Indonesian secular vs. *Madrasah* schools: assessing the discrepancy in English reading and listening tests

Abu Nawas[1*] , I Gusti Ngurah Darmawan[1] and Nina Maadad[1]

*Correspondence:
abu.nawas@adelaide.edu.au

[1] School of Education,
Faculty of Arts, Business, Law
and Economics, The University
of Adelaide, Adelaide, SA 5000,
Australia

Abstract

The greater emphasis on the significance and difference in English performance between the school types has mainly been investigated across Asian countries. However, not much is known about what language skills differentiate their overall language achievement. Using a quantitative study with comparative analysis, this study measured the reading and listening skills of 1319 Indonesian students who were selected using a stratified sample design and grouped them into secular (*Sekolah*, $n = 726$) and Islamic (*Madrasah*, $n = 593$) groups. The samples were selected from 9205 of the total population of secondary school students, in Bone Regency, South Sulawesi Indonesia. The three-way ANOVA results showed a significant difference ($p < 0.05$) in reading and listening subskills between the groups. Highly significant results of *Madrasah* students in reading and listening subskills indicate they are better at constructing what text means in a variety of contexts, as a literary experience in reading texts and obtaining general and specific information from listening tests compared to those attending secular schools. Poor performance of boys and students who enrolled in public secular schools may become the main explanation for achievement gaps across the groups. The main and interaction effects of the school system, sectors, and gender on the tested subskills were also explained in this study. Additionally, the result of the DIF test confirmed that the equity of the tested items between them was supported.

Keywords: *Madrasah*, Secular, Islamic, English, Reading, Listening

Introduction

The majority of Asian countries including Indonesia have reformed their English language curriculum, including reading and listening literacy since it is part of the economic competitiveness that is shaping the world (Pajarwati et al. 2021; Isadaud et al. 2022). This claim has been highlighted by the Organization for Economic Co-operation and Development (OECD) about the prominent roles of foreign languages (e.g., English skills) today. According to the OECD (2021), a foreign language is not solely used as a tool of communication but is in fact developed for the purposes of cross-cultural

understanding, economic growth, and cognitive thinking. Learning a foreign language is not only for interacting with people from other countries, it is also for understanding and developing their cultural awareness and cross-cultural communicative skills (Porto et al. 2018). Moreover, English receptive skills such as reading and listening competencies have become essential in the workplace to help with economic progress. People with better English reading and listening expertise are more likely to be employed as they are considered to possess superior communication skills in cooperating and negotiating with their work colleagues in more than one country (OECD 2021; Araújo et al. 2015; Mohamed et al. 2014; Longweni and Kroon 2018). Simultaneously, reading and listening skills enhance metalinguistic understanding and critical thinking. People with better reading (Mart 2012; Mermelstein 2015) and listening (Ahmadi 2016; Leong and Ahmadi 2017; Bozorgian 2012) tend to do better in tasks such as writing and speaking. Similarly, those performing better in reading (Whitten, Labby, and Sullivan 2016; Duru and Koklu 2011) and listening (Zhang et al. 2017; Arthur et al. 2017) skills are associated with high critical thinking and problem-solving. Therefore, the roles of English skills, such as reading and listening competencies are now recognized as life-long learning skills which apply to many domains.

Over the years, theories regarding reading and listening skills have taken various forms. In 1968, Davis initially defined reading comprehension as the ability to critically understand written text, emphasizing aspects like text meaning, drawing conclusions, recognizing writing techniques, discerning mood, and answering questions. Grabe (1991) presented a technical definition of reading that included recognition skills, vocabulary and text knowledge, content comprehension, and evaluation abilities. These concepts align with Keenan et al. (2008) notion that reading comprehension is a holistic process involving the interaction between passage meaning, emphasizing understanding the entire text rather than individual words and sentences. Similarly, listening skills, historically defined as the ability to comprehend spoken language (Dirven and Taylor 1984), have been further developed in second language studies. Linguists like Bowen et al. (1985) describe listening comprehension as a process involving speech comprehension, recognition, and perception, explaining it as receiving and understanding spoken language, including sound recognition and message comprehension. These ideas correspond with the definition of listening skills as a cognitive, metacognitive, and socio-affective process (Bingol et al. 2014). The cognitive process decodes incoming information for memory, metacognitive skills assist in recognizing aspects of oral input, such as planning and evaluation, and socio-affective skills involve cooperation and reduced anxiety during listening, all influenced by factors like language and prior knowledge.

Due to the development of the essential roles of reading and listening skills today, several scholars have developed and emphasised reading and listening literacy into several subskills. For example, the OECD (2019) particularly highlights the prominent subskills in reading assessment. Firstly, locating information involves comprehension skills to get the main ideas and reflect on the entire text. It draws on the reader's understanding of what the text demands; the text organizes knowledge and evaluates the relevance of the passage. Secondly, text understanding is seen by the reader as the construction of understanding the meaning conveyed by the text. Specifically, this skill is based on the core process of attaining a representation of the literal meaning of the passage and

constructing an integrated text with prior knowledge through mapping and inference. Thirdly, text evaluation and reflection require readers to assess the quality of the information in the passage and reflect on the writing style. This process enables the readers to make justifications, draw their interpretations, and evaluate their understanding of the texts. Overall, these comprehension processes acknowledge the goal-driven, critical, and intertextual nature of reading skills and practices. Readers are required to construct what text means in a variety of contexts and for numerous reasons as a literary experience.

Similar to reading skills, listening competency is not just understanding the spoken language, it also involves some language process and learning acquisition: the ability to get a general idea, specific information, and every detail, and to make inferences (Solak and Erdem 2016). Three abilities are key, and they are as follows. First is the ability to get a general idea or listen to the gist involves general thematic understanding, without focusing on detailed information. Listeners are only expected to understand the main idea of the speaker or general information rather than comprehend the entire text. Second is the ability to obtain a specific piece of information or listen for specific information requires the listeners to discover one piece of information uttered by the speaker. It involves a listening process to establish whether the information is stated or not; thus, they should have some idea before and during the listening process. Third is the ability to comprehend every detail and understand how listeners feel or hear; here the inference focuses on a specific kind of information from the speaker. Listeners are expected to narrow down and get the details they need and ignore anything which does not sound relevant. Simultaneously, they are supposed to extract information that is not explained and any unfamiliar meaning that appears in the listening material. Those subskills emphasize the combination of knowledge, skill, and prior knowledge of listeners.

With the extended roles and specific subskills in English competencies established and needed today, concerns in exploring possible factors affecting students' English reading and listening skills differently have grown markedly in many countries. For example, in Indonesia, the disparity in English performance between and within different school systems and contexts has been noted as the main problem. This issue has been recognized by Newhouse and Beegle (2006) who examine the effect of school types on student performance in Indonesia, highlighting a major difference in what secular or non-Islamic and *Madrasah* or Islamic school students achieved. Using the national examination data, the study revealed that public secular schools did better than private secular and Islamic school students in three subjects including the English test. This finding has been supported by Hendajany (2016) who provides evidence of disparities, showing that private secular schools were superior in performance compared to those attending private Islamic institutions. However, an investigation by Asadullah, Chaudhury, and Dar (2007) tends to slightly discard the early findings. Their study comparing the religious and secular secondary schools in Bangladesh concludes that even though no significant difference was noted between secular and *Madrasah* schools, students who attend Islamic schools tended to perform worse compared to those in non-*Madrasah* schools.

In this case, several studies have reported some explanations leading to poor school performance in Islamic schools. For instance, Stern and Smith (2016) suggest that school funding and resources have become the main issues in the poor performance of students

from Indonesian Islamic schools (mostly in the private sector). Most secular schools operate under public or government authority, they receive consistent government funding and most of their teachers are civil-servant teachers who are paid standard wages or higher by the government. Contrarily, the majority of teachers in *Madrasah* schools are non-permanent teachers; they receive low salaries which hinge on the availability of funding subsidies from the government given to these schools. This claim is confirmed in several investigations which have acknowledged that low standards of resources, such as facilities and learning materials (Ali et al. 2011; Ependi 2020), low-paid teachers (Muhajir 2016; ADB 2014) as well as untrained teachers (Kholis and Murwanti 2019; ADB 2014) are still problematic in Islamic schools in Indonesia. This can undermine teaching effectiveness and student performance. Similarly, different investigations looking at the same issues have noted that lack of school funding, poor school infrastructure, and unqualified teachers in other countries, such as England (Ameli et al. 2006) and the Philippines (Lamla 2018), have become serious problems in Muslim or *Madrasah* schools. For this reason, Islamic schools tend to struggle to deliver a high-quality education compared to secular schools.

Furthermore, several investigations (Ali et al. 2011; Muttaqin et al. 2019) conducted in Indonesian *Madrasah* schools have noted that evidence of the discrepancy in learning achievement has existed in all school sectors. The studies revealed that students from public Islamic schools obtained high scores in English which indicates they did better in English learning compared to students attending private *Madrasah* schools. Consistent with prior literature, the studies acknowledge that the advantages of public *Madrasah* schools as government-funded entities enable them to have better resources and outcomes. This claim is echoed in the study by Asadullah, Chaudhury, and Dar (2007) who sampled secondary school students in Bangladesh, confirming that Islamic schools in the private sector did poorly in language subjects compared to non-Islamic schools. Conversely, a conflicting result suggests that better English scores, including reading, were recorded in public schools and independent schools in non-religious contexts (Magulod 2017). However, other analyses revealed a different trend and contended the type of school did not influence students' English skills as far as secular education was concerned (Nyarko et al. 2018; Berends and Waddington 2018; Eng, Mohamed, and Javed 2013).

