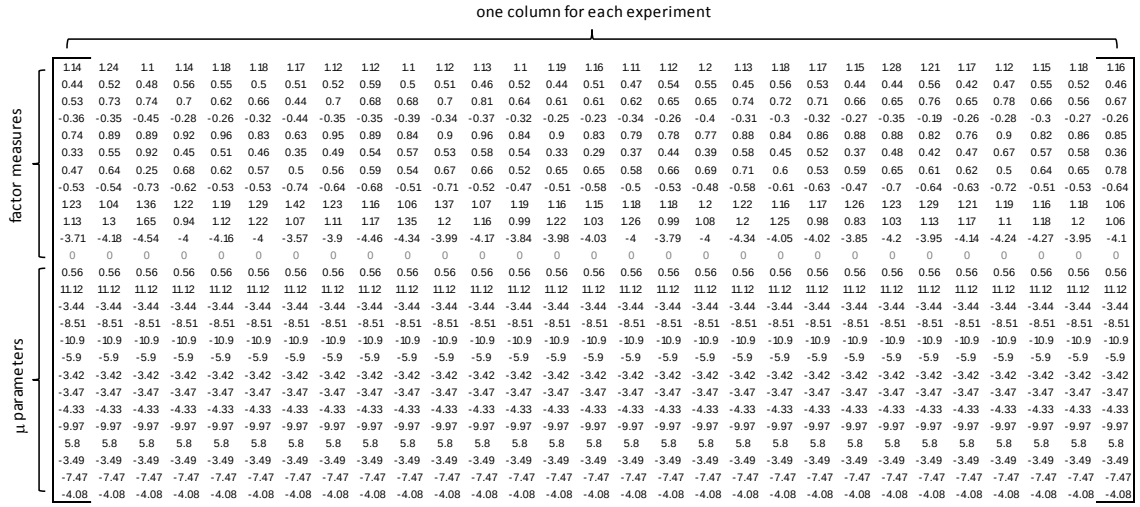


Figure S 57: LLTM basic parameter matrix H

As can be seen in Equation 41, the product of the matrices Q and H yields a new matrix containing all student independent measures for each evaluated experiment, approximately equal to those which would be estimated from a fully connected Partial Credit Model. This (annotated) matrix of δ_{LLTM} values is presented in full below.

