Organisation for Economic Co-operation and Development

EDU/WKP(2023)2

Unclassified **English text only**

22 February 2023

DIRECTORATE FOR EDUCATION AND SKILLS

Cancels & replaces the same document of 16 February 2023

AI SCORING FOR INTERNATIONAL LARGE-SCALE ASSESSMENTS USING A DEEP LEARNING MODEL AND MULTILINGUAL DATA

OECD Education Working Paper No. 287

Tomoya OKUBO (OECD), Wayne HOULDEN (Janison), Paul MONTUORO (Janison), Nate REINERTSEN (OECD), Chi Sum TSE (OECD), Tanja BASTIANIC (OECD)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Tomoya OKUBO, tomoya.okubo@oecd.org
Wayne HOULDEN, whoulden@janison.com

JT03512875

OECD EDUCATION WORKING PAPERS SERIES

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Acknowledgements

The authors would like to thank Elena Govorova[1] and Leandro Marino[2] for their constructive comments and careful reviews. Jenny Baracaldo Fernández and Stephen Flynn contributed to the editing and production of this paper. The preliminary version of this paper was presented in the 54th PISA Governing Board technical session.

[1] 2E Estudios y Evaluaciones

[2] Fundação Cesgranrio

Abstract

Artificial Intelligence (AI) scoring for constructed-response items using recent advancements in multilingual, deep learning techniques utilising models pre-trained with a massive multilingual text corpus, is examined using international large-scale assessment data. Historical student responses to Reading and Science literacy cognitive items developed under the PISA analytical framework are used as training data for deep learning together with multilingual data to construct an AI model. The trained AI models are then used to score and the results compared with human-scored data. The score distributions estimated based on the AI-scored data and the human-scored data are highly consistent with each other; furthermore, even item-level psychometric properties of the majority of items showed high levels of agreement, although a few items showed discrepancies. This study demonstrates a practical procedure for using a multilingual data approach, and this new AI-scoring methodology reached a practical level of quality, even in the context of an international large-scale assessment.

Table of Contents

1. Introduction

1.1. Constructed-response item in educational assessment

Educational assessments are conducted primarily for achievement surveys or the selection of test takers; however, they are not limited only to such purposes. In the literature, various effects of educational assessments are reported. For example, the design of a test affects test takers' learning strategy, motivation, and achievement (Crooks, 1988[1]; Harkins, 2001[2]). Furthermore, consolidation (Halpin and Halpin, 1982[3]; Nungester and Duchastel, 1982[4]) and self-regulation of learning are considered important benefits of taking educational assessments (Bloom, Hastings and Madaus, 1971[5]; Butler and Winne, 1995[6]). Hence, an assessment of learners can be influential, especially in a large-scale assessment. It is thus argued that the design of large-scale assessments should be based not only on the validity and reliability of assessments, but also on consideration of the social and ethical impact of how the results are interpreted and used (Frederiksen and Collins, 1989[7]; Messick, 1989[8]).

Mainly, there are two types of response formats in educational assessments: selected-response formats and constructed-response formats. The single-best answer type for multiple-choice questions is the most frequently used of the selected-response formats, which can be automatically scored if it is a computer-based assessment (CBA) without being subject to human error. Regarding the single-best type, test takers are not required to construct their own responses or show a process of leading students' responses based on their organised knowledge; thus, this type of item is occasionally criticised for the validity of measures. Multiple-choice formats are sometimes considered unfavourable because of the potential for guessing, which prejudices the reliability of the proficiency estimates. However, some research has revealed that test takers do not always guess, and instead select responses from a subset of the total item options, depending on proficiency (Downing, 2003[9]; Haladyna, 2004[10]). Therefore, the criticisms of the selected-response formats in terms of reliability are not widely justifiable.

Constructed-response formats demand that test takers write their responses in the form of one or more sentences, in which many response patterns can be correct. It is known that only skilled and well-trained raters can make reliable and valid judgements (Stalnaker, 1951[11]). Among constructed-response formats, the essay format (e.g., over 200 words) sometimes aims to measure writing skills in addition to the proficiency of the target domain; however, it is not the focus of this paper. Note that not all constructed-response items measure complex and higher-order cognitive ability, despite being generally regarded as drawing on higher-order cognitive processes. What is measured depends not only on the response format of an item, but also on its content (Jolly, 2010[12]). Some studies have revealed that the reliabilities of constructed-response items, including performance assessments, have a large variance in the interaction term between respondents and items (Brennan, 2001[13]), which implies that the scores of a respondent depend on the item contents involved in the assessment. Therefore, constructed-response items should cover a wide range of content to ensure both the validity and reliability of measures, meaning many items are needed in test forms.

Although assessing test takers with constructed-response formats has meaningful effects on learners, some assessments and examinations employ only multiple-choice items to reduce costs, improve the accuracy of scoring in practice, and reduce the duration between test taking and the provision of results. In the OECD Programme for International Student Assessment (PISA), approximately one-third of cognitive items are constructed-response items (OECD, n.d.[14]). In PISA, in order to reduce the cost of scoring cognitive items and coding a few open-ended student questionnaires, the machine-supported coding system (MSCS) is implemented (Yamamoto et al., 2018[15]), where it automatically scores the student responses based on the judgements given by the human raters in the past PISA cycles only if the student responses appear in the historical data. MSCS has succeeded in improving efficiencies by about 20%-32% depending on domains in PISA 2018 (OECD, n.d.[14]). The MSCS employs perfect pattern-matching techniques; therefore, many student responses that are almost the same as historical data are not scored/coded automatically, although the contents of the items are equivalent.

1.2. Deep learning technology and AI scoring

Artificial intelligence (AI) scoring for essay-type items using natural language processing (NLP) techniques has been used in large-scale assessment (Shermis, Burstein and Leacock, 2006[16]). In most cases, intermediate indices are calculated based on the student's essay, with a focus on writing elements such as the number of total words, word complexity, and proportion of grammar errors so that a student's final score is computed based on the intermediate indices. In AI scoring for essays, the scoring methodology is more accurate when the student's essay is long enough to contain sufficient information. For details of AI scoring for essay-type items, see Ramesh and Sanampudi (2022[17]).

In PISA, constructed-response items are not designed to measure writing skills, but to assess the quality of a student's thinking rather than the student's final response itself (OECD, 2019[18]). Hence, the techniques and methodologies developed for scoring essay-type items are not aligned with PISA cognitive items. Although the MSCS is implemented in PISA, only the pattern-matching technique is applied, and deep learning technology is not yet applied for the item scorings in PISA. Leacock and Chodorow (2003[19]) introduced an automated scoring system for short constructed-response items; however, the applicable area of the technique was limited.

Recent technological developments in deep learning algorithms have been tremendous. Deep learning has been applied in many areas and has shown high performance on various tasks such as classification. Transformer is a deep learning technique proposed in Vaswani, et al. (2017[20]), which is a neural network model that learns context and meaning by tracking relationships in sequential data. In Transformer, Self-Attention and Position-wise Feed-Forward Network are incorporated in an encoder-decoder model, it has proven to perform better at some tasks than earlier models (Hu et al., 2020[21]). Therefore, Transformer models can be considered as a model that suits the task of automatically scoring constructed-response items.

In the current study, a deep learning model is applied for the scoring task of PISA-type, short constructed-response items, and the psychometric property of the AI scoring is examined. Comparability of the scores across countries/economies is crucial in international large-scale assessments (ILSAs); hence, the item functioning of each country/economy is examined carefully for all items. The same logic should be applied for scoring methodologies, where the item functioning given by the AI model should be equivalent to that by the human raters in every context since it directly impacts the student proficiency distribution.