The study by Ali et al. (2011) also discovered that student diversity, such as gender, was found to differ in English achievement in Indonesian *Madrasah* schools. Female students achieved higher scores in English examination tests, including reading and listening than male students. This finding is echoed by Murtafi'ah and Putro (2020) who report that boys were more likely to be less motivated and achieved poorly compared to boys doing English in Islamic schools. This is in line with several studies conducted in secular or global school contexts. For example, a specific investigation by Mirizon, Diem, and Vianty (2018) in Indonesian schools has reaffirmed that female students performed better in English reading comprehension. Focusing on the reading subskills, the study revealed that females obtained higher scores than males in comprehension tasks, such as inferring the main idea, details, and cause and effect. In the same education context, a systematic review by Trang (2022) on gender gaps in listening skills concludes that males seem better at listening than females. Boys tend to focus more on specific ideas

in listening tests than females. This is not denied by some studies in different countries (Mulualem, Mulu, and Gebremeskal 2022; Musa et al. 2016) showing that boys outperformed girls in English. However, other research seems to reject prior findings which found that boys and girls performed similarly (Attah and Ita 2017; Akinwumi 2017; Rahman et al. 2021). Additionally, several investigations have acknowledged that students' learning motivation (Saaty 2022; Bećirović 2017) and anxiety (Al-Sohbani 2018; Hussain, Shahid, and Zaman 2011) influence how differently boys and girls perform. In the research on this topic, the reviewed evidence seems to generally confirm the discrepancy and the inconsistent results of students' English performance across the school system, school sector, and student gender. However, they tend to fail to quantify what language skills differentiate their overall English performance across the groups.

Although the disparities in language tests have become a growing issue between secular and Islamic schools, public and private schools, as well as male and female students, the prior investigations primarily focused on general English literacy, while some research empirically examined separate school settings and was concerned with the possible external factors affecting the disparities. Published investigations on what language subskills differentiate their overall English language achievement between them remain scarce. Therefore, this study aims to address the issues by offering the following research questions:

1) What are the differences in reading and listening subskills between the students attending secular and Islamic schools as well as private and public?
2) How do the school system and school sector interact with female and male students' reading and listening performances differently?

As the country implementing both a secular and Islamic education system, this current investigation addresses the gaps in our knowledge by conducting a comparative investigation between secular or *Sekolah* and Islamic or *Madrasah* schools in Indonesia. Using the same cognitive tests, i.e., English reading and listening tests, this study aims to provide strong evidence for the presence of discrepancies in reading and listening subskills, such as in locating information, understanding ideas and information, evaluating text content and textual element in the reading test as well as listening for gist or general, specific information and detail in the listening test between students enrolled in *Sekolah* and *Madrasah* schools. Simultaneously, this study explains the discrepancy in the tested skills across the school sector and student gender as well as the interaction effects of the school system, sector, and student gender on their reading and listening subskills. To support reliable and meaningful cross-group comparison, the fairness of the measurement tools must be established to guarantee that these tools were used in the same way for both groups.

Study context: schooling system in Indonesia
This study was conducted in Indonesia which has a dual system of secular and Islamic education managed by two separate governments. Secular or non-Islamic schools are governed by the Ministry of Education and Culture (MoEC), while Islamic schools are under the Ministry of Religious Affairs (MoRA). The dualistic system is historically due

to the Muslim and secular nationalists' reactions and political considerations in 1945 once independence was achieved, concerning the character of education for national and religious reasons (Sirozi 2004). The majority (84%) of Indonesian schools are secular in nature, while a small portion (16%) are Islamic schools. According to the national education system number 20 in 2003, all schools in Indonesia, including public and private schools under MoEC and MoRA operate with the same regulations and policies. As an example, both secular and Islamic schools have the same schooling levels—basic (*Sekolah Dasar/*SD and *Madrasah Ibtidaiyah*/MI), lower secondary (*Sekolah Menengah Pertama/SMP* and *Madrasah Tsanawiyah/MTs*), upper secondary (*Sekolah Menengah Atas/SMA* and *Madrasah Aliyah/MA*), and university/college on to Islamic university/college as the higher education level (MoEC 2017a). Additionally, both secular and Islamic schools adhere to curriculum guidelines set forth by the Ministry of Education and Culture (MoEC) and are obligated to align with specific educational standards encompassing moral, cognitive, affective, and psychometric developmental domains. Beginning in the 2000s, the central government, in cooperation with the MoEC and the Ministry of Religious Affairs (MoRA), has initiated efforts to devolve their control over education policy, while preserving the foundational framework of the national education system. This shift aims to enhance the delivery, effectiveness, quality, and equity of education across diverse school types and geographical areas within Indonesia. However, as explained earlier, several studies have identified a discrepancy in school resources (Ali et al. 2011; Stern and Smith 2016; ADB 2014), such as school funds, facilities, and teacher quality between secular and Islamic schools which leads to different student outcomes. As most Islamic or *Madrasash* schools are managed by non-government authorities, they receive insufficient financial support from the government, while secular schools receive consistent government funding. In secular schools, they have more access to school facilities while the availability and quality of school facilities in Islamic schools remain problematic. Additionally, most secular school teachers are civil-servant teachers who are paid standard wages or higher by the government and have more access to participate in teacher training. In contrast, low-paid and untrained teachers are mostly found in *Madrasah* schools which struggle to deliver a high-quality education compared to secular schools. For this reason, an investigation which addresses the issues aligned with student diversity is strongly suggested.

Methodology

Participants

In this study, the population of interest encompassed secondary schools in Bone Regency, South Sulawesi, Indonesia, which totalled about 84 schools (36 secular and 48 Islamic schools) and 16,021 with 9205 students from secular and 6816 Islamic/*Madrasah* students. To construct a representative sample, a two-stage stratified sampling design by categorizing them into similar groups and randomly choosing from separate strata (Cohen, Manion, and Morrison 2002; Mills and Gay 2016). The stratification procedure concerned multilevel phases at the district and school levels. In the first phase, 12 districts were chosen based on the probability of each district comprising at least one secular (*Sekolah*) and one Islamic (*Madrasah*) school. Moreover, the total number of student samples was nominated within 30 schools in the second phase. Specifically, as presented

Table 1 Distribution of student participants between the groups

| School Sector | Gender | School System | | Total |
		Sekolah (SS)	*Madrasah* (MS)	
Public School	Female	428	119	547
	Male	193	46	239
	Total	621	165	786
Private School	Female	59	279	338
	Male	46	149	195
	Total	105	428	533
Total	Female	487	398	885
	Male	239	195	434
	Total	726	593	1319

in Table 1, 726 students were from secular schools grouped into *Sekolah*, while 593 Islamic school students were clustered into the *Madrasah* group. In the *Sekolah* group, 621 students enrolled in public schools, while 105 secular students were in private institutions. In contrast, the majority of *Madrasah* students ($n = 428$) were administered in private schools and only 165 students were in public Islamic schools. Simultaneously, about 487 students in *Sekolah* and 398 students in *Madrasah* are females, while 239 (SS) and 195 (MS) are males.

By dividing the population into distinct subgroups or strata, this sampling technique can enhance the accuracy, representativeness, and generalizability of research findings (Mills and Gay 2016; Ross 2005). More specifically, this approach ensures that each subgroup within the population is adequately represented in the sample. It also can improve the reliability of the research findings by addressing the potential biases and providing a precise reflection of the entire population. Likewise, stratification allows for better insights into specific subgroups and enables more meaningful comparisons which lead to more robust and trustworthy conclusions. For this reason, the use of a multi-stratified sample design used in this study is to ensure the adequate representation of secular and Islamic school students as the target population and to offer meaningful research outcomes.

Measures: reading and listening tests

The student's achievements—English reading and listening proficiency—were measured using the standardized English National Test developed by the Department of National Standard Education of the Ministry of Education and Culture of Indonesia (MoEC 2017b). The multiple choice test consisted of 20 items, i.e., 10 items of reading and 10 items of listening were selected and as part of the item analysis, student age, grade, experience, task requirement, and content were taken into account (Cohen, Manion, and Morrison 2002). As shown in Table 2, 10 items of reading proficiency covered three reading subskills (OECD 2019), these being: READ01 or locating information (reading items 1, 4, and 7); READ02 or understanding the ideas and information (reading items 2, 6, 9, and 10); and READ03 or evaluating the text content and textual elements (reading items 3, 5, and 8). Simultaneously, 10 items in the listening test were classified into three listening subskills (Solak and Erdem 2016), namely: LIST01 or listening for gist (listening

Table 2 English reading and listening subskills and items

Skill	Subskills	Item
Reading	Locating information (READ01)	READ1, READ4, READ7
	Understanding the ideas and information (READ02)	READ2, READ6, READ9, READ10
	Evaluating the text content and textual elements (READ03)	READ3, READ5, READ8
Listening	Listening for gist (LIST01)	LIST1, LIST7, LIST9
	Listening for specific information (LIST02)	LIST5, LIST6, LIST8
	Listening in detail (LIST03)	LIST2, LIST3, LIST4, LIST10

items 1, 7, and 9); LIST02 or listening for specific information (listening items 5, 6, and 8); and LIST03 or listening in detail (listening items 2, 3, 4, and 10). All items within the reading and listening scenarios were measured using multidimensional item analysis, transformed into six derived weighted likelihood estimate (WLE) scores through Rash analysis to reduce or remove any scoring bias (Warm 1989) and identified as dependent or outcome variables in this study. In addition, three categorical variables of the school system (SCSYSTM, 0 = Secular, 1 = Islamic), school sector (SCSECTOR, 0 = Public, 1 = Private), and student gender (GENDER, 0 Female, 1 = Male) were recognized as independent variables in this study. These variables were used to compare the students attending secular and Islamic/*Madrasah* schools, enrolled in the public/government and private/non-government schools as well as female and male students regarding their English reading and listening subskills.