ASLE survey item	Biological Buffers	Thermochemistry	Vapour Pressure	Melting Points and Recrystallisation	Quantitative techniques	Reaction Kinetics	Introductory experiment	Liquid-Liquid Extraction and TLC	Synthesis of Aspirin	Coloured Complexes of Iron	Analysis of Spinach Extracts	Ion exchange Chromatography	Copper(II) Ion Absorption Spectrophotometry	Quantitative Techniques (revised: for foundations)	Determination of Vitamin C concentration	Equilibrium and Le Chatalier's Principle (revised: for foundations)	Visible Absorption Spectrophotometry	Aromachemistry	Analysis of Spinach Extracts (revised: for foundation - in pairs)	Metal Activity Series	Thermochemistry (revised: for foundations)	Reaction Kinetics (revised: for foundations)	Introductory experiment (revised: observations video)	Reaction Kinetics (revised: question order and phrasing)	Biological Buffers (revised: laptop)	Vapour Pressure (revised: laptop)	Copper(II) Ion Absorption Spectrophotometry (revised: laptop)	Introductory Experiment (revised: pipetting)	Quantitative Techniques (revised: no pipetting)
1	-1.51	-0.95	-0.9	-0.72	-0.58	-0.93	-2.01	-1.16	-1.02	-1.19	-0.96	-1.02	-0.8	-0.82	-0.69	-1.25	-0.51	-1.58	-0.59	-0.76	-0.79	-0.81	-1.16	0.01	-0.43	-0.18	-0.56	-0.47	-0.72
2	-1.72	-1.37	-0.74	0.03	0.37	-1.08	-3.91	0.02	-0.05	-0.08	0.07	-0.01	-0.54	0.37	0.08	-0.16	-0.26	-2.56	0.16	-0.52	-1.48	-0.85	-2.95	-0.49	-0.87	-0.05	-0.41	-0.16	0.18
3	-2.42	-1.76	-1.73	-0.45	-1.23	-1.33	-3.01	-0.9	-0.33	-0.88	-0.39	-0.61	-1.44	-1.47	-0.98	-0.89	-0.9	-1.17	-0.43	-0.84	-1.66	-1.44	-2.76	-0.72	-1.74	-1.28	-0.9	-2.06	-0.9
4	-1.7	-0.55	-1.01	-0.28	-0.47	-0.78	-1.07	-0.73	-0.29	-1.02	-0.39	-0.84	-0.88	-1.28	-0.94	-1.25	-0.56	-0.39	-0.72	-0.44	-0.7	-1.25	-0.93	-0.25	-1.25	-0.93	-0.39	-0.59	-0.94
5	-1.5	-0.21	-1.01	-0.48	-0.27	-0.57	-1.39	-0.83	-0.73	-1.14	-0.68	-0.71	-0.93	-0.77	-0.88	-1.27	-0.79	-0.46	-0.69	-0.49	-0.65	-0.84	-0.37	-0.13	-1	-0.88	-0.49	-0.44	-0.93
6	-1.68	-0.47	-0.95	-0.67	-1.3	-0.89	-2.12	-0.8	-0.86	-0.95	-0.8	-0.18	-1.18	-1.23	-1.08	-1.14	-0.91	-0.76	-0.6	-0.41	-0.61	-0.94	-0.99	-0.05	-0.99	-0.42	-0.97	-1.72	-0.8
7	-1.55	-0.81	-0.42	-0.55	-0.43	-0.97	-1.24	-0.71	-0.22	-0.77	-0.48	-1.05	-0.94	-1.1	-0.88	-1.09	-0.73	-1.11	-0.84	-0.55	-0.94	-1.58	-1.39	-0.56	-1.18	-0.76	-0.51	-0.46	-0.94
8	-0.79	0.23	0.67	-0.02	0.15	-0.16	-0.68	-0.15	0.1	-0.01	0.08	0.15	-0.03	-0.45	-0.51	-0.54	-0.09	-0.26	0.25	-0.08	-0	-0.45	-0.01	-0.02	-0.11	0.35	0.28	0.35	-0.25
9	-1.35	-0.15	-1.35	-0.35	-0.1	-0.94	-1.4	-0.6	-0.04	-0.47	-1.14	-0.7	-0.56	-1.01	-0.69	-1.02	-0.61	-0.32	-1.29	-0.34	-0.56	-1.28	-1.23	-0.79	-1.72	-1.25	-0.33	-0.45	-0.96
10	-0.93	0.11	-0.25	-0.97	-0.64	-0.39	-0.89	-0.96	-0.84	-0.83	-0.93	-0.81	-1.16	-0.56	-0.8	-0.94	-1	-0.5	-0.83	-0.2	-0.56	-1.07	-0.12	0.09	-0.53	-0.68	-0.63	-0.65	-0.88
11	-0.52	0.05	1.23	-1.88	-2.23	-0.01	-2.53	-1.69	-1.92	0.37	-2.04	0.48	0.19	-1.6	-1.7	-0.21	-0.25	-0.2	-0.13	-0.21	-0.16	-0.14	-1.87	0.53	0.11	0.42	-0.11	-1.88	-1.87
12	-1.02	-0.49	-0.4	-0.17	0.06	-0.68	-1.44	-0.04	-0.25	-0.57	-0.3	-0.17	-0.51	-0.35	-0.6	-0.8	-0.82	-0.94	-0.51	-0.63	-0.54	-0.46	-0.58	-0.8	-1.09	-0.42	-0.65	-0.41	-0.61
13	-3.12	-0.77	-3.63	-2.24	-3.07	-2.36	0.76	-2.76	-4	-3.83	-1.06	-2.08	-2.14	-3.9	-4.36	-4	-1.17	-0.38	-2.04	-2.54	-2.14	-3.5	-0.36	-2.53	-2.27	-3.09	-1.9	0.04	-2.22
14	-2.09	-0.68	-1.79	-0.3	-0.52	-1.16	-2.01	-0.9	-0.57	-0.99	-0.33	-0.37	-1.04	-1.02	-0.95	-1.27	-0.54	-0.78	-0.16	-0.88	-1.13	-1.2	-1.02	-0.86	-1.04	-0.92	-0.42	-0.4	-0.27