1.3. International large-scale assessment and multilingual AI model

ILSAs are frequently delivered in multiple languages, so that test candidates can be tested in their language of instruction. Verified translations of coding guides are taken to be equivalent and are the lynchpin of training human raters for the constructed-response items in the test instruments. The invariance of item functioning is taken as evidence of the equivalence of the national scoring panels (OECD, n.d.[14]). However, there is little doubt that maintaining scoring standards across different panels of raters, whether from year to year or country to country, is complex and difficult. As a result, there is likely to be some variation in reliability and accuracy from panel to panel.

The use of AI scoring for constructed-response items in ILSAs raises several questions about whether equivalence can be improved across the different languages in which an ILSA is delivered, given that AI modelling of responses does not change on its own over time. However, other questions arise because AI scoring models are not based on verified translations of coding guides but on corpora of scored responses, and there may be idiosyncrasies in the human scoring. This study examines the ability of a deep learning model to accurately predict scores across multiple languages.

Modelling approaches that aim at cross-lingual understanding are increasing, where low-resource languages in terms of training data can be supported by the other language data and existing datasets such as BERT (Devlin et al., 2018[22]), RoBERTa (Liu et al., 2019[23]), T5 (Raffel et al., 2020[24]), and XLM-R (Conneau et al., 2019[25]). These modelling approaches can be trained with student responses from various languages in an ILSA, the resulting model can then be used to predict scores from responses in the various languages, and this has the potential to increase the international comparability of the scores as well as the accuracy of scoring. This can be a turning point for ILSAs. Before these modelling approaches appeared, multilingual data was considered a risk to comparability; however, their introduction may facilitate mutual support in data analysis across countries.

1.4. Objective and overview

In this study, the psychometric property of AI scoring using deep learning technology is examined. ILSA data is used with a particular focus on the accuracy of AI scoring statistics at the country/economy level. To analyse differences in the psychometric property of the AI model between languages, various sets of training data by languages are prepared. Furthermore, a modelling approach that incorporates a multilingual dataset is also considered in this study. A new scaling model is also proposed, which aims to increase the reliability and validity of the scoring by updating the methodology frequently used in ILSAs.

The overarching research question is, how can deep learning technology contribute to improvements in the assessment processes and subsequent outcomes of international large-scale assessments? The concrete questions identified and examined in this study are as follows.

1. When compared to human marking, how accurate is AI scoring using the recent deep learning technology compared with human scoring? Is language the main factor that impacts scoring accuracy? How many student responses should be prepared to train deep learning models in ILSAs?
2. Is it possible to build an efficient and accurate multilingual scoring model that can be applied to all of the languages that respondents use? Is such a model more accurate than using a set of language-specific models?
3. Are there ways to refine and improve the validity of initial deep learning scores? From a modelling perspective, should certainty of predicted scores be taken into account when estimating student proficiencies?

This study aims to examine the above-mentioned questions by using data from different language versions of the PISA-based Test for Schools (PBTS). The cognitive items of PBTS were developed by the Australian Council of Education Research (ACER, 2012[26]) under the analytical framework of PISA and were validated in a field trial. The PBTS is currently only delivered as a CBA, and it has collected student responses across more than 15 countries. The scale of PBTS is aligned with that of PISA; thus, the scales and constructs are comparable with PISA (Okubo et al., 2021[27]).

The remainder of this paper is structured as follows. In section 2, the data and methodology used in the present study are introduced. Information on the PBTS is also provided in this section. Section 3 reports the psychometric properties of the AI scoring applied to the PBTS data, the accuracy of the country-level score estimates of different groups (i.e., different languages) are investigated, and the item functioning in each AI training condition is assessed. A new scaling model that fits AI scoring is proposed and examined in section 4, and lastly, the summary and discussions are provided in section 5.

2. Data illustration

2.1. Constructed-response items used in the PISA-based Test for Schools

In the PBTS, there are 65 constructed-response items, of which 39 items are scored by human raters, and the rest are scored automatically or semi-automatically. In semi-automatic scoring, an exact matching technique is applied to student responses based on historical data that the scoring system references. Otherwise, if the number of correct response strings is limited, such as the keyword-input items in mathematics (MATH), the list of correct response strings is used for exact matching. The student responses that do not match the historical data or the correct response list will be scored by humans. The semi-automatic method improves the efficiency of the human scoring depending on the complexity of the student responses. Thus, the only technique used for improving the efficiency of scoring is exact pattern matching in PBTS.

Table 1 shows the proportions of the human-scored items and the constructed-response items in PBTS. In PBTS, the proportion of constructed-response items across mathematics (MATH), reading (READ), and science (SCIE) is 46.4%, while that of human-scored items is 27.9%; some short constructed-response items are scored automatically in PBTS (OECD, 2016[28]). Since the proportion of human-scored items is small in MATH, only READ and SCIE are the focus of this study. Specifically, the 37 human-scored items in READ and SCIE are the targets to be scored by AI. The items scored automatically (e.g., multiple-choice items, etc.) were kept in the dataset and used for the scaling analyses, which follows that no human resource was used at all for the scoring tasks during the experiment settings since the AI models scored the student responses of the constructed-response items.

Table 1: Proportion of human-scored items and all constructed-response items in PBTS

Domain	Subdomain	*N* of human-scored items (%)	*N* of constructed-response items (%)	*N* of items
MATH	Employing	2 (10.0%)	15 (75.0%)	20
	Formulating	0 (0.0%)	7 (63.6%)	11
	Interpreting	0 (0.0%)	3 (33.3%)	9
	Sub-total	2 (5.0%)	25 (62.5%)	40
READ	Access and retrieve	2 (11.8%)	4 (23.5%)	17
	Integrate and interpret	9 (50.0%)	9 (50.0%)	18
	Reflect and evaluate	7 (63.6%)	7 (63.6%)	11
	Sub-total	18 (39.1%)	20 (43.5%)	46
SCIE	Evaluate and plan	3 (25.0%)	3 (25.0%)	12
	Explain	11 (50.0%)	11 (50.0%)	22
	Scientifically interpret	5 (25.0%)	6 (30.0%)	20
	Sub-total	19 (35.2%)	20 (37.0%)	54
Total		39 (27.9%)	65 (46.4%)	140

2.2. Training data for AI scoring

In PBTS, the student's raw responses to the human-scored items are presented to the human raters on the scoring platform provided by the international platform provider (i.e., Janison). The student's raw responses are randomly assigned to the human raters, and more than 20% of the student responses are presented multiple times to different human raters in order to check raters' scoring consistency on each task. When double-scored responses are discrepant, senior raters adjudicate the final score. In this study, double-scored student responses that received consistent scores from both raters were used as the

training data. The scoring accuracy of human-scored items in PBTS was measured by dividing the proportion of consistently scored responses to the double-coded responses by the total number of the double-scored items.

In this study, datasets from different countries were used. These datasets were initially combined into four datasets (groups), as some of them share the same languages, although they are not completely the same. In addition to this, two distinct populations (cycle-1 and cycle-2) were prepared in one language group. Consequently, in total, five datasets (groups) were prepared for this study. Table 2 shows the descriptive statistics of the human-scored responses of the five groups, which includes the following datasets; language-A (group-A), language-B (group-B), language-C (C1 and C2; groups C1 and C2), and language-D (group-D). Single-scored student responses were scored by one rater, whereas double-and triple-scored responses were scored by multiple raters. The percentage of double-scored responses also refers to the percentage of double-scored responses in each country that received consistent ratings from both human raters. The percentage of triple-scored responses refers to the percentage of double-scored responses that received inconsistent ratings from the human raters. The definition of scoring accuracy used in this study is defined in section 2.4.

The proportions of the triple-scored responses varied from group to group; group-A had the highest proportion of triple-scored responses, meaning that this dataset had the lowest level of inter-rater agreement. On the other hand, group-C2 had the highest agreement among the groups. One considerable reason for this is that the human raters improved their scoring skills from the first to the second testing cycle. Details of the scoring accuracy are provided in section 3.1. The proportions of blank responses are also distributed widely, which is the function of the group's proficiency levels in the domains. In this study, the blank responses are treated in accordance with the PISA methodology (See PISA technical report (OECD, n.d.[14]) for details). The average percentage of single-scored, double-scored, and triple-scored responses across the groups in READ were 57.8%, 31.4 %, and 2.8%, respectively, and in SCIE the comparative average percentages were 55.5%, 26.8%, and 3.6%.