Methods of analysis

Item validity: Rasch measurement model (RMM)

The Rasch measurement model (RMM) is generally employed to measure how well the test items are distributed regarding the test-takers' ability (Bond and Fox 2015). This analysis explicitly enables researchers to use the participants' scores or responses to measure their performance on a linear scale that accounts for the unequal difficulties between the test items. For this reason, RMM is important as it provides an estimate of the difficulty of the item according to the frequency of the sample's response to the measured items. In this study, Rasch techniques including differential item functioning (DIF) and multidimensional analysis of dichotomously scored items using the Conquest software were carried out. The differential item functioning technique confirmed the fairness and equity of the test item between the compared groups (Bond and Fox 2015; T. Brown and Bonsaksen 2019). Looking at the level of difficulty concerning the element between the *Sekolah* and *Madrasah* groups for all 20 measured items, this analysis makes it possible to determine whether the tested items work the same for both groups. Furthermore, the multidimensional analysis consisted of a subset of items measured as a single latent dimension (Adams et al. 2017), i.e., 10 items of reading and 10 items of listening measured into six dimensions (see the previous section).

Fit statistics indices served to determine whether the items fit the expected Rasch model. Following the suggestion made in other research (Alagumalai et al. 2005; Wu et al. 2016; Bond and Fox 2015), the fit of the tested items was established based on item logit (expected value = 1), discrimination (±2), and item differentiation for DIF

Nawas *et al. Language Testing in Asia*   (2023) 13:52

Page 9 of 23

analysis (0.5). For their mean square scores/MNSQ, this study adopts the acceptable range between 0.8 and 1.2 (Wright and Linacre 1994). The items with their infit MNSQ values which are greater than 1.00 specify the underfit model with large residuals. In contrast, the values which are less than 1.00 indicate an overfit model and their residuals are smaller than expected. Moreover, the items with positive logit scales denote difficult items while the negative logit scores mean that the items can be endorsed. The tested items whose item discrimination was greater than 0.2 are specified as good items, while less than 0.2 designates them as misleading items. Additionally, t-statistics values which are less than −2 and greater than 2 indicate unacceptable values, but the studies also argue that t-statistics values are sensitive to the sample size. For the acceptable item differentiation in DIF analysis, a value of ±0.5 means that the items work in the same way for both groups. A study by Dorans and Holland (1992) argues that item differentiation values greater than 0.5 are still acceptable. In this case, misfitting items are typically acknowledged and removed from the model. As well, more focus is given here to the acceptable MNSQ values and item discrimination since the items fit with the Rasch model. In addition, before performing the DIF analysis, item fit analysis needs to be undertaken to ensure the tests (English reading and listening) function properly and confirm the quality and validity of measurement instruments. This analysis assesses the alignment between the individual item tests and the underlying measurement model, ensuring the tests effectively contribute to accurate measurement.

Three-way analysis of variance (ANOVA)

In this study, a series of comparative analyses using SPSS software was conducted. First, a descriptive statistical analysis was undertaken using the exploration method to compare central tendencies of the observed measures (WLE), such as reading and listening subskills between the students from secular and Islamic schools and those from different school sectors and gender within the groups. Moreover, a three-way analysis of variance (ANOVA) was undertaken to determine if there is an interaction effect between three predictors—school system, school sectors, and student gender—on reading and listening subskills as the outcome variables. The significance of the mean differences between the groups is according to their *p*-value of 0.05. The interaction effects hold a unique significance in understanding the complex relationships between multiple independent variables (Pallant 2016; Jaccard 1998). However, Field (2013) suggests that when significant interaction effects are observed, interpreting the main effects in such a context often leads to ambiguity. By more focusing on the interaction effects than the main effects, this study can gain a deeper and more nuanced understanding of the factors influencing their dependent variables, leading to more robust and informed research findings.

In addition, before conducting the three-way ANOVA, the initial tests of normality and homogeneity of variance as assumptions of the tests are conducted to determine whether the data used follows a normal distribution and to assess whether the variability of the dependent variable is approximately constant across different levels of independent variables. In this study, the normality of the data is assessed using 2 and ±10 for its skewness and kurtosis (Griffin and Steinbrecher 2013; T.A. Brown 2015), while the homogeneity of variance assumption is based on Levene's test results of significant value less than 0.5 suggesting the variance of independent variables across the groups is not

equal (Pallant 2016). However, a study by Pallant (2016) points out that the main output of the ANOVA test is the results of tests of between-subjects effects which explain the main and interaction effects of the tested variables.

Results

Rasch modelling: item analysis and differential item functioning (DIF)

Before performing the differential item functioning (DIF) tests, the initial run of the fit analysis shown in Table 3 was executed to examine how well the reading and listening items are distributed regarding the level of the test-takers. The results of 20 items of reading and listening indicate that the items are acceptable. This is evident with the item discrimination revealed of greater than 0.2 and the INFIT MNSQ are within the 0.8–1.2 range, signifying that the tested items fit the Rasch model well. Furthermore, the DIF was undertaken to assess the fairness of the test items as applied to the *Sekolah* (SS) and *Madrasah* (MS) groups. As documented in Table 4, similar results for the 20 items of reading and listening indicating the acceptable values of item discrimination (>0.2) and INFIT MNSQ (0.8–1.2) are listed for both

Table 3 Results of item fit analysis of reading and listening subskills

	Item	Estimate	S. E	Infit			Item Delta	Item Discr.
				MNSQ	CI	t		
READING	Dimension 1							
	READ1	-0.028	0.04	1.00	(0.96, 1.04)	0.20	-0.03	0.44
	READ4	-0.054	0.04	1.04	(0.96, 1.04)	2.10	-0.05	0.35
	READ7	0.082[a]	0.06	0.96	(0.96, 1.04)	-1.70	0.08	0.52
	Dimension 2							
	READ2	-0.181	0.04	1.00	(0.96, 1.04)	0.00	-0.18	0.45
	READ6	0.527	0.04	1.00	(0.94, 1.06)	-0.10	0.53	0.37
	READ9	-0.256	0.04	1.01	(0.97, 1.03)	0.30	-0.26	0.37
	READ10	-0.091[a]	0.07	1.00	(0.96, 1.04)	0.20	-0.09	0.35
	Dimension 3							
	READ3	0.544	0.04	1.03	(0.95, 1.05)	1.30	0.54	0.39
	READ5	-0.204	0.04	1.01	(0.97, 1.03)	0.70	-0.20	0.39
	READ8	-0.340[a]	0.06	0.98	(0.97, 1.03)	-1.30	-0.34	0.50
LISTENING	Dimension 1							
	LIST1	-0.806	0.04	1.00	(0.96, 1.04)	0.10	-0.81	0.42
	LIST7	0.292	0.04	1.00	(0.96, 1.04)	0.00	0.29	0.34
	LIST9	0.514[a]	0.06	1.00	(0.96, 1.04)	0.10	0.51	0.34
	Dimension 2							
	LIST5	0.188	0.04	0.93	(0.95, 1.05)	-3.10	0.19	0.50
	LIST6	-0.501	0.04	0.99	(0.96, 1.04)	-0.80	-0.50	0.43
	LIST8	0.313[a]	0.06	1.03	(0.95, 1.05)	1.10	0.31	0.33
	Dimension 3							
	LIST2	0.005	0.04	0.95	(0.96, 1.04)	-2.40	0.00	0.48
	LIST3	0.167	0.04	1.01	(0.96, 1.04)	0.40	0.17	0.39
	LIST4	-0.691	0.04	0.97	(0.96, 1.04)	-1.80	-0.69	0.48
	LIST10	0.520[a]	0.08	1.03	(0.95, 1.05)	1.20	0.52	0.33