Figure S 58: Matrix of δ_{LLTM} measures specific to survey item and experiment

7.6.4 Measures for basic factors contributing to ASLE survey responses

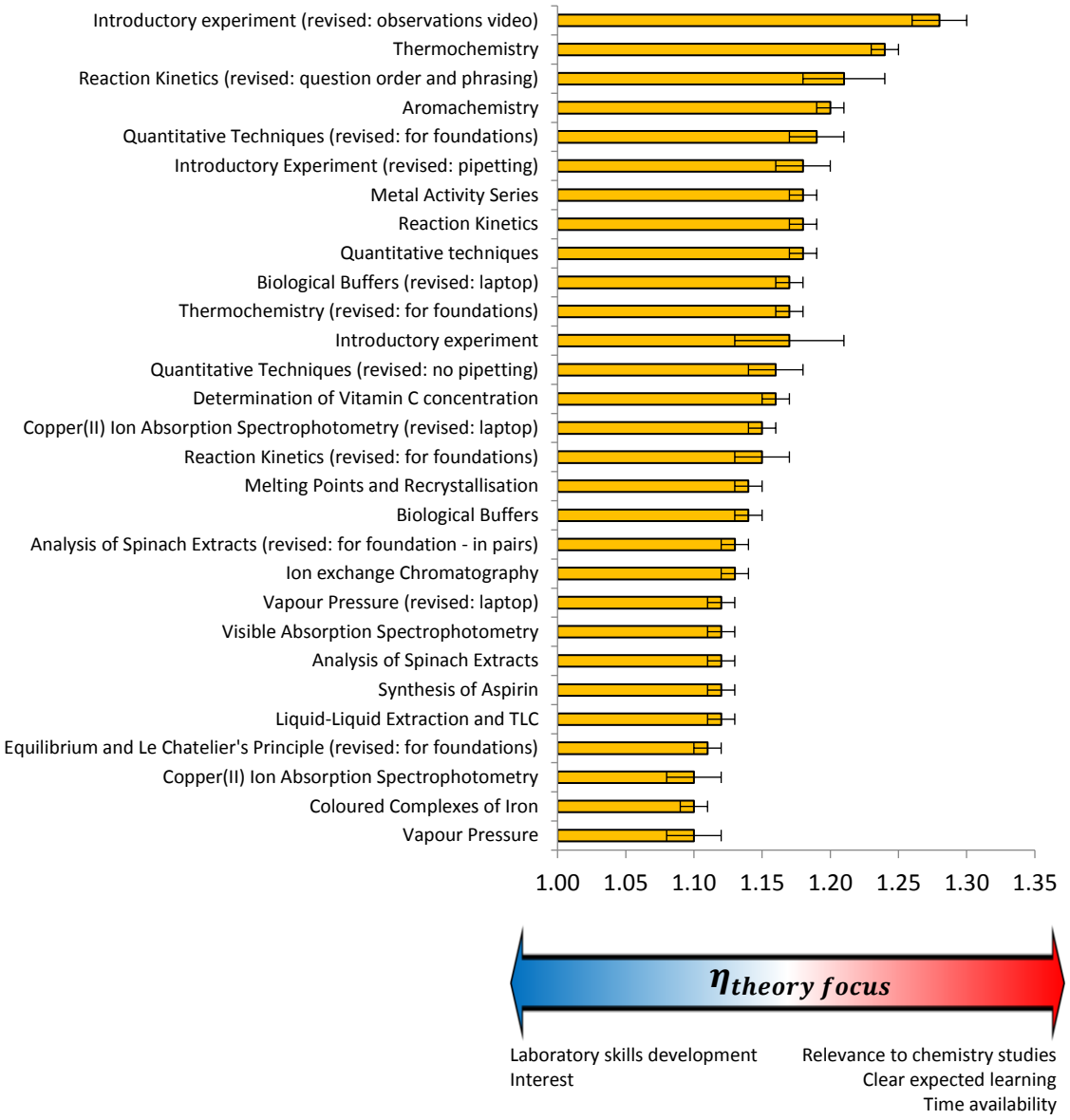


Figure S 59: Measures of latent factor 1 for all experiments

Analyses suggest this factor is largely related to the perceived connection between theoretical content in the experiment and content presented in the lecture course. The factor appears to be related to the *perceived* connection with *recognisable* theory from the perspective of the students. Not mere presence of any theory at all or presence of course content regardless of the familiarity of presentation. Experiments with high measures for this factor have an easily recognised relevance to the course and consequently clear expected learning outcomes, whereas experiments with low values for this factor are perceived as more "skills-based" and time consuming. A strong "boredom" response is seen for experiments inclusive of familiar lecture content, and so student interest also correlates with this factor.