Table 2: Frequencies of scoring of students' responses by groups

Condition	Domain	Single-scored responses	Double-scored responses	Triple-scored responses	Blank responses
Human-A	READ	42.7%	44.6%	7.6%	5.1%
	SCIE	50.4%	33.9%	8.1%	7.7%
Human-B	READ	71.1%	16.0%	1.9%	11.0%
	SCIE	65.5%	15.3%	1.3%	17.9%
Human-C1	READ	72.2%	17.6%	1.7%	8.5%
	SCIE	64.1%	17.3%	1.9%	16.8%
Human-C2	READ	72.3%	17.6%	0.9%	9.2%
	SCIE	66.1%	15.6%	1.0%	17.3%
Human-D	READ	30.9%	61.1%	1.8%	6.2%
	SCIE	31.2%	52.5%	5.5%	10.8%
Average	READ	57.8%	31.4%	2.8%	8.0%
	SCIE	55.5%	26.9%	3.6%	14.1%

2.3. Training data and target data to be analysed

In order to examine the psychometric properties of the AI scoring, different conditions were prepared. There were three factors in these conditions, namely, the training data language, the number of responses used to train the AI, and the target data to be scored by trained AI. The training data included four languages from different language families. In two groups (i.e., two different languages; groups A and B), three different conditions of training responses were set in order to see the impact of the size of data used for training. Further, the different target data were analysed in language-C in order to assess the stability of the AI scoring and the human scoring between the two different testing cycles.

In addition to the factors mentioned above, four training datasets that combined different languages were prepared (i.e., multilingual training datasets). Here, responses were used without translation into a single language. In total, there were 15 different conditions set in the training data (Table 3). Note that the students in each group were not representative of a country or a language, and that they were taken partly from the original datasets. Furthermore, the different scaling coefficients are applied to the scores for the groups. As such, the statistical indices and scores calculated in this study should not be interpreted as representative of any country's/economy's broader characteristics. Please note that the student responses used for the AI training contained some identical (or quite similar) student responses within a training dataset; therefore, the number of training responses was not equal to the number of students' response patterns in the dataset.

Table 3: Training data and target data to be analysed

#	Condition	Training data	Language	*N* of training responses	Target data	*N* of target students
1	AI300-A	Group-A	A	300	Group-A	6 809
2	AI600-A			600		
3	AI800-A			800		
4	AI300-B	Group-B	B	300	Group-B	20 347
5	AI600-B			600		
6	AI1000-B			1 000		
7	AI1000C1-C1	Group-C1	C	1 000	Group-C1	51 237
8	AI1000C2-C1	Group-C2		1 000		
9	AI1000C1-C2	Group-C1		1 000	Group-C2	47 401
10	AI1000C2-C2	Group-C2		1 000		
11	AI300-D	Group-D	D	360	Group-D	5 244
12	AImix-A	Mix of groups	mixed	600 * 3 (A, B, C2) + 360 (D)	Group-A	6 809
13	AImix-B	Mix of groups	mixed	600 * 3 (A, B, C2) + 360 (D)	Group-B	20 347
14	AImix-C1	Mix of groups	mixed	600 * 3 (A, B, C2) + 360 (D)	Group-C1	51 237
15	AImix-D	Mix of groups	mixed	600 * 3 (A, B, C2) + 360 (D)	Group-D	5 244

AI training was performed item by item, and no information, other than student responses to the cognitive items, was included in the training data. Predicted scores of the items by the trained AI were saved in the cognitive dataset together with the other automatically scored items, such as the multiple-choice items. Data analyses were performed for each dataset. In the present study, the student responses used for AI training were also included in the target data to be analysed. This is because the trained AI model did not always result in perfectly accurate predictions, even on training dataset responses. Moreover, excluding student responses used for AI training may have jeopardised the stability of conditional distributions of student proficiency, leading to undesirable errors in these estimations. Furthermore, across test cycles, a substantial number of student responses were highly similar.

Therefore, excluding the trained responses risked losing this nuance in the data. Note, however, that in Table 3, the target data in conditions 8 and 9 did not contain any training data in target data to be scored. These two conditions allowed the AI models to be tested without any training data.

Table 4 reports the average number of human-scored items that students actually responded to in the test, and the total number of items that students responded to in booklets on average. Although all groups used the same booklet design, the number of completed items differed slightly between conditions because the number of items that students did not reach differed in each condition. Table 4 also reveals that in READ about 40% of the items were human-scored, and in SCIE about 34% of the items were human-scored. Note that these percentages are different from those reported in Table 1 because assignment rates of constructed-response items in the booklets were higher than that of multiple-choice items. The percentages of the human-scored items (i.e., AI-scored items) are approximately proportional[3] to the impact on the student proficiency estimation using AI scoring, which implies that the impacts of different scoring methodologies (i.e., AI scoring and human scoring) to the estimates of score distributions were about 40% and 34% in READ and SCIE, respectively.

Table 4: Average number of human-scored items students responded

Condition	Domain	*N* of human-scored items	*N* of responded items	*%*
Human-A	READ	7.6	19.4	39.6%
	SCIE	8.0	22.8	33.9%
Human-B	READ	7.5	19.2	39.6%
	SCIE	7.9	22.6	33.9%
Human-C1	READ	7.7	19.7	39.6%
	SCIE	8.1	23.0	33.9%
Human-C2	READ	7.6	19.5	39.6%
	SCIE	8.0	22.8	33.9%
Human-D	READ	7.5	19.2	39.6%
	SCIE	7.9	22.6	33.9%
Average	READ	7.6	19.4	39.6%
	SCIE	8.0	22.8	33.9%

2.4. Statistical model and index

In order to evaluate the psychometric properties of AI scoring, the following statistical indices were employed in this study.

Index for scoring accuracy of human-scored item

The scoring accuracy of the human-scored items, $\kappa_{gj}^{(H)}$ is defined as:

Equation 1

$$\kappa_{gj}^{(H)} = \frac{m_{gj}^{(2)}}{m_{gj}^{(2)} + m_{gj}^{(3)}}$$

[3] Strictly speaking, item information function defined with item parameters affect score estimates.

where $m_{gj}^{(3)}$ is the total number of triple-scored responses (i.e., inconsistent double-scored responses) of item $j(=1,\dots,J)$ in group $g(=1,\dots,G)$, and $m_{gj}^{(2)}$ is the number of double-scored responses (i.e., consistent double-scored responses). Equation 1 is a simple statistic which requires only two variables to compute; however, $\kappa_{gj}^{(H)}$ does not provide us with any information on how it impacts the point estimate of the mean μ_g and variance σ_g^2 of a group. Moreover, it is an index that is dependent on score distribution $N(\mu_g, \sigma_g^2)$, and may not reflect the reliability of scoring solely. This index also has a risk of overestimating accuracy since $m_{gj}^{(2)}$ contains the pattern that both human raters scored responses wrongly.

Index for scoring accuracy of AI-scored items

The scoring accuracy of AI-scored items is calculated differently from that for human-scored items since there are no triple-scored responses $m_{gj}^{(3)}$ in AI-scored responses. The definition of AI scoring accuracy is given by:

Equation 2

$$\kappa_{gj}^{(AI)} = \frac{m_{gj}^{*}}{m_{gj}}$$

where m_{gj}^{*} is the number of matched scores between the human-scored responses and the AI-scored responses, while m_{gj} denotes the total number of scored responses of item j in group g. $\kappa_{gj}^{(AI)}$ treats human-scored responses as "true" scores, which is obviously a limitation of the index. As Equation 1 and Equation 2 are not functions of student proficiency θ, it is not possible to evaluate the scoring accuracy at each level of student proficiency θ, which means the index does not provide us with information on scoring accuracy at each student proficiency level. Note that blank responses are excluded from calculations.