Chi-square test of parameter equality = 1854.05, *df* = 19, Sig Level = 0.000

[a] Constraint

Table 4 Reading and listening item fit differences between groups

	Item	Group	Estimate	S. E	Infit			Item Delta	Item Discr.	Item Diff.
					MNSQ	CI	*t*			
Reading	Dimension 1—Locating information									
	READ1	SS	-0.083	0.041	1.00	(0.95, 1.05)	-0.2	0.04	0.39	0.17
		MS	0.083[a]	0.041	1.02	(0.95, 1.05)	0.9	0.04	0.33	
	READ4	SS	0.185	0.042	1.04	(0.94, 1.06)	1.3	0.27	0.23	0.37
		MS	-0.185[a]	0.042	1.01	(0.96, 1.04)	0.6	-0.27	0.21	
	READ7	SS	0.112	0.042	0.93	(0.94, 1.06)	-2.3	0.33	0.44	0.22
		MS	-0.112[a]	0.042	0.98	(0.95, 1.05)	-0.7	-0.07	0.30	
	Dimension 2—Understanding the ideas and information									
	READ2	SS	-0.067	0.041	0.95	(0.95, 1.05)	-2.1	-0.04	0.46	0.13
		MS	0.067[a]	0.041	1.00	(0.95, 1.05)	-0.2	-0.08	0.36	
	READ6	SS	-0.229	0.043	0.96	(0.93, 1.07)	-1.1	0.51	0.34	0.46
		MS	0.229[a]	0.043	1.07	(0.91, 1.09)	1.4	0.80	0.26	
	READ9	SS	-0.088	0.041	0.98	(0.95, 1.05)	-0.8	-0.14	0.32	0.18
		MS	0.088[a]	0.041	1.07	(0.96, 1.04)	3.1	-0.14	0.21	
	READ10	SS	-0.006	0.125	0.98	(0.95, 1.05)	-0.6	0.10	0.34	0.12
		MS	0.006[a]	0.125	1.05	(0.95, 1.05)	2.2	-0.06	0.24	
	Dimension 3—Evaluating the text content and textual elements									
	READ5	SS	-0.003	0.041	1.02	(0.96, 1.04)	1.0	-0.32	0.40	0.01
		MS	0.003[a]	0.041	0.98	(0.96, 1.04)	-1.1	-0.49	0.26	
	READ3	SS	0.171	0.042	1.05	(0.92, 1.08)	1.3	0.57	0.31	0.34
		MS	-0.171[a]	0.042	1.01	(0.95, 1.05)	0.5	0.05	0.28	
	READ8	SS	0.009	0.041	0.95	(0.96, 1.04)	-2.3	-0.45	0.43	0.02
		MS	-0.009[a]	0.041	0.98	(0.96, 1.04)	-1.0	-0.64	0.30	
Listening	Dimension 1—Listening for gist or getting general idea									
	LIST1	SS	-0.022	0.043	0.95	(0.95, 1.05)	-2.0	-0.95	0.40	0.04
		MS	0.022[a]	0.043	0.99	(0.94, 1.06)	-0.3	-1.08	0.31	
	LIST7	SS	-0.090	0.043	1.05	(0.95, 1.05)	1.9	0.12	0.32	0.18
		MS	0.090[a]	0.043	1.00	(0.94, 1.06)	-0.1	0.13	0.22	
	LIST9	SS	0.326	0.044	1.05	(0.91, 1.09)	1.2	0.75	0.24	0.65
		MS	-0.326[a]	0.044	1.02	(0.95, 1.05)	0.9	-0.08	0.22	
	Dimension 2— Listening for specific information									
	LIST5	SS	-0.015	0.044	0.98	(0.93, 1.07)	-0.6	0.40	0.42	0.03
		MS	0.015[a]	0.044	0.97	(0.94, 1.06)	-1.1	0.25	0.35	
	LIST6	SS	-0.066	0.043	0.99	(0.96, 1.04)	-0.5	-0.31	0.33	0.13
		MS	0.066[a]	0.043	0.97	(0.96, 1.04)	-1.3	-0.35	0.35	
	LIST8	SS	0.114	0.044	1.03	(0.92, 1.08)	0.8	0.64	0.26	0.23
		MS	-0.114[a]	0.044	1.01	(0.94, 1.06)	0.3	0.24	0.27	
	Dimension 3—Listening in detail									
	LIST2	SS	-0.158	0.043	0.95	(0.95, 1.05)	-2.2	-0.03	0.45	0.32
		MS	0.158[a]	0.043	0.98	(0.94, 1.06)	-0.8	0.11	0.38	
	LIST3	SS	-0.209	0.044	1.07	(0.95, 1.05)	2.6	0.08	0.25	0.42
		MS	0.209[a]	0.044	0.95	(0.93, 1.07)	-1.4	0.33	0.31	
	LIST4	SS	-0.179	0.043	0.97	(0.96, 1.04)	-1.5	-0.73	0.40	0.36
		MS	0.179[a]	0.043	0.98	(0.96, 1.04)	-1.1	-0.54	0.35	
	LIST10	SS	0.300	0.131	1.08	(0.90, 1.10)	1.6	0.91	0.24	0.60
		MS	-0.300[a]	0.131	1.03	(0.94, 1.06)	1.1	0.13	0.24	

Chi-square test of parameter equality = 90.57, *df* = 18, Sig Level = 0.000

[a] Constraint

respondent groups. Two items of READ2 and READ7 for the MS group and the item of READ1 for the SS group have INFIT MNSQ values of 1.00, which is the expected value of the infit mean square. The INFIT MNSQ values of other items for both groups range between 0.95 and 1.08 and they are close to 1.00. The items with infit values greater than 1.00 indicate an underfit model whose residuals are larger than expected. The overfit model, in contrast, is revealed from those items with infit values lower than 1.00 and has low residuals which are exposed. Similarly, the estimate and standard error of measurement for the items are summarized in Table 4 which presents the position of the logit scale. As 0 (zero) is the average value for the difficulty level of the tested items, this shows that most of the items are close to the average estimate. Positive logit values of READ3, READ4, READ7, and READ8 for the SS group indicate that those reading test items are more difficult for the *Sekolah* students than the *Madrasah* group. More difficult items, positive logit values, are revealed for the *Madrasah* group in the listening test except for items LIST8, LIST9, and LIST10, which indicate that the other seven items in the listening test are easier for *Sekolah* students.

Three-way analysis of variance (ANOVA)

Main effects of school system (SCSYTM), school sector (SCSECTOR) and student gender (GENDER)

Tables 5 and 6 display differences in students' reading and listening skills based on the type of school system (SCSYTM). In the reading test, significant distinctions emerged in locating information (READ01, $p < 0.05$), understanding ideas and information (READ02, $p < 0.05$), and evaluating text content and elements (READ03, $p < 0.05$) between students from different school systems. Madrasah students outperformed secular school students in READ01 (Madrasah: $M = -0.20$, SD $= 1.18$; Sekolah: $M = -0.57$, SD $= 1.14$), READ02 (Madrasah: $M = -0.41$, SD $= 1.06$; Sekolah: $M = -0.53$, SD $= 1.19$), and READ03 (Madrasah: $M = -0.48$, SD $= 1.14$; Sekolah: $M = -0.53$, SD $= 1.14$), indicating their better skills in locating information, comprehending ideas, and evaluating text in reading tests (see Fig. 1a). In listening skills, differences across SCSYTM were observed only in listening for the main idea (LIST01, $p < 0.05$) and listening for specific information (LIST02, $p < 0.05$). Sekolah students scored lower in LIST01 (Sekolah: $M = -0.29$, SD $= 1.15$; Madrasah: $M = -0.15$, SD $= 1.19$) and LIST02 (Sekolah: $M = -0.53$, SD $= 1.14$; Madrasah: $M = -0.48$, SD $= 1.14$) compared to Madrasah students (see Fig. 1d), indicating poorer performance among secular-school students in grasping the main idea and specific details during listening tests. No significant difference ($p > 0.05$) was found in LIST03, suggesting that Sekolah and Madrasah students performed similarly in listening tests when it came to detailed listening skills.

Furthermore, as shown in the tables above, only LIST01 ($p < 0.05$) and LIST02 ($p < 0.05$) display significant differences among students from different school systems. Public school students excelled in LIST01 (Public: $M = -0.19$, SD $= 1.14$; Private: $M = -0.28$, SD $= 1.21$) and LIST02 (Public: $M = -0.46$, SD $= 1.20$; Private: $M = -0.58$,

Table 5 Results of Three-way ANOVA concerning the Effects of School System, School Sector and Gender on Reading Subskills

Tests of Between-Subjects Effects

Source	Dependent variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected model	READ01	86.038[a]	7	12.291	9.330	<.001	.047
	READ02	62.974[b]	7	8.996	7.221	<.001	.037
	READ03	123.036[c]	7	17.577	13.490	<.001	.067
Intercept	READ01	132.884	1	132.884	100.865	<.001	.071
	READ02	191.580	1	191.580	153.778	<.001	.105
	READ03	46.771	1	46.771	35.897	<.001	.027
SCSYTM	READ01	31.848	1	31.848	24.174	<.001	.018
	READ02	9.474	1	9.474	7.604	.006	.006
	READ03	36.177	1	36.177	27.766	<.001	.021
SCSECTOR	READ01	3.775	1	3.775	2.865	.091	.002
	READ02	.034	1	.034	.028	.868	.000
	READ03	2.756	1	2.756	2.116	.146	.002
GENDER	READ01	.074	1	.074	.056	.813	.000
	READ02	12.841	1	12.841	10.307	.001	.008
	READ03	13.081	1	13.081	10.040	.002	.008
SCSYTM * SCSECTOR	READ01	2.758	1	2.758	2.094	.148	.002
	READ02	.477	1	.477	.383	.536	.000
	READ03	4.358	1	4.358	3.345	.068	.003
SCSYTM * GENDER	READ01	8.425	1	8.425	6.395	.012	.005
	READ02	14.787	1	14.787	11.869	<.001	.009
	READ03	13.518	1	13.518	10.375	.001	.008
SCSECTOR * GENDER	READ01	5.310	1	5.310	4.031	.045	.003
	READ02	.073	1	.073	.059	.808	.000
	READ03	5.052	1	5.052	3.877	.049	.003
SCSYTM * SCSECTOR * GENDER	READ01	.000	1	.000	.000	.993	.000
	READ02	1.476	1	1.476	1.185	.277	.001
	READ03	.012	1	.012	.009	.924	.000
Error	READ01	1727.170	1311	1.317			
	READ02	1633.271	1311	1.246			
	READ03	1708.123	1311	1.303			
Total	READ01	2029.680	1319				
	READ02	1995.452	1319				
	READ03	1874.220	1319				
Corrected Total	READ01	1813.208	1318				
	READ02	1696.245	1318				
	READ03	1831.159	1318				

[a] R Squared = .047 (Adjusted R Squared = .042)

[b] R Squared = .037 (Adjusted R Squared = .032)

[c] R Squared = .067 (Adjusted R Squared = .062)

Note: * = Interaction Effect

SD = 1.17) compared to their private school counterparts (see Fig. 1e). This suggests that private school students performed less well in understanding the main idea and specific information during listening tests than public school students. Additionally,

Nawas *et al. Language Testing in Asia*   (2023) 13:52

Page 14 of 23

Table 6 Results of Three-way ANOVA concerning the Effects of School System, School Sector and Gender on Listening Subskills