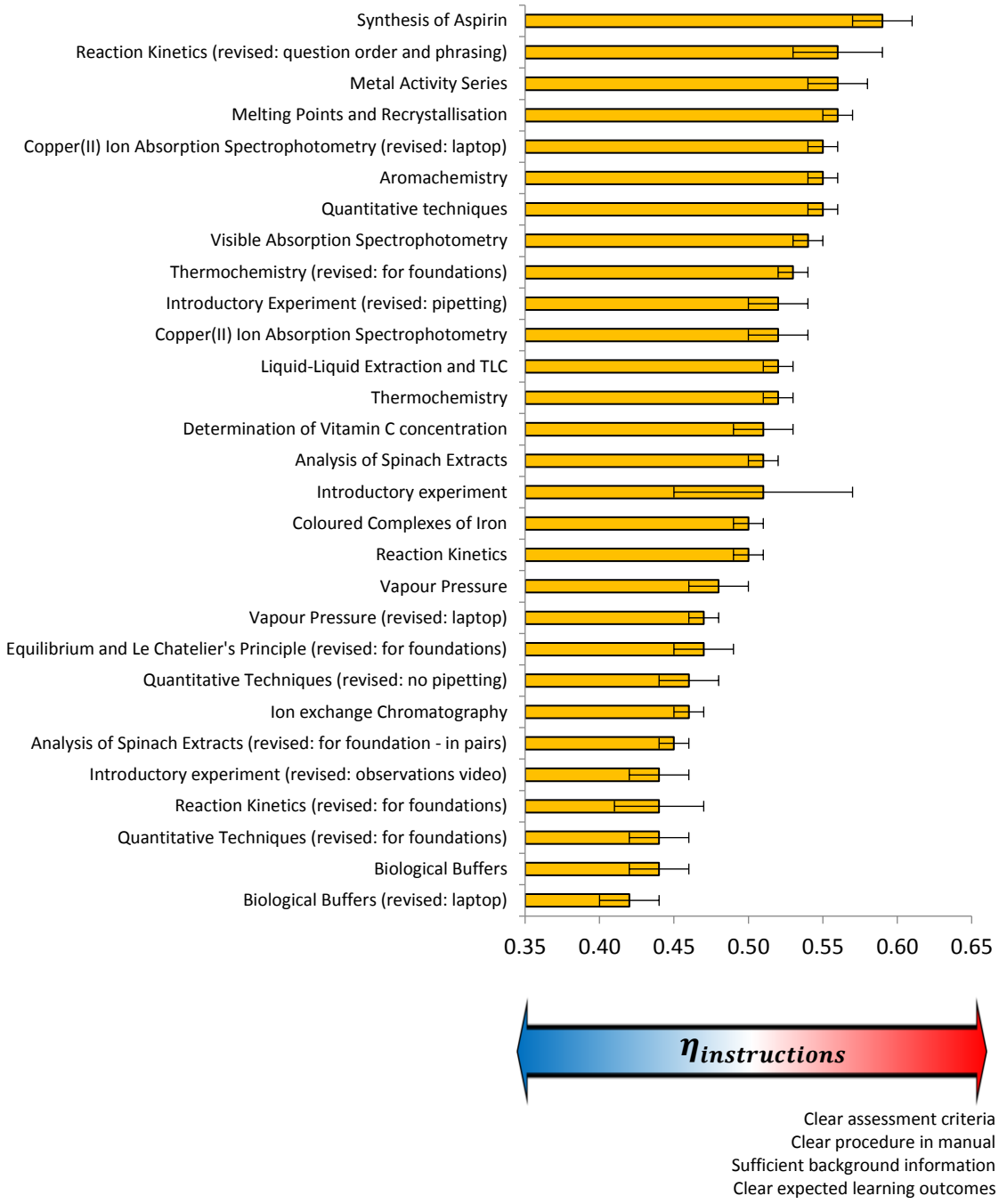


Figure S 60: Measures of latent factor 2 for all experiments

This factor most closely resembles an overall quality of instructional and guiding material provided to students conducting the experiment. Most ASLE item responses correlating with this factor's values involve a sense of clarity in the material. Patterns observed in values above, notably different iterations of the "Reaction kinetics" experiment, suggest that not only are the instructional notes relevant to this measure, but also the order and phrasing of questions asked within the laboratory notebook that are submitted for assessment.