Root Mean Squared Deviation between expected frequency on human-scored and AI-scored responses

In PISA, Item Response Theory (Lord and Novick, 1968[29]) is employed for scaling, which is a latent variable modelling from which various analyses and methodologies have been developed since the late 1960s. In order to overcome the limitations of $\kappa_{gj}^{(H)}$ and $\kappa_{gj}^{(AI)}$, model-based evaluation is also conducted in this study. Concretely, root mean squared deviation (RMSD) is employed for the model-based evaluation of item functioning of AI-scored items.

In this study, the difference in the expected frequency conditioned on human-scored data and that of AI-scored data is employed to evaluate the accuracy of AI scoring. The deviation of the expected frequency of AI scoring from human scoring, $\mathrm{Dev}_{(gg')j}(\theta)$, is defined as follows:

Equation 3

$$\mathrm{Dev}_{(gg')j}(\theta) = o_{gjk}(\theta) - o_{g'jk}(\theta)$$

Here, g denotes a group data scored by AI, while g' denotes that by human. $o_{gjk}(\theta)$ is the expected frequency of AI scoring of category $k(=1,\dots,K_j)$ in item j, which is defined as:

Equation 4

$$o_{gjk}(\theta) = \frac{\sum_{i=1}^{N_g} u_{gijk} h_{gi}(\theta|\boldsymbol{\varphi}_g)}{\sum_{k'=0}^{K_j-1} \sum_{i=1}^{N_g} u_{gijk'} h_{gi}(\theta|\boldsymbol{\varphi}_g)}$$
$$o_{g'jk}(\theta) = \frac{\sum_{i=1}^{N_g} u_{g'ijk} h_{g'i}(\theta|\boldsymbol{\varphi}_{g'})}{\sum_{k'=0}^{K_j-1} \sum_{i=1}^{N_g} u_{g'ijk'} h_{g'i}(\theta|\boldsymbol{\varphi}_{g'})}$$

where u_{gijk} is the binary scored data of student i to the category k of item j in AI-scored group g, $h_{gi}(\theta)$ is the conditional distribution of θ given by the parameters, and $\boldsymbol{\varphi}_g = [\mu_g, \sigma_g^2]$ are the parameters of the proficiency distribution of AI-scored group g. On the other hand, $o_{g'jk}(\theta)$ is expected frequency based on $\boldsymbol{\varphi}_{g'} = [\mu_{g'}, \sigma_{g'}^2]$ and $u_{g'ijk}$, which is the binary scored data given by human raters.

Unlike Equation 1 and Equation 2, because item response theory (IRT)-based item parameters are given in this study, the degree of discrepancy in Equation 3 is defined independently of the student proficiency, θ. Since Equation 3 is a function of θ and does not provide the accuracy of AI scoring across all proficiency levels, an RMSD index is calculated to measure the level of the item equivalence:

Equation 5

$$\text{RMSD}_{(gg')j} = \frac{1}{K_j} \sum_{k=0}^{K_j-1} \sqrt{\int_\theta \left(o_{gjk}(\theta) - o_{g'jk}(\theta)\right)^2 f_g(\theta)\, \mathrm{d}\theta}$$

where f_g is the proficiency distribution of the population.

The RMSD (Equation 5) is used to analyse differential item functioning (DIF) in PISA and PBTS, where the index is defined as the weighted average of the discrepancy between the model-based conditional probability and the expected frequency of student responses. Furthermore, in PISA and PBTS, 0.12 is set as the threshold for country/economy DIF for cognitive items (OECD, n.d.[14]).

Distribution of θ

In this study, the parameters of the distribution of θ, $h(\theta|\boldsymbol{\varphi}_g)$, are estimated in order to investigate the difference between the human-scored data and the AI-scored data. In IRT, under the given item parameter $\boldsymbol{\Lambda}_g^*$, the likelihood function to be maximised is defined as follows:

Equation 6

$$L(\boldsymbol{\Lambda}_g^*, \boldsymbol{\varphi} \mid \boldsymbol{u}_g) = \prod_{i=1}^{N} \int_\theta f(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}_g^*) h(\theta|\boldsymbol{\varphi}_g)\, \mathrm{d}\theta$$

Here, since each group has a different parameter set (i.e., national parameters and unique parameters), a subscript which denotes the group is attached to $\boldsymbol{\Lambda}^*$. The marginal likelihood is maximised via the expectation (E) and maximisation (M) algorithm (Dempster, Laird and Rubin, 1977[30]) is employed in the scaling phase of PISA and PBTS (See Baker and Kim (2004[31]) for details).

2.5. Pre-trained cross-lingual model and AI scoring

The last few years have seen significant advancements in the field of natural language processing (NLP). The two most important advancements being the development of Transformers (a new model that uses 'attention mechanisms' to track the relations between words across long text sequences in forward and reverse directions) and Transfer Learning (the use of large pre-trained models, which are then adjusted to perform a specific task).

In this study, the base XLM-R model is used for AI scoring. XLM-R is an NLP model that has been pre-trained with 2.5TB of filtered CommonCrawl data in 100 languages. The training data sets described in Table 3 are used to further train the base XLM-R model to predict scores. Each subset has different characteristics, the number of responses to be used for training and the languages of responses within the data set. The pre-trained cross-lingual XLM-R model captures generic language characteristics that ideally support the requirements for training scoring tasks in ILSA.

3. Psychometric property of AI coding

3.1. Accuracy of human scoring and AI scoring

Calculated averages of $\kappa_{gj}^{(H)}$ and $\kappa_{gj}^{(AI)}$ over the items of READ and SCIE are shown in Table 5. The average scoring accuracies of human-scored items across the groups range from 85% to 97% in READ and from 80% to 94% in SCIE., The average scoring accuracies for AI-scored items were almost equivalently spread from 84% to 94% in READ and from 80% to 94% in SCIE. In language-C, the scoring accuracy of the group-C2 (cycle-2) data was better than that of group-C1 (cycle-1) data, which is because the same scoring team performed the scoring tasks, and therefore, their scoring quality was improved.

Table 5 indicates that the accuracy of the AI scoring reaches that of the human scoring in the largest training data conditions in group-B and group-C1. Furthermore, the AI scoring accuracy overcame the human scoring accuracy in group-A, when the training data for each item included 800 resposnes. However, the AI scoring accuracy was always lower than the human scoring accuracy when the training data for each item included 300 responses. For groups C1 and C2, the AI scoring was more accurate when the group C2 data was used for AI training, compared to when the C1 data was used for AI training. This demonstrates the importance of the quality of AI training data for AI scoring accuracy.

For multilingual conditions, the number of AI training responses used for each language was 360 or 600. Nonetheless, in the multilingual conditions, the scoring accuracy improved. This indicates that, at least in this scenario, multilingual data trained for deep learning contribute productively to prediction models without the need to translate student responses into a single language.