Tests of between-subjects effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected model	LIST01	37.528[a]	7	5.361	4.005	<.001	.021
	LIST02	30.977[b]	7	4.425	3.179	.002	.017
	LIST03	15.386[c]	7	2.198	1.594	.133	.008
Intercept	LIST01	70.577	1	70.577	52.723	<.001	.039
	LIST02	261.931	1	261.931	188.157	<.001	.126
	LIST03	149.278	1	149.278	108.264	<.001	.076
SCSYTM	LIST01	14.582	1	14.582	10.893	<.001	.008
	LIST02	7.413	1	7.413	5.325	.021	.004
	LIST03	1.376	1	1.376	.998	.318	.001
SCSECTOR	LIST01	10.758	1	10.758	8.037	.005	.006
	LIST02	8.669	1	8.669	6.227	.013	.005
	LIST03	.289	1	.289	.209	.647	.000
GENDER	LIST01	2.604	1	2.604	1.945	.163	.001
	LIST02	.590	1	.590	.424	.515	.000
	LIST03	3.686	1	3.686	2.673	.102	.002
SCSYTM * SCSECTOR	LIST01	3.087	1	3.087	2.306	.129	.002
	LIST02	16.792	1	16.792	12.062	<.001	.009
	LIST03	.635	1	.635	.461	.497	.000
SCSYTM * GENDER	LIST01	3.852	1	3.852	2.878	.090	.002
	LIST02	1.086	1	1.086	.780	.377	.001
	LIST03	.485	1	.485	.352	.553	.000
SCSECTOR * GENDER	LIST01	2.926	1	2.926	2.186	.140	.002
	LIST02	2.911	1	2.911	2.091	.148	.002
	LIST03	1.720	1	1.720	1.247	.264	.001
SCSYTM * SCSECTOR * GENDER	LIST01	.919	1	.919	.686	.408	.001
	LIST02	1.077	1	1.077	.774	.379	.001
	LIST03	.894	1	.894	.649	.421	.000
Error	LIST01	1754.950	1311	1.339			
	LIST02	1825.022	1311	1.392			
	LIST03	1807.656	1311	1.379			
Total	LIST01	1861.516	1319				
	LIST02	2194.993	1319				
	LIST03	2069.623	1319				
Corrected Total	LIST01	1792.479	1318				
	LIST02	1855.999	1318				
	LIST03	1823.042	1318				

[a] R Squared = .021 (Adjusted R Squared = .016)

[b] R Squared = .017 (Adjusted R Squared = .011)

[c] R Squared = .008 (Adjusted R Squared = .003)

Note: * = Interaction Effect

there were no significant achievement gaps ($p > 0.05$) in LIST03, as well as all three reading subskills shown in Fig. 1b, indicating that students in public and private schools perform similarly when it comes to detailed comprehension in listening tests
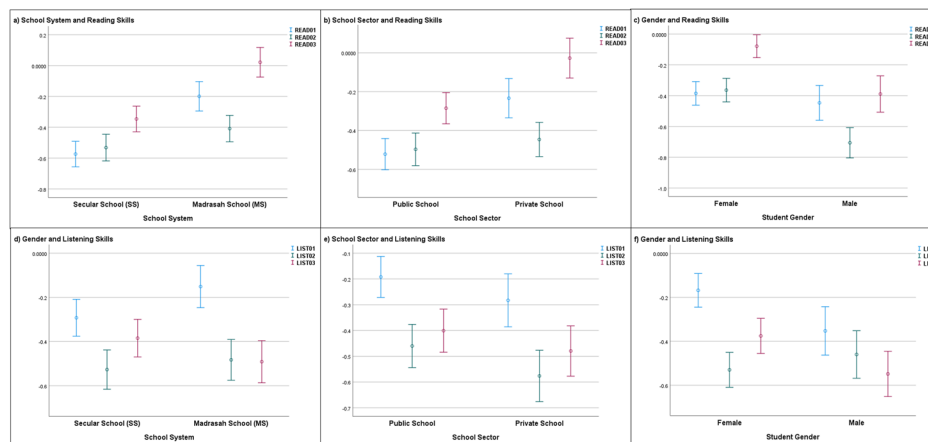
Fig. 1 The main effects of school system, school sector, and student gender on reading and listening subskills

and finding information, understanding ideas, and critiquing text content and elements in reading assessments. Regarding students' reading and listening performance based on gender (GENDER), significant differences were observed in READ02 ($p < 0.05$) and READ03 ($p < 0.05$). As depicted in Fig. 1c, female students scored higher in READ02 (Female: $M = -0.36$, SD = 1.16; Male: $M = -0.71$, SD = 1.04) and READ03 (Female: $M = -0.08$, SD = 1.13; Male: $M = -0.39$, SD = 1.25) compared to male students. This indicates that males tend to struggle with reading skills, especially in understanding ideas and information and evaluating text content and elements. No significant differences ($p > 0.05$) were found in READ01 and all listening subskills (Fig. 1f), indicating that male and female students performed similarly in locating information in reading tests and listening subskills. As suggested by Field (2013), interpreting main effects in the presence of significant interaction effects tends to be ambiguous when interaction effects are significant. Thus, deeper and more nuanced explanations of the interaction effects of the school system, sector, and gender influencing their dependent variables, leading to more robust and informed research findings are discussed in the next section.

Interaction effects of school system, school sector and student gender on reading and listening subskills

The interaction effects of the predictors on reading and listening subskills are separately illustrated in Tables 5 and 6. In the reading tests, the significant moderation effects of the school system and gender (SCSYTM*GENDER), as well as the school sector and gender (SCSECTOR *GENDER), are revealed. More specifically, the significant interaction effects of SCSYTM*GENDER on READ01 ($F (1,1311) = 6.395$, $p \leq 0.05$), READ02 ($F (1,1311) = 11.869$, $p \leq 0.05$), and READ03 ($F (1,1311) = 10.375$, $p \leq 0.05$) indicate there were significantly different reading skills between girls and boys in the different school systems. As shown in Fig. 2a–c in secular schools, females achieved higher scores than males, while boys did better than girls in *Madrasah* schools. This suggests that female students from secular schools and males from Islamic schools did better in three subskills;
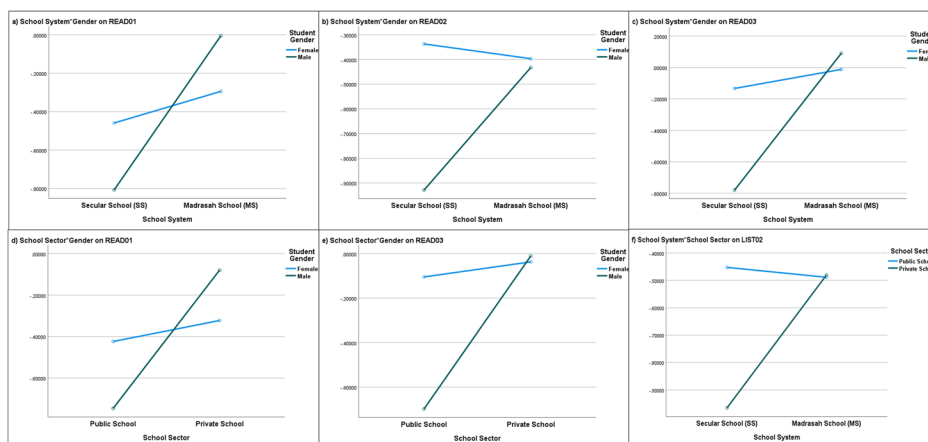
Fig. 2 The significant interaction effects among SCSYTM*GENDER on reading and listening subskills

they did better in locating information, understanding the ideas and information as well and evaluating the text content and textual elements in reading tests compared to boys in the secular group and girls in the *Madrasah* group. Regarding the reading achievement discrepancies between female and male students according to the school system, girls in *Sekolah* (READ01, $M = -0.46$, SD $= 1.18$; READ02, $M = -0.34$, SD $= 1.21$; READ03, $M = -0.13$, SD $= 1.10$) and *Madrasah* (READ01, $M = -0.29$, SD $= 1.13$; READ02, $M = -0.40$, SD $= 1.10$; READ03, $M = -0.01$, SD $= 1.16$) performed slightly similarly in reading subskills. On the other hand, the biggest differences in reading subskills are illustrated between boys in secular (READ01, $M = -0.81$, SD $= 1.01$; READ02, $M = -0.93$, SD $= 1.04$; READ03, $M = -0.78$, SD $= 1.11$) and Islamic schools (READ01, $M = 0.00$, SD $= 1.27$; READ02, $M = -0.43$, SD $= 0.98$; READ03, $M = 0.09$, SD $= 1.25$), favouring males in *Madrasah* group. The lowers scores of males in secular schools might become the key issue of the poor overall scores attained by the *Sekolah* than *Madrasah* schools.

The moderation effects of SCSECTOR*GENDER are detected in READ01 (F (1,1311) $=$ 4.031, $p \leq 0.05$) and READ03 (F (1,1311) $=$ 3.877, p ≤ 0.05) signalling the gaps revealed in those reading subskills between females and males from different school sectors. This study found that girls obtained higher scores in READ01 and READ03 than boys in public schools, while females in public schools obtained lower scores than males in private schools. The findings indicate that girls in public schools and boys in private schools obtained high scores in READ01 and READ03 (find Fig. 2d, e) signalling that males in public schools and females in private schools did not achieve well in locating information and evaluating text content also textual elements in reading assessments. Simultaneously, the results also present that female students from the public (READ01, $M = -0.42$, SD $= 1.17$; READ03, $M = -0.10$, SD $= 1.11$) and private (READ01, $M = -0.32$, SD $= 1.15$; READ03, $M = -0.04$, SD $= 1.16$) schools are more likely to perform similarly, on the contrary, boys in public (READ01, $M = -0.75$, SD $= 1.07$; READ03, $M = -0.70$, SD $= 1.14$) and private (READ01, $M = -0.08$, SD $= 1.24$; READ03, $M = -0.01$, SD $= 1.28$) sectors are shown to have biggest discrepancies in favour of male students attending private institutions.