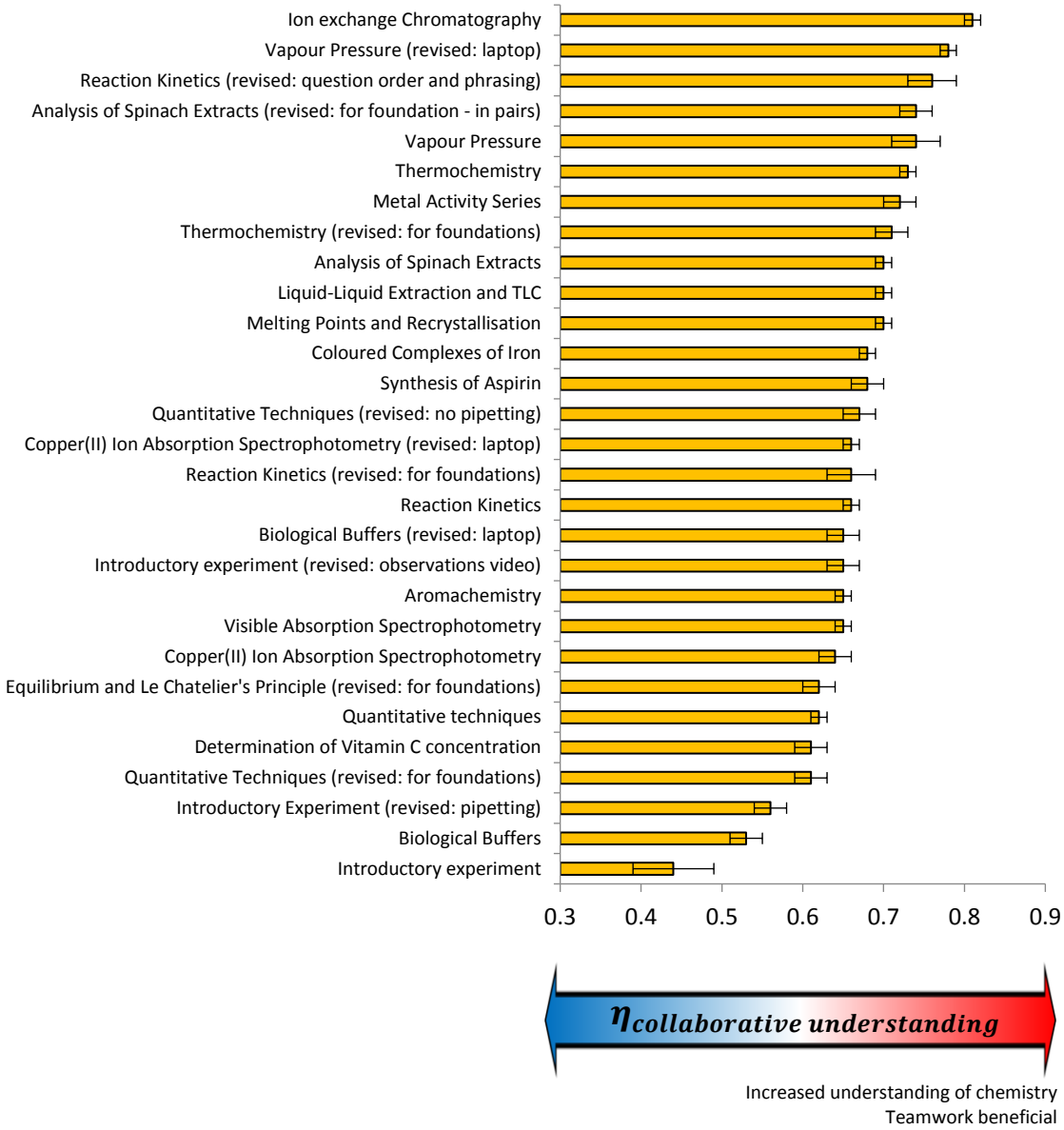


Figure S 61: Measures of latent factor 3 for all experiments

This factor exists as a manifestation of the fact that perceived benefits of teamwork and perceived increases in understanding are so closely relates as to manifest as a singular, irreducible factor within this data set. It is not clear from this analysis alone why this occurs: teamwork could promote understanding, understanding could prompt students to be more willing to help their classmates, or a confounding factor could cause the two to correlate for other reasons.