Table 5: Scoring accuracies of human scoring and AI scoring

Condition	Training data	READ (Human)	READ (AI)	SCIE (Human)	SCIE (AI)
AI300-A	Group-A	85.4%	83.5%	80.7%	79.5%
AI600-A			88.3%		84.9%
AI800-A			89.3%		85.7%
AI300-B	Group-B	89.6%	86.5%	91.9%	87.2%
AI600-B			89.0%		89.6%
AI1000-B			89.2%		92.3%
AI1000C1-C1	Group-C1	91.3%	87.1%	90.1%	91.2%
AI1000C2-C1			91.3%		89.1%
AI1000C1-C2	Group-C2	95.4%	87.7%	93.7%	90.7%
AI1000C2-C2			94.0%		93.8%
AI360-D	Group-D	93.0%	90.7%	90.1%	88.1%
AImix-A	Group-A	85.4%	89.0%	80.7%	86.4%
AImix-B	Group-B	89.6%	89.2%	91.9%	92.3%
AImix-C1	Group-C1	91.3%	89.6%	90.1%	88.1%
AImix-D	Group-D	93.0%	91.0%	90.1%	90.0%

3.2. Item functioning of AI-scored data

This section compares the item functioning between human-scored items and the AI-scored items. Hence, $\text{Dev}_{(gg')j}(\theta)$ (Equation 3) was calculated for each item between the human-scored data and AI-scored data. Figure 1 shows $\text{Dev}_{(gg')j}(\theta)$ of the READ and SCIE conditions in group-A. Namely, the figures compare condition Human-A to conditions AI300-A, AI600-A, and AI800-A. Figure 2 reports the conditions in group-B. That is, the figures compare condition Human-B to conditions AI300-B, AI600-B, and AI1000-B. Figure 3 and Figure 4 present conditions in groups C1 and C2, which were based on the experimental design. Figure 5 compares human scoring with condition AI360-D. And, finally, Figure 6 to Figure 9 report the $\text{Dev}_{(gg')j}(\theta)$ of the READ and SCIE conditions in the multilingual conditions.

Each line in the figures depicts the $\text{Dev}_{(gg')j}(\theta)$ of an item, and the grey dashed lines denote the threshold of the RMSD index, which is set at 0.12 in accordance with PISA and PBTS. Since the RMSD is a weighted average deviation across θ (Equation 5), instances where lines partially cross the threshold line do not necessarily indicate DIF. Also note that the horizontal bar shown at the bottom of each figure represents the mean density function with regard to θ of the two groups (g and g'). Therefore, areas outside the black or grey bands might have insufficient student responses, and insufficient Fisher information to gain accurate insights about item functioning and AI scoring accuracy. As such, readers should focus their attention on the range of results within the black and grey bands, where approximately 95% of student responses lie.

In a $\text{Dev}_{(gg')j}(\theta)$ calculation, the expected frequency of human-scored items, $o_{g'jk}(\theta)$, is placed to the right of that of AI-coded items (Equation 3). Hence, the deviation, $\text{Dev}_{(gg')j}(\theta)$, takes a positive value (> 0) if human scoring is more severe than AI scoring, and takes a negative value (< 0) if the AI scoring is more lenient than the human scoring. $\text{Dev}_{(gg')j}(\theta)$ is slightly biased towards the positive side across the conditions, which implies the AI scoring is slightly more lenient than the human scoring.

Figure 1: Deviations of item functioning between Human-A and (AI300-A, AI600-A, AI800-A)

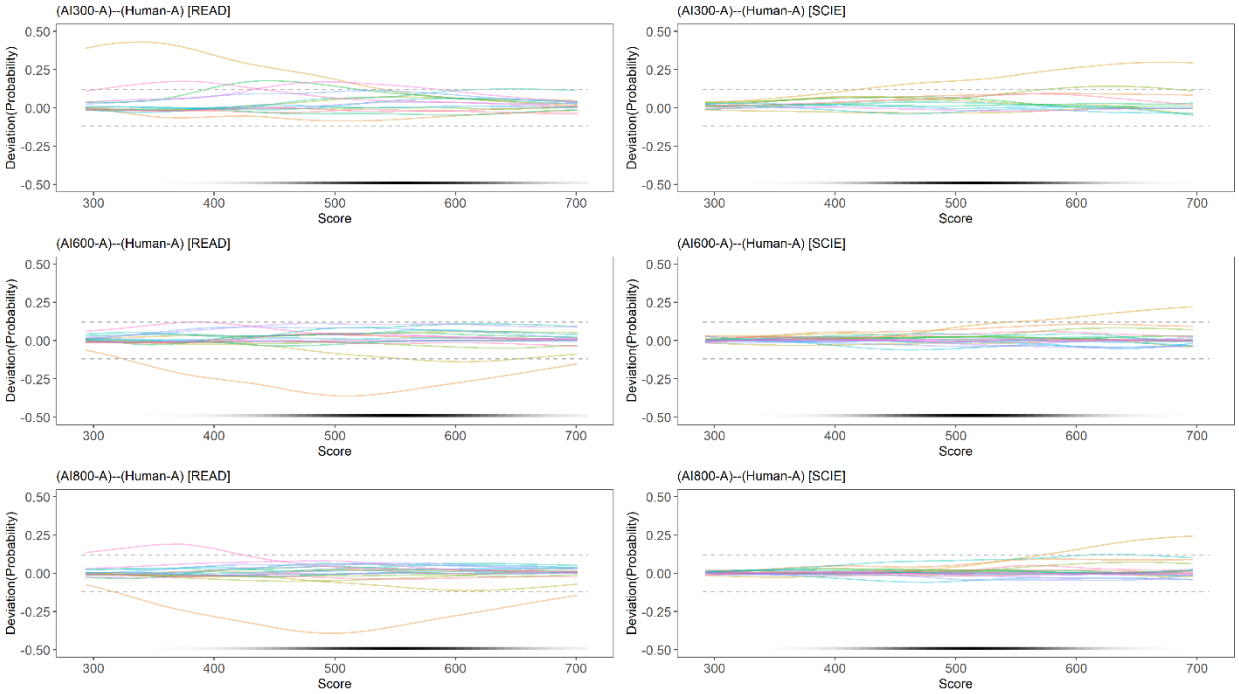


Figure 2: Deviations of item functioning between Human-B and (AI300-B, AI600-B, AI1000-B)

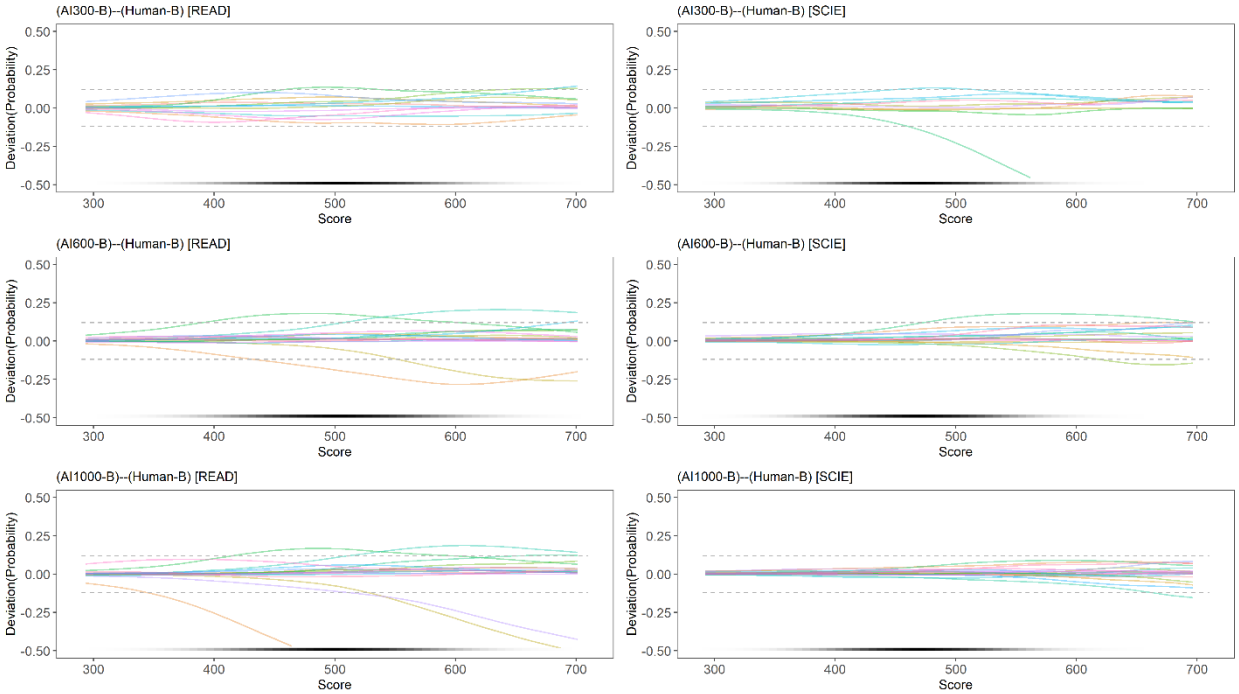


Figure 3: Deviations of item functioning between Human-C1 and (AI1000C1-C1, AI1000C2-C1)