Different from the previous findings, in listening tests, the interaction effects are only revealed between the school system and school sector (SCSTYM*SCSECTOR) on

LIST02 (F (1,1311) $= 12.062$, $p \leq 0.05$). The effect of SCSTYM*SCSECTOR indicates that there was a significant discrepancy between the secular and Islamic students across different school sectors. As documented in Fig. 2f, students attending public secular did better than those in private secular schools, while public Islamic schools achieved lower than students in private *Madrasah* schools. The results signify that the students from public secular and private *Madrasah* schools are better at listening for specific information in listening tests compared to the other groups, such as private secular and public *Madrasah* students. Additionally, it is also shown that students from public *Sekolah* (LIST02, $M = -0.45$, SD $= 1.22$) and *Madrasah* (LIST02, $M = -0.49$, SD $= 1.12$) groups tend to perform similarly. On the other hand, the biggest gap in LIST02 is shown between private secular (LIST02, $M = -0.97$, SD $= 1.15$) and Islamic (LIST02, $M = -0.48$, SD $= 1.16$) schools in favour of private *Madrasah* schools.

Overall, the results also show that adjusted R^2 for the corrected model of 0.42 for READ01, 0.32 for READ02, and 0.62 for READ03. This concludes that around 4% of the variance of the student's scores in locating information, 3% in understanding the ideas and information, and 6% in evaluating the text content and textual elements are explained by the predictors of the school system, school sector, and student gender. For the listening subskills, the adjusted R^2 results indicate that around 2% or 0.016 on the variable of student scores in listening for gist (LIST01), 1% or 0.011 in listening for specific information (LIST02), and 0.3% or 0.003 in listening in detail (LIST03) are explained by the three-way variables of SCSYTM, SCSECTOR, and GENDER. Regarding the effect size of the three-way interaction among variables on reading and listening subskills, the partial eta square (partial η2) is <0.01 which represents a small partial eta squared (Cohen et al. 2002). Additionally, the estimates within ± 3 and ± 10 for its skewness and kurtosis of the data indicate that a normal distribution is revealed. Likewise, Levelne's test was not statically significant shown in the listening and reading achievement >0.05, indicating that homogeneity of variance is evident across the groups (see Appendix).

Discussion

This paper was motivated by published findings (Newhouse and Beegle 2006; Hendajany 2016) which claim there is a discrepancy between secular and Islamic schools in English performance. However, such studies fail to identify what language skills differentiate their overall language performance. For this reason, this present research sets out to specifically prove there are disparities in English language reading and listening subskills between the students from secular and Islamic schools, public and private schools as well as male and female students in Indonesia. Simultaneously, the interaction effects of the school system, school sector, and gender on students' English skills were investigated. Before assessing the main and interaction effects of the predictors of students' reading and listening skills, the equality of measurement tools using differential item functioning (DIF) analysis was first checked to ensure they functioned equally for both groups. The results confirmed that the equity of the test items for the *Sekolah* (SS) and *Madrasah* (MS) groups was supported; it was evident by obtaining the acceptable thresholds of item fit statistics garnered from previous studies (Alagumalai et al. 2005; Wu et al. 2016; Bond and Fox 2015). Moreover, the findings verify that the implemented English proficiency tests work in the same way as designed by MoEC (2017b) on the listening and reading test for grade 12 students in

different types of schools in Indonesia. Concurrently, the multidimensional item analysis of the instrument used for the reading test adapted from the OECD (2019) and the listening test from Solak and Erdem (2016) were also relevant to this study. As a result, the three-way ANOVA analysis used here does offer some interesting findings:

Firstly, the findings of this study illuminate compelling discrepancies in the tested items as dependent variables between the school system 'secular and Islamic schools', favouring the students enrolled in Islamic schools. This study serves as a pivotal indicator, signalling that *Madrasah* students demonstrate better English proficiency, especially in the complex task of interpreting textual meanings in reading tests and discovering specific information in listening tests. The findings are aligned with the theories of reading and listening skills highlighting that students with better reading comprehension (e.g., Islamic school students) tend to have an ability to critically understand written text, emphasizing aspects like text meaning, drawing conclusions, recognizing writing techniques, discerning mood, and answering questions (Davis 1968a); they can interact with passage meaning, emphasize and understand the entire text rather than individual words and sentences (Keenan et al. 2008). With better listening literacy, the students in *Madrasah* schools did well in understanding spoken language, including sound recognition and message comprehension (Bowen et al. 1985) as well as decoding incoming information for memory in listening tests (Bingol et al. 2014). This new trend of better English achievement in Islamic schools has changed the prior tendencies and rejects the previous studies (Newhouse and Beegle 2006; Hendajany 2016) showing that secular students performed better in English than Islamic school students. Therefore, the outcomes of this study carry significant implications which extend beyond the scope of educational assessment. Likewise, this study sheds light on the distinctive pedagogical methods and approaches employed in Islamic schools, which contribute to fostering students' advanced skills in English reading and listening subjects.

Other findings of this study offer an intriguing insight into the academic performance landscape by indicating that students enrolled in public schools outperformed their counterparts attending private schools. Public school students excelled in discerning both general and specific pieces of information in listening assessments challenging preconceived notions about the superiority of private schools. Better performance achieved by students enrolled in public schools is associated with greater resources, especially government funds (Stern and Smith 2016), leading to better learning outcomes. The findings also corroborate the prior studies (Ali et al. 2011; Stern and Smith 2016; ADB 2014) highlighting that Indonesian schools managed by the government authority are beneficial with the school facilities and teacher quality. Public schools have greater access to educational resources and infrastructure, whereas the availability and quality of school facilities in private schools continue to pose challenges. The majority of educators in public schools are government-employed civil-servant teachers who receive standard or higher government wages. They also have increased opportunities to engage in teacher training programs, enabling them to provide high-quality education, resulting in enhanced learning outcomes when compared to their counterparts in private schools.

It is important to note that the specific results of the discrepancies in language achievement across the school system and sectors revealed in this study are strongly affected and moderated by the student gender. Higher scores among girls in secular schools and boys

in Islamic schools underscore their superior performance in key aspects of reading comprehension. The enhanced scores among girls in secular schools in terms of understanding the text's demands, comprehending the conveyed meaning, justifying points, and drawing conclusions reflect a nuanced interaction between gender and the educational setting. Moreover, a notable trend of underperformance among boys attending public schools and girls in Islamic schools in terms of locating information and evaluating text content during reading assessments is also shown in this study. These inconsistent results on the effect of student gender on their performance have been stated in the prior studies (Mulualem et al. 2022; Musa et al. 2016; Ali et al. 2011) generally recognizing whether boys or girls are better in language performances. Simultaneously, other studies supported that contradicting results on boys and girls across school settings, such as Islamic and non-Islamic schools are strongly affected by their learning behaviours, such as language learning motivation (Saaty 2022; Bećirović 2017) and anxiety (Al-Sohbani 2018; Hussain et al. 2011) possibly lead to different outcomes for students' English achievement. Recognizing the strengths and challenges present in each context is essential for designing targeted interventions that can address the girls' and boys' learning behaviours across different school settings through effective teaching and learning.

Altogether, the existence of poor English performance by boys in secular schools was identified as the key factor contributing to lower English achievement in secular schools than those in Islamic schools. In order to ensure fairness and parity in education for both male and female students in both secular and Islamic schools, this research yields practical results. One of these outcomes involves the pivotal role that *Sekolah* teachers play in high-quality instruction that accommodates student diversity, such as a wide range of learning needs and preferences. This encompasses the application of tailored teaching approaches, offering individualized assistance to tackle the distinct learning needs of male students, and fostering collaborative and peer-based learning endeavours that enable mutual knowledge exchange. Additionally, these efforts can boost students' enthusiasm and motivation for learning English and mitigate learning obstacles, such as English difficulty and anxiety. Consequently, it is highly recommended to formulate a comprehensive educational policy and receive governmental backing to address imbalances in educational quality based on student gender and school attributes. This could involve allocating sufficient resources, enhancing teacher quality, implementing evidence-based methodologies, and closely monitoring progress in school achievements to ensure that every Indonesian student enjoys equitable access and accomplishments. Moreover, the findings reflect practical evidence for the consistency of measurement tools using the differential item functioning (DIF) technique with Rasch analysis employed in the cross-group comparison. Unfortunately, the generalizability of these findings is limited by the scope of the research, as it only looked at the secular and Islamic education systems in Indonesia. Further studies on this topic should explore other contexts, measures, and methods and use more varieties and sizes of samples.

Conclusion

This study proves notable differences in English proficiency between secular and Islamic schools, favouring Islamic school students. It highlights the impressive language skills of *Madrasah* students, showcasing their ability to grasp intricate language

tasks like understanding text nuances and extracting precise information. This finding breaks from past trends where secular students excelled over their Islamic peers, showcasing significant progress in *Madrasah* language learning despite limited resources. Moreover, this study also noted that the performance of the students from *Sekolah* and *Madrasah* schools varied depending on school sectors and gender. Girls in secular schools and boys in Islamic schools perform better in English reading comprehension. Girls in secular schools excel in understanding text demands, text meanings, and text conclusions. Conversely, boys in public schools and girls in Islamic schools struggle with tasks like locating information and evaluating text content in reading assessments. Thus, the poor performance of boys and those in public secular schools are the main contributors to the overall low scores obtained by secular schools. For this reason, recognizing these dynamics is crucial for designing effective interventions tailored to their learning behaviours and needs are urgently needed in all school settings, including in secular-Islamic schools. This perspective shift encourages further exploration of students' learning attitudes, and teaching methods across school types, enhancing our understanding of diverse factors affecting English proficiency.