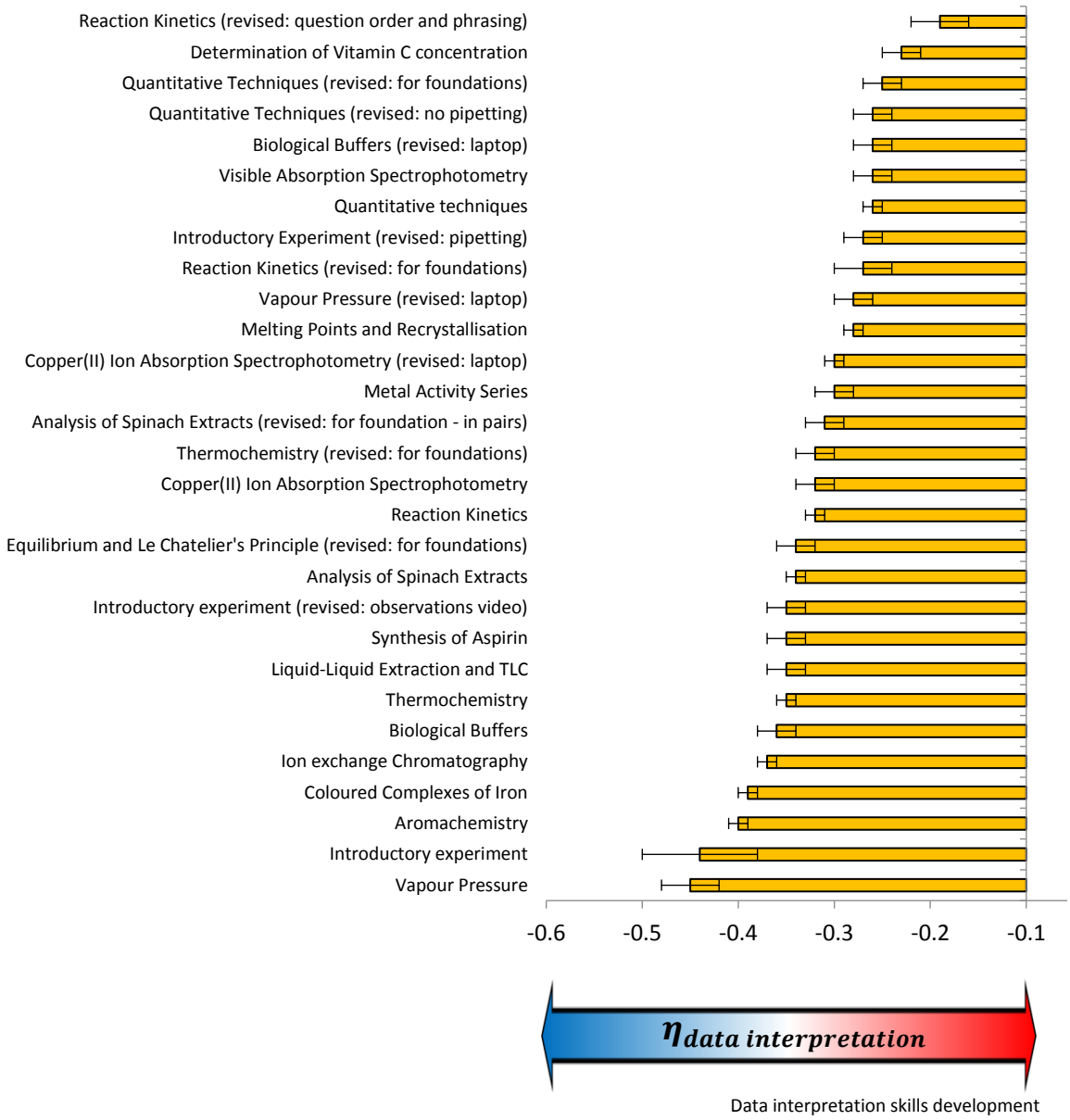


Figure S 62: Measures of latent factor 4 for all experiments

This factor most closely reflects the perceived increase in data interpretation skills from the students' perspective. More positive values of this factor (increased perception of skills development) does not necessarily arise merely from the inclusion of data interpretation tasks: the *development* of skills, rather than mere *use* of existing skills, appears to be important based on patterns in the observed measure values. Notably, the figure above shows higher measure values for experiments presented to students of lower ability (the Foundations of Chemistry cohort), presumably because skills utilised in the experiments were more often new, and therefore developed through the exercise. Simply including new analytical procedures is not sufficient either, however: "Vapour pressure" receives the lowest measure of all, likely because the unfamiliar graphing data logger involved was met with confusion and frustration.