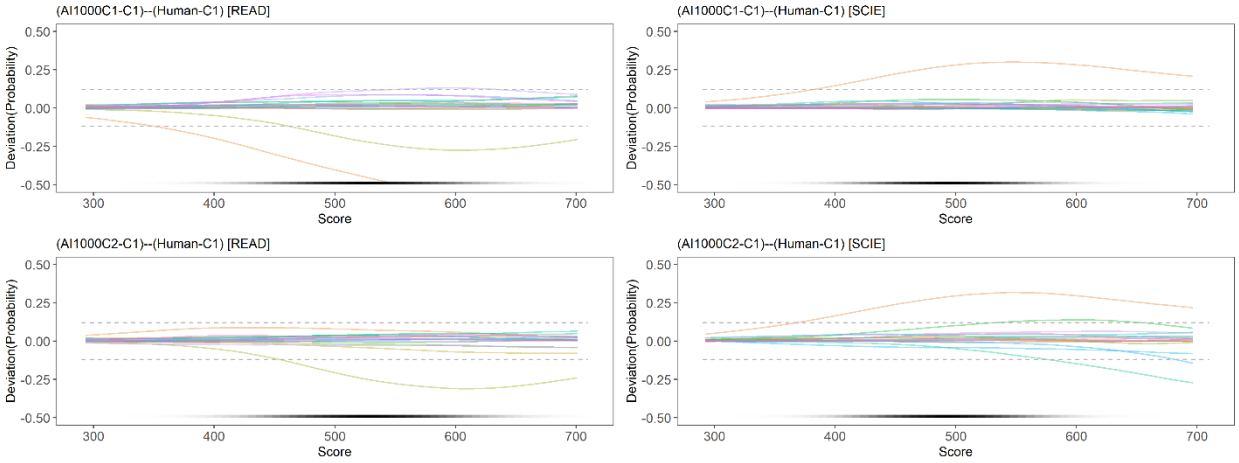


Figure 4: Deviations of item functioning between Human-C2 and (AI1000C1-C2, AI1000C2-C2)