Appendix

Table 7 Distribution and homogeneity test of the students' reading and listening performance in Sekolah and Madrasah groups

		N	Min	Max	Mean	SD	Skewness		Kurtosis	
							Statistic	SE	Statistic	SE
Distribution of Student English Achievement										
Secular/Sekolah School (SS)	Listening	726	-3.18	3.15	-0.44	0.90	-0.11	0.09	0.77	0.18
	Reading	726	-3.10	3.10	-0.54	0.99	0.42	0.09	1.50	0.18
Islamic/Madrasah School (MS)	Listening	593	-3.18	1.93	-0.43	0.90	-0.30	0.10	0.62	0.20
	Reading	593	-3.10	3.10	-0.24	0.81	-0.03	0.10	1.11	0.20
Homogeneity Test of Student English Achievement (Levene's Test of Equality of Error Variances)										
English Achievement	F	Sig.								
Listening	3.32	0.07								
Reading	0.07	0.80								

Abbreviations

OECD   Organization for Economic Co-operation and Development
MoEC   Ministry of Education and Culture
MoRA   Ministry of Religious Affairs
WLE   Weighted likelihood estimate
RMM   Rasch measurement model
DIF   Differential item functioning
MNSQ   Mean square scores
ANOVA   Analysis of variance
SS   Sekolah schools
MS   Madrasah schools
SECTOR School sector
GENDER Student gender
M   Mean score
SD   Standard deviation

Authors' contributions
These authors contributed equally.

Authors' information
Abu Nawas
Abu Nawas is currently a PhD candidate at the School of Education, The University of Adelaide, South Australia. His research is focused on comparative studies in education, the school system, religious education, and language and literacy studies. His current project is the effectiveness of secular and Islamic education on student outputs. Nawas' ORCID: https://orcid.org/0000-0003-0365-1020.
I Gusti Ngurah Darmawan
Dr Igusti Ngurah Darmawan is a researcher with a strong interest in ICT, Science and Mathematics education, as well as quantitative methods and measurement. He has 20 years of experience in analysing, evaluating, and reporting large-scale assessment data. He recently completed a project, funded by SA's DfE, that focused on the evaluation of the Brightpath program in public schools across South Australia; and is currently involved in an Australian Federal Government's Emerging Priorities Program, focusing on Preparing for Parenting in a Post-Pandemic world - School seminars to skill parents and teachers to support the wellbeing, behaviour and self-regulation of students. Darmawan's ORCID: https://orcid.org/0000-0002-7628-6434.
Nina Maadad
Dr Nina Maadad research interest is in comparative studies, refugee education, culture, education and languages and has taught these at a tertiary and secondary school levels. She is currently involved in a longitudinal study focusing on "Schooling and Education for Refugees". She has published a number of books, including the following titles: '*The Education of Arabic Speaking Refugee children and Young Adults'* (2021). '*Syrian Refugee Children in Australia and Sweden: Education and Survival among the Displaced, Dispossessedand Disrupted'* (2020); '*Schooling and Education in Lebanon for Syrian and Palestinian Refugees Inside and Outside the Camps'* (2017); '*Academic Mobility: International Perspectives on Higher Education Research'* (2014); and '*The Adaptation of Arab Immigrant to Australia: Psychological, Social, Cultural and Educational Aspects'* (2007). Maadad'd ORCID: https://orcid.org/0000-0002-4903-8400.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by the University of Adelaide's Office of Research Ethics, Compliance and Integrity (Approval No: H-2020-038).

Competing interests
The authors declare that they have no competing interests.

References
Adams, Raymond J, Margaret L Wu, and Mark R Wilson. (2017). "ACER conquest." In *Handbook of Item Response Theory*, 495-506. Chapman and Hall/CRC.
ADB. (2014). *Indonesia: Madrasah Education Development Project*. Jakarta: Asian Development Bank.
Ahmadi, Seyedeh Masoumeh. (2016). The importance of listening comprehension in language learning. *International Journal of Research in English Education, 1*(1), 7–10. www.ijreeonline.com.
Akinwumi, Julius Olaitan. (2017). Effects of gender and school location on the Ekiti State secondary schools students' achievement in reading comprehension in English language. *Journal of Education and Practice, 8*(5), 50–55. www.iiste.org.
Alagumalai, Sivakumar, Curtis, David D., & Hungi, Njora. (2005). *Applied Rasch measurement: a book of exemplars*. The Netherlands: Springer. https://link.springer.com/book/10.1007/1-4020-3076-2 .
Ali, Mohammad, Kos, Julie, Lietz, Petra, Nugroho, Dita, Furqon, Zainul, Asmawi, & Emilia, Emi. (2011). *Quality of education in Madrasah*. Indonesia: The World Bank.
Al-Sohbani, Yehia Ahmed Y. (2018). Foreign language reading anxiety among Yemeni secondary school students. *International Journal of English Language & Translation Studies, 6*(1), 57–65. http://www.eltsjournal.org/archive/value6%20issue1/7-6-1-18.pdf .
Ameli, Saied Reza, Azam, Aliya, & Merali, Arzu. (2006). *Great Britain*. Great Britain: Islamic Human Rights Commission.
Araújo, L., Flisi, S., Dinis, P., & da Costa, and E Soto-Calvo. (2015). *Languages and employability*. Publications Office of the European Union. https://doi.org/10.2788/860807

Arthur, Courtney, Badertscher, Eden, Goldenberg, Paul, Moeller, Babette, McLeod, Matt, Nikula, Johannah, & Reed, Kristen. (2017). *Strategies to improve all students' Mathematics learning and achievement*. Education Development Center Inc.

Asadullah, Nazmul Chaudhury, & Dar, Amit. (2007). Student achievement conditioned upon school selection: religious and secular secondary school quality in Bangladesh. *Economics of Education Review, 26*(6), 648–659. https://doi.org/10.1016/j.econedurev.2007.10.004

Attah, R. F., & Ita, P. M. (2017). Gender as predictor of academic achievement in English among senior secondary school two students in Calabar Metropolis, Cross River State. *Global Journal of Educational Research, 16*(2), 149–153. https://doi.org/10.4314/gjedr.v16i2.9

Bećirović, Senad. 2017. "The relationship between gender, motivation and achievement in learning English as a foreign language." European Journal of Contemporary Education 6 (2): 210-220. https://doi.org/10.13187/ejced.2017.2.210

Berends, Mark, Joseph, R., & Waddington. (2018). School choice in Indianapolis: effects of charter, magnet, private, and traditional public schools. *Education Finance and Policy, 13*(2), 227–255. https://doi.org/10.1162/edfp_a_00225

Bingol, Mustafa Azmi, Celik, Behcet, Yildiz, Naci, & Mart, Cagri Tugrul. (2014). Listening comprehension difficulties encountered by students in second language learning class. *Journal of Educational & Instructional Studies in the World, 4*(4), 1–6.

Bond, Trevor G., & Fox, Christine M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences*. New York and London: Routledge, Taylor and Francis Group.

Bowen, Jean Donald, Madsen, Harold Stanley, & Hilferty, Ann. (1985). *TESOL Techniques and Procedures*. London: Newbury House Pub.

Bozorgian, Hossein. (2012). The relationship between listening and other language skills in international English language testing system. *Theory and Practice in Language Studies, 2*(4), 657–663. https://doi.org/10.4304/tpls.2.4.657-663

Brown, Timothy A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications Inc.

Brown, Ted, & Bonsaksen, Tore. (2019). An examination of the structural validity of the Physical Self-Description Questionnaire-Short Form (PSDQ–S) using the Rasch measurement model. *Cogent Education, 6*(1), 1–28. https://doi.org/10.1080/2331186X.2019.1571146

Cohen, Louis, Manion, Lawrence, & Morrison, Keith. (2002). *Research methods in education* (8th ed.). Routledge.

Davis, Frederick B. 1968. Research in comprehension in reading. *Reading Research Quarterly, 3*(4), 499-545. https://www.jstor.org/stable/747153.

Dirven, René, & Taylor, John Oakeshott. (1984). Listening comprehension (Part I). *Language Teaching, 17*(4), 326–343. https://doi.org/10.1017/S026144480001082X

Dorans, Neil J, and Paul W Holland. 1992. DIF Detection and Description: Mantel-Haenszel and Standardization. *ETS Research Report Series* (1), 1-40. https://files.eric.ed.gov/fulltext/ED387526.pdf.

Duru, Adem, & Koklu, Onder. (2011). Middle school students' reading comprehension of mathematical texts and algebraic equations. *International Journal of Mathematical Education in Science and Technology, 42*(4), 447–468. https://doi.org/10.1080/0020739X.2010.550938

Eng, Lin Siew, Mohamed, Abdul Rashid, & Javed, Muhammad. (2013). Analysis of students' competency in listening comprehension of the English language at Pakistani secondary school level. *Middle-East Journal of Scientific Research, 16*(3), 331–341. https://doi.org/10.5829/idosi.mejsr.2013.16.03.2298

Ependi, Rustam. (2020). Solution to Islamic education problems transitive Islamic Perspective. *International Journal of Creative Research Thoughts (IJRT), 8*(11), 730–738. www.ijcrt.org.