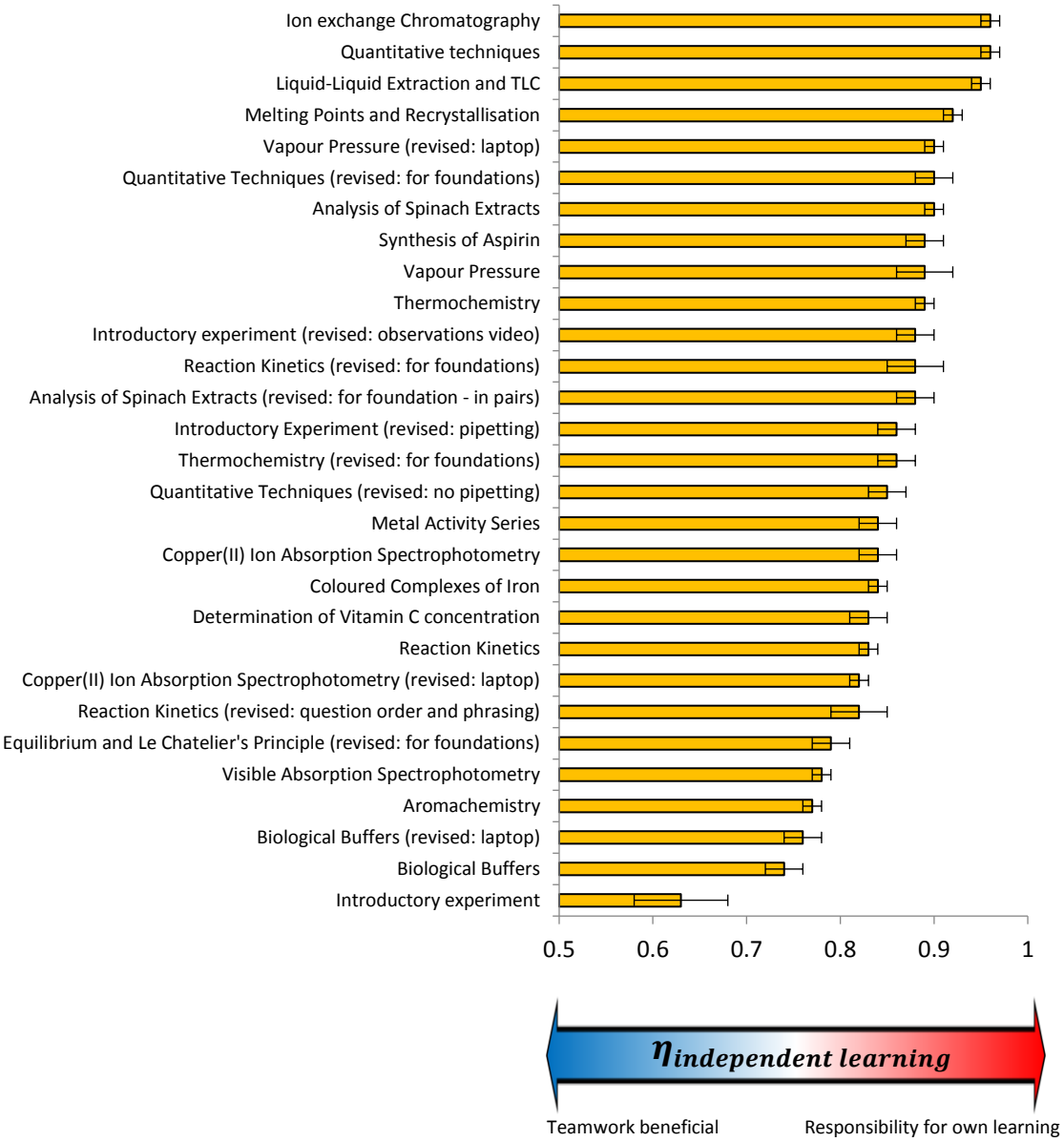


Figure S 63: Measures of latent factor 5 for all experiments

This factor most clearly resembles a spectrum from working individually (more positive values) to benefiting from working with others (less positive values), and maps reasonably closely (though not perfectly) to whether students worked in pairs (low values) or individually (high values). The value of this factor appears to depend on factors other than how students are required to work, however, given the appearance of an experiment conducted individually ("introductory experiment") as the lowest value and an experiment conducted in pairs ("Ion Exchange Chromatography") as the highest value, both in contrast to the broader trend. The prior knowledge of the student cohorts (Foundations of Chemistry IA/B or Chemistry IA/B) may play a role also, as experiments conducted by the less experienced Foundations cohort appear to cluster at the lower end of the spectrum, representing that teamwork is beneficial in these cases.

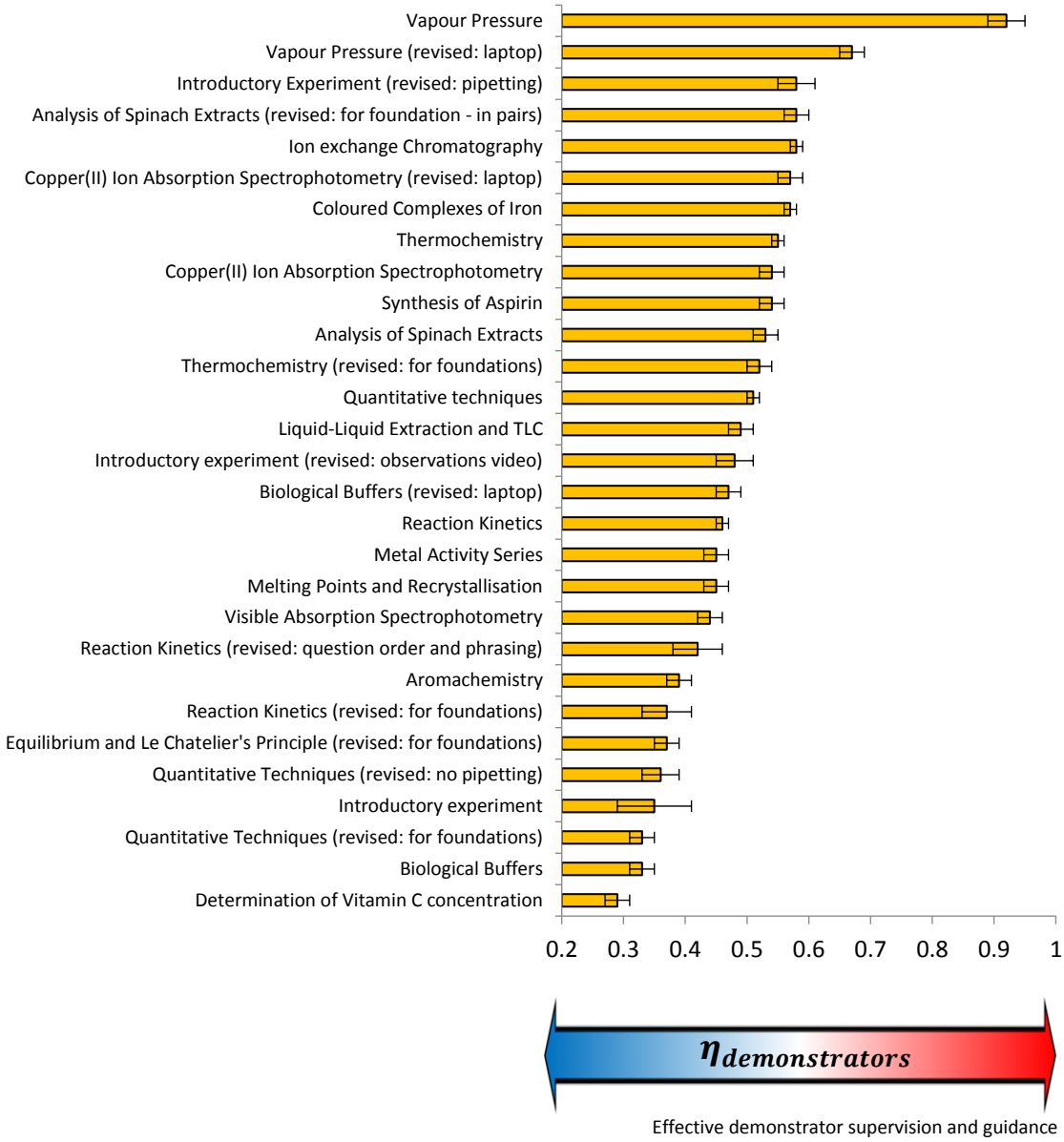


Figure S 64: Measures of latent factor 6 for all experiments

This factor most closely resembles the perception that the demonstrator's assistance in the experiment was effective. It is important to note that these values were extracted from data sets where multiple different demonstrators taught different subsets of the student groups, and so values of this factor cannot be interpreted as reflecting teacher quality. Rather, they reflect an attribute specific to the experiment itself which influences the perceived appreciation of the demonstrator (or possibly an aggregate view of the range of demonstrators who taught each experiment). Key to this interpretation is the "Vapour pressure" experiment as the highest value: demonstrators were appreciated to a far greater extent in this generally poorly received experiment than in any other case. However, when the identical experiment was run in absence of the poorly received handheld data logger device, this extremely positive perception of the demonstrators vanished. This factor appears to reflect a reliance on and appreciation for the demonstrator's help (for example as a result of a poor experiment), not the demonstrator's teaching ability.