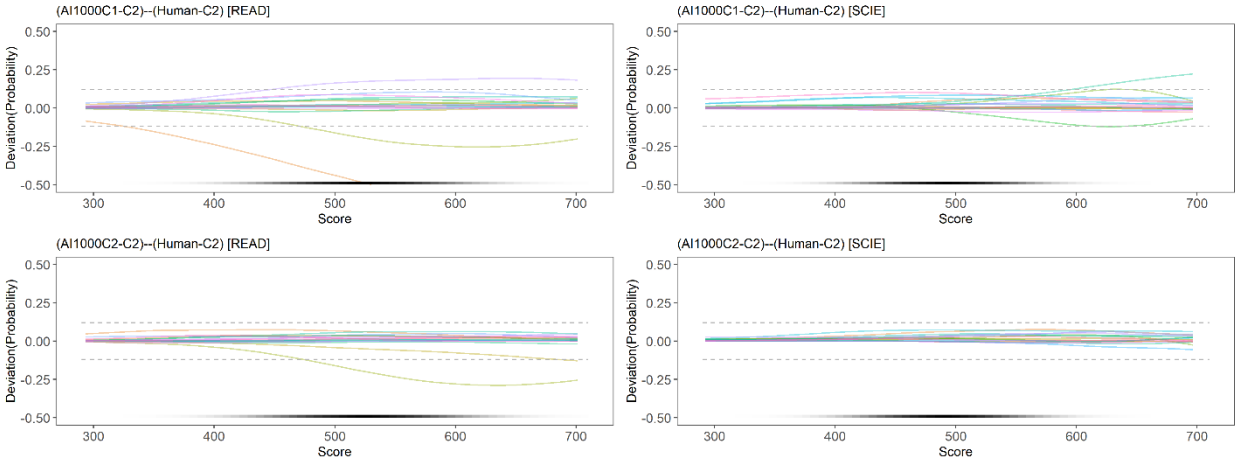


Figure 5: Deviations of item functioning between Human-D and AI360-D

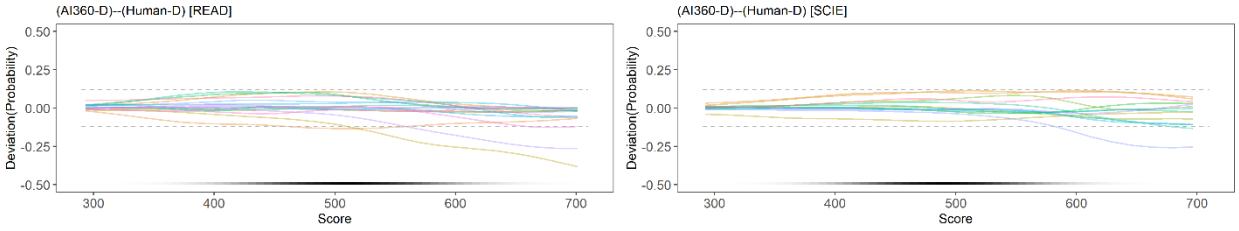


Figure 6: Deviations of item functioning between Human-A and AImix-A

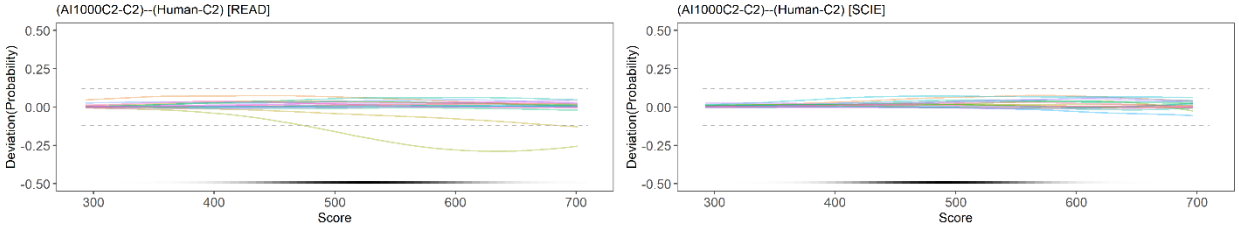


Figure 7: Deviations of item functioning between Human-B and AImix-B

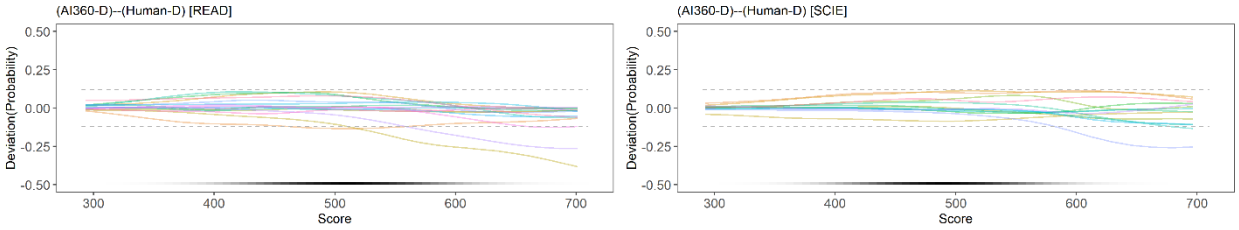


Figure 8: Deviations of item functioning between Human-C1 and AImix-C1

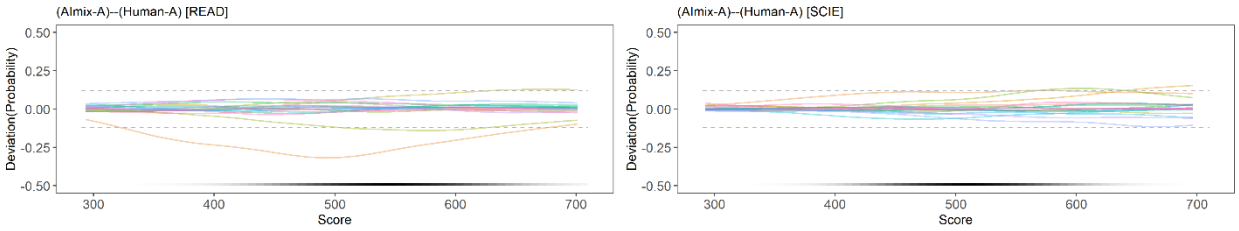


Figure 9: Deviations of item functioning between Human-D and AImix-D

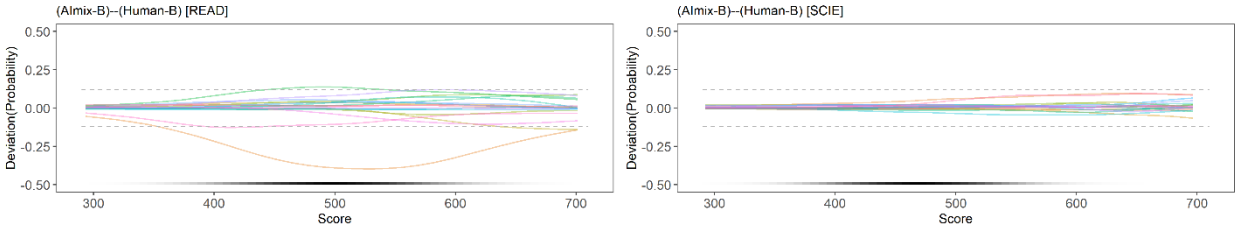


Figure 1 to Figure 9 indicate that the more student responses are used in AI training, the more the deviation lines conform to RMSD = 0.0. For the three different numbers of AI training response in Figure 1 and Figure 2, item functioning is not stable when the number of training responses is only 300 per item. In READ, the items represented in orange showed $\left|\mathrm{Dev}_{(gg')j}(\theta)\right| > 0.12$ for a wide range of proficiency (Figure 1 and Figure 2), regardless of the number of AI training responses. In SCIE, the item represented in orange showed different items functioning in many conditions. These items had DIF (i.e., scoring-mode DIF), in which case we can conclude that the AI scoring functioned differently from the human scoring. Here the group proficiency distribution $h\left(\theta|\boldsymbol{\varphi}_g\right)$ will be biased by the items if the items that show different functioning do not get appropriate treatments.

Deviation of the expected frequency $\mathrm{Dev}_{(gg')j}(\theta)$ in groups C1 (Figure 3) and C2 (Figure 4) showed different AI scoring behaviours in both READ and SCIE. The conditions using group-C2 data were more stable than conditions using the group-C1 data. Thus, comparisons of figures in Figure 4 and Figure 5 indicate that the quality of AI training data is essential to obtain an accurate AI scoring model. Figure 5 and Figure 9 indicate that the human scoring in group-D was more lenient than the AI scoring. The figures show that the human scoring was more lenient especially among high-performance students. In this study, only 360 student responses were used for the training; thus, further investigation on group-D is required to understand this result.

The item characteristics given by AI-scored data did not fit that of human-scored data in two READ items and one SCIE item. However, most items scored by AI had the same item functioning as those scored by human raters in both READ and SCIE domains. The number of items that showed discrepancies was limited regardless of the language. In this study, all items were retained for scaling in order to assess the impact of the precision of the AI scoring.

3.3. Impact on score distribution

In ILSAs, constructed-response items are scored by multiple human raters to ensure the reliability of scoring. Human raters use rubrics to score student responses to items. However, in educational assessment, restrictions in scoring schedules and financial costs often mean that not all student responses to constructed-response items are scored by multiple raters. In most cases, the aim of conducting an ILSA is understanding the broader population, not the proficiency of individual students. Therefore, the parameter of the interest is the proficiency distribution of the population, $\boldsymbol{\varphi}_g = [\mu_g, \sigma_g^2]$, not the proficiencies of any single student θ_i. Therefore, the main purpose of investigating human scoring quality is to check the bias of human raters on the population proficiency.

In this section, the distribution of θ of a group, $h(\theta|\boldsymbol{\varphi}_g)$, is estimated for each condition. Table 6 shows the estimates of the mean proficiency distribution of READ and SCIE. The estimates in parentheses represent the standard deviation of the distribution. Moreover, Figure 10 to Figure 18 show the distributions of θ, which compare scores based on human-scored data and AI-scored data. Note also that all items are included in the estimations, even though some items showed scoring-mode DIF.

Table 6: Estimated score distributions of READ and SCIE

Condition	READ (Human)	READ (AI)	Difference (AI-Human)	SCIE (Human)	SCIE (AI)	Difference (AI-Human)
AI300-A	530.7(113.6)	543.2(111.5)	+12.5(−2.1)	520.3(105.2)	533.7(115.5)	+13.4(+10.3)
AI600-A		534.0(113.3)	+3.0(−0.3)		528.0(108.7)	+7.7(+3.5)
AI800-A		531.1(112.9)	−0.4(−0.7)		526.8(108.7)	+6.5(+3.5)
AI300-B	435.2(105.2)	440.8(108.2)	+5.6(+3.0)	416.9(97.5)	424.9(96.6)	+8.0(−0.9)
AI600-B		441.5(106.6)	+6.3(+1.4)		422.8(97.8)	+5.9(+0.3)
AI1000-B		434.9(102.5)	−0.3(−2.7)		422.3(98.0)	+5.4(+0.5)
AI1000C1-C1	489.7(94.3)	487.6(92.0)	−2.1(−2.3)	470.3(83.5)	479.4(84.8)	+9.1(+1.3)
AI1000C2-C1		490.7(90.3)	+1.0(−4.0)		475.6(85.8)	+5.3(+2.3)
AI1000C1-C2	482.8(92.5)	481.5(90.1)	−1.3(−2.4)	467.5(82.1)	476.8(82.0)	+9.3(−0.1)
AI1000C2-C2		484.6(90.8)	+1.8(−1.7)		472.3(83.1)	+4.8(+1.0)
AI300-D	470.3(104.7)	456.3(93.0)	−14.0(−11.7)	480.3(104.9)	470.9(101.6)	−9.4(−3.3)
AImix-A	530.7(113.6)	524.5(114.9)	−6.2(+1.3)	520.3(105.2)	522.4(109.4)	+2.1(+4.2)
AImix-B	435.2(105.2)	435.4(103.8)	+0.2(−1.4)	416.9(97.5)	421.1(98.2)	+4.2(+0.7)
AImix-C1	489.7(94.3)	489.2(94.6)	−0.5(+0.3)	470.3(83.5)	475.4(85.2)	+5.1(+1.7)
AImix-D	470.3(104.7)	452.5(92.6)	−17.8(−12.1)	480.3(104.9)	471.0(97.4)	−9.3(−7.5)

In most cases, the distributions of θ based on the AI-scored data almost overlapped the distributions of θ based on the human-scored data, with the exception of conditions that only included 300 AI training responses. In most of these conditions, the AI scores overestimated student proficiencies in both READ and SCIE. However, this tendency was stronger in SCIE than READ. In other conditions (i.e., AI training data > 300), the AI scores in SCIE were overestimated, even though the number of training responses was as high as 1 000, which was equivalent to the results of the item functioning analysis reported in section 3.2. In READ, the variances of the score distributions estimated based on the AI-scored data were close to the expected score distributions, regardless of the conditions. Nonetheless, in SCIE, the variances were slightly overestimated in the AI scorings. Considering that standard errors of the mean in PISA 2018 ranged from about 1.0 to 3.0 in most countries/economies, the gap reported in the present study between human and AI scoring can be considered almost equivalent in READ when the number of training responses is large enough, but it was slightly overestimated in SCIE.