Field, Andy. (2013). *Discovering statistics using IBM SPSS statistics*. Sage Publishing.

Grabe, William. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375–406.

Griffin, Megan M., and Trisha D. Steinbrecher. (2013). "Chapter four - large-scale datasets in special education research." In *International Review of Research in Developmental Disabilities*, Edited by Richard C. Urbano, 155-183. Academic Press.

Hendajany, Nenny. (2016). The effectiveness of public vs private schools in Indonesia. *Journal of Indonesian Applied Economics, 6*(1), 66–89. https://doi.org/10.21776/ub.jiae.2016.006.01.4

Hussain, Muhammad Athar, Shahid, Saeed, & Zaman, Amir. (2011). Anxiety and attitude of secondary school students towards foreign language learning. *Social and Behavioral Sciences, 29*, 583–590. https://doi.org/10.1016/j.sbspro.2011.11.279

Isadaud, Dzulkifli, Fikri, M Dzikrul, & Bukhari, Muhammad Imam. (2022). The urgency of English in the curriculum in Indonesia to prepare human resources for global competitiveness. *DIAJAR: Jurnal Pendidikan dan Pembelajaran [Education and Learning Journal], 1*(1), 51–58. https://journal.yp3a.org/index.php/DIAJAR .

Jaccard, James. (1998). *Interaction effects in factorial analysis of variance*. Sage Publishing.

Keenan, Janice M., Betjemann, Rebecca S., & Olson, Richard K. (2008). Reading comprehension tests vary in the skills they assess: differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281–300. https://doi.org/10.1080/10888430802132279

Kholis, Nur, & Murwanti, Murwanti. (2019). Teacher professionalism in Indonesia, Malaysia, and New Zealand. *TARBIYA: Journal of Education in Muslim Society, 6*(2), 179–196. https://doi.org/10.!5408/tjems.v6i2.11487

Lamla, M. U. H. M. I. N. T. (2018). Issues and concerns on madrasah education in Basilan, Philippines: the Asaatiz perspectives. *International Journal of Novel Research in Education and Learning, 5*(4), 1–4.

Leong, Lai-Mei., & Ahmadi, Seyedeh Masoumeh. (2017). An analysis of factors influencing learners' English speaking skill. *International Journal of Research in English Education, 2*(1), 34–41. www.ijreeonline.com.

Longweni, Mpumelelo, & Kroon, Japie. (2018). Managers' listening skills, feedback skills and ability to deal with interference: a subordinate perspective. *Acta Commercii, 18*(1), 1–12. https://doi.org/10.4102/ac.v18i1.533

Magulod, Jr Gilbert, & C. (2017). Factors of school effectiveness and performance of selected public and private elementary schools: implications on educational planning in the Philippines. *Asia Pacific Journal of Multidisciplinary Research, 5*(1), 73–83.

Mart, Çagri Tugrul. (2012). Developing speaking skills through reading. *International Journal of English Linguistics, 2*(6), 91–96. https://doi.org/10.5539/ijel.v2n6p91

Mermelstein, Aaron David. (2015). Improving EFL learners' writing through enhanced extensive reading. *Reading in a Foreign Language, 27*(2), 182–198. http://nflrc.hawaii.edu/rfl .

Mills, Geoffrey E., & Gay, L. R. (2016). *Educational research: competencies for analysis and applications* (11th ed.). Pearson Education Limited.

Mirizon, Soni, Diem, Chuzaimah Dahlan, & Vianty, Machdalena. (2018). Students' specific comprehension skills in English based on school locations, grades, and gender. *Indonesian Journal of Applied Linguistics, 7*(3), 538–548. https://doi.org/10.17509/ijal.v7i3.9799

MoEC. (2017a). *Indonesia Educational Statistics in Brief 2016/2017*. edited by Center for Educational Data and Statistics and Culture. Jakarta: Minister of Education and Culture (MoEC).

MoEC. (2017b). *Ujian Nasional Tahun Pelajaran 2016/2017 [National Examination Test for Academic Year 2016/2017]*. Edited by Badan Standar Nasional Pendidikan [National Education Standards Agency]. Jakarta, Indonesia: Ministry of Education and Culture.

Mohamed, Abdullah Adnan, Radzuan, Noor Raha Mohd., Kassim, Hafizoah, & Ali, Mohammad Musab Azmat. (2014). Conceptualizing English workplace communication needs of professional engineers: the challenges for English language tertiary educators. *Selangor Business Review (SBR), 1*(1), 1–9.

Muhajir. (2016). The implementation of Madrasah-based management (MBM) at Man 1 and Man 2 Serang City, Banten, Indonesia: a comparative study. *Higher Education Studies, 6*(2), 50–60. https://doi.org/10.5539/hes.v6n2p50

Mulualem, Yihun Getachew, Mulu, Yalew Endawoke, & Gebremeskal, Tilahun Gidey. (2022). Effects of English learning beliefs on english achievement: academic emotions as mediators. *Heliyon, 8*(7), 1–14. https://doi.org/10.1016/j.heliyon.2022.e09829

Murtafi'ah, Banatul, & Putro, Nur Hidayanto Pancoro Setyo. (2020). Gender differences in Santri's reading attitude and motivation. *EduLite: Journal of English Education, Literature and Culture, 5*(2), 251–262. https://doi.org/10.30659/e.5.2

Musa, Alice KJ., Bala, Dauda, & Umar, Mohammad A. (2016). Gender differences in achievement goals and performances in English language and mathematics of senior secondary schools students in Borno State, Nigeria. *Journal of Education and Practice, 7*(27), 165–175. www.iiste.org.

Muttaqin, Tatang, Wittek, Rafael, Heyse, Liesbet, & van Duijn, Marijtje. (2019). The achievement gap in Indonesia? Organizational and ideological differences between private Islamic schools. *School Effectiveness and School Improvement, 4*(7), 1–31. https://doi.org/10.1080/09243453.2019.1644352

Newhouse, David, & Beegle, Kathleen. (2006). The Effect of School Type on Academic Achievement Evidence from Indonesia. *Journal of Human Resources, 41*(3), 529–557. https://www.jstor.org/stable/40057268 .

Nyarko, Kingsley, Kugbey, Nuworza, Kofi, Collins Courage, Cole, Yaa Adubea, & Adentwi, Kobina Impraim. (2018). English reading proficiency and academic performance among lower primary school children in Ghana. *SAGE Open, 8*(3), 1–10. https://doi.org/10.1177/2158244018797019

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2021). *ISA 2025 Foreign Language Assessment Framework*. Paris: OECD Publishing. https://www.oecd.org/pisa/foreign-language/ .

Pajarwati, Dian, Mardiah, Hikmatul, Harahap, Ronitama Parsaulian, Siagian, Rusliani Octavia, & Ihsan, Muhammad Taufik. (2021). Curriculum reform in Indonesia: English education toward the global competitiveness. *ETDC: Indonesian Journal of Research and Educational Review, 1*(1), 28–36. https://doi.org/10.51574/ijrer.v1i1.51

Pallant, Julie. (2016). *SPSS manual: a step by step guide to data analysis using IBM SPSS*. Sydney: Allen & Unwin.

Porto, Melina, Houghton, Stephanie Ann, & Byram, Michael. (2018). Intercultural citizenship in the (foreign) language classroom. *Language Teaching Research, 22* (5), 484–498. https://doi.org/10.1177/1362168817718580

Rahman, Abu Rashed Md Mahbuber., Jalaluddin, Ilyana, Kasim, Zalina Mohd, & Darmi, Ramiza. (2021). Attitudes towards learning english among the Aliya Madrasah students in Bangladesh. *Indonesian Journal of Applied Linguistics, 11*(2), 269–280. https://doi.org/10.17509/ijal.v11i2.34121

Ross, Kenneth Norman. (2005). *Sample design for educational survey research*. Pergamon Press.

Saaty, Ahdab Abdalelah. (2022). The role of motivation and gender in English language learning for Saudi students. *International Journal of English Linguistics, 12*(3), 76–88. https://doi.org/10.5539/ijel.v12n3p76

Sirozi, Muhammad. (2004). Secular–religious debates on the Indonesian national education system: colonial legacy and a search for national identity in education. *Intercultural Education, 15*(2), 123–137. https://doi.org/10.1080/1467598042000224998

Solak, Ekrem, & Erdem, Gamze. (2016). *Teaching language skills for prospective English teachers*. Ankara: Pelkan.

Stern, Jonathan MB., & Smith, Thomas M. (2016). Private secondary schools in Indonesia: what is driving the demand? *International Journal of Educational Development, 46*, 1–11. https://doi.org/10.1016/j.ijedudev.2015.11.002

Trang, Nguyen Minh. (2022). Gender differences in listening: research perspectives. *Indian Journal of Langauge and Linguistics, 3*(3), 12–24. https://doi.org/10.54392/ijll2233

Warm, Thomas A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Whitten, Christy, Labby, Sandra, & Sullivan, Sam L. (2016). The impact of pleasure reading on academic success. *The Journal of Multidisciplinary Graduate Research, 2*(4), 48–64.

Wright, Ben, & Linacre, John Michael. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm .

Wu, Margaret, Hak Ping Tam, and Tsung-Hau Jen. 2016. *Theory into practice educational measurement for applied researchers*. Springer Nature Singapore Pte Ltd: Springer. https://link.springer.com/book/10.1007/978-981-10-3302-5.

Zhang, Juan, Fan, Xitao, Cheung, Sum Kwing, Meng, Yaxuan, Cai, Zhihui, & Bi Ying, Hu. (2017). The role of early language abilities on math skills among Chinese children. *PloS One, 12*(7), 1–14. https://doi.org/10.1371/journal.pone.0181074

Publisher's Note