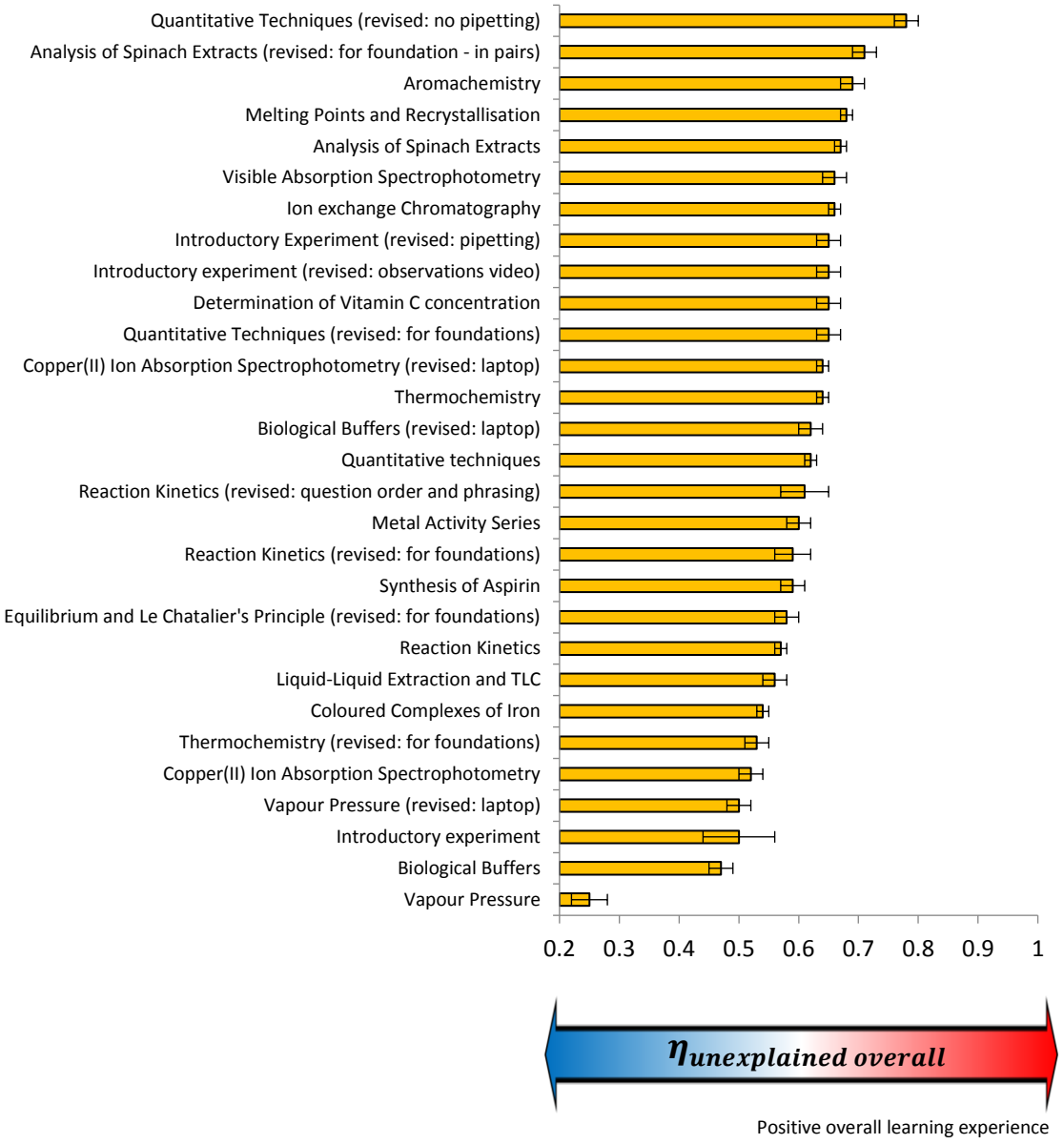


Figure S 65: Measures of latent factor 7 for all experiments

This factor has an unknown relation to laboratory activity design, but most closely resembles a perception of positive overall experience. Critically, it is a contribution to overall experience not resembling any items included on the ASLE survey instrument, and therefore represents a variation in this perception which is, as yet, unaccounted for. Were other items to be included on the ASLE survey and used to identify more factors than those comprising this LLTM, it is anticipated that measures for this factor would vary to a far lesser extent.