Figure 10: Score distributions of Human-A, AI300-A, AI600-A, and AI800-A

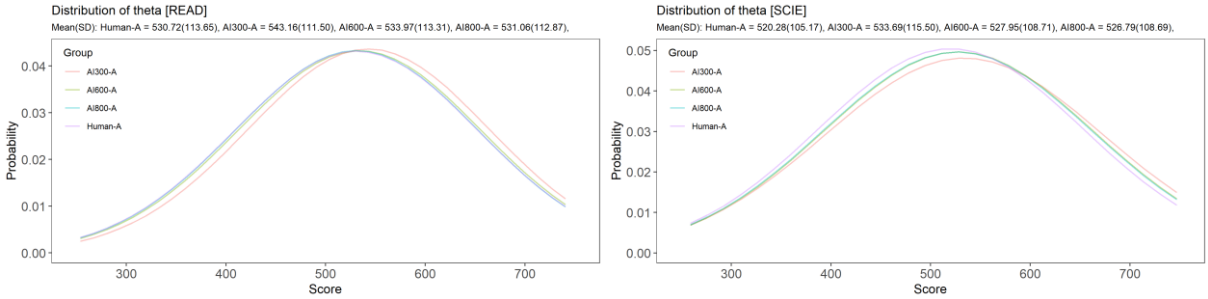


Figure 11: Score distributions of Human-B, AI300-B, AI600-B, and AI1000-B

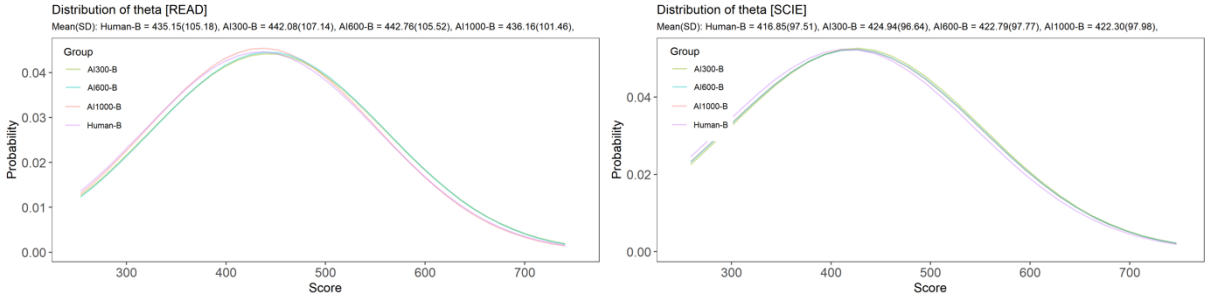


Figure 12: Score distributions of Human-C1, AI1000C1-C1, and AI1000C2-C1

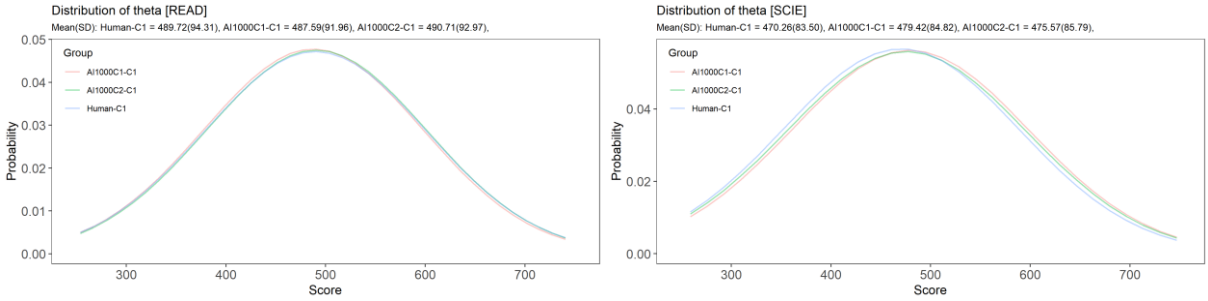


Figure 13: Score distributions of Human-C2, AI1000C1-C2 and AI1000C2-C2

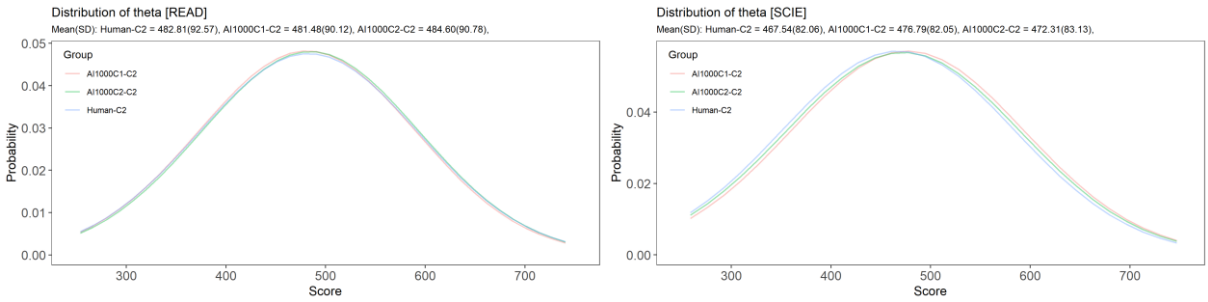


Figure 14: Score distributions of Human-D and AI360-D

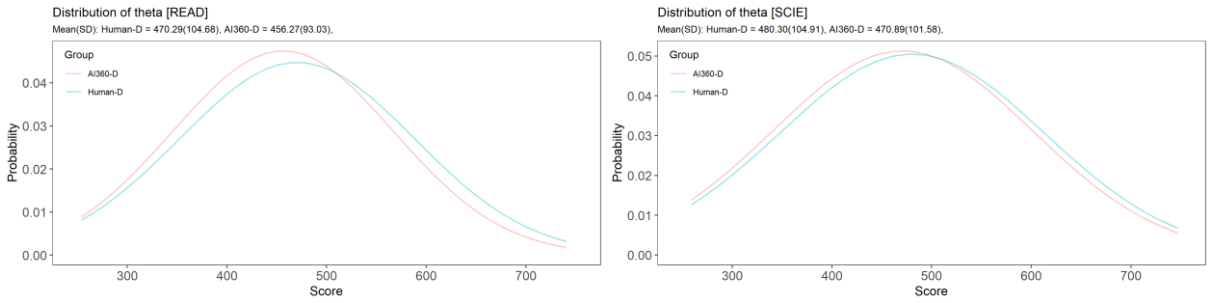


Figure 15: Score distributions of Human-A and AImix-A

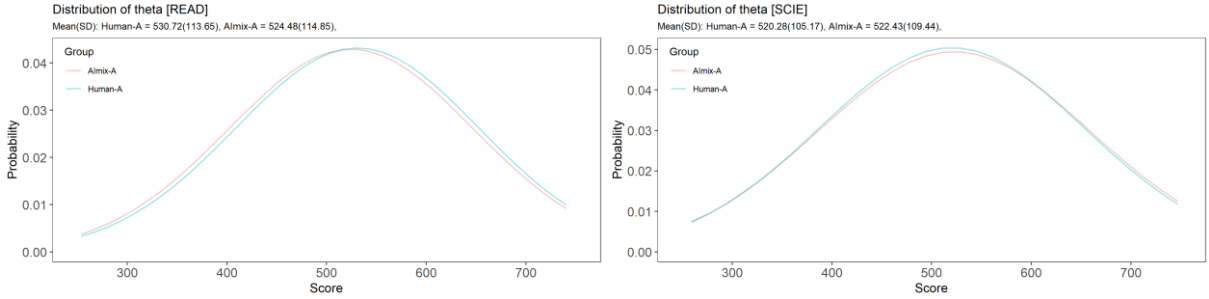


Figure 16: Score distributions of Human-B and AImix-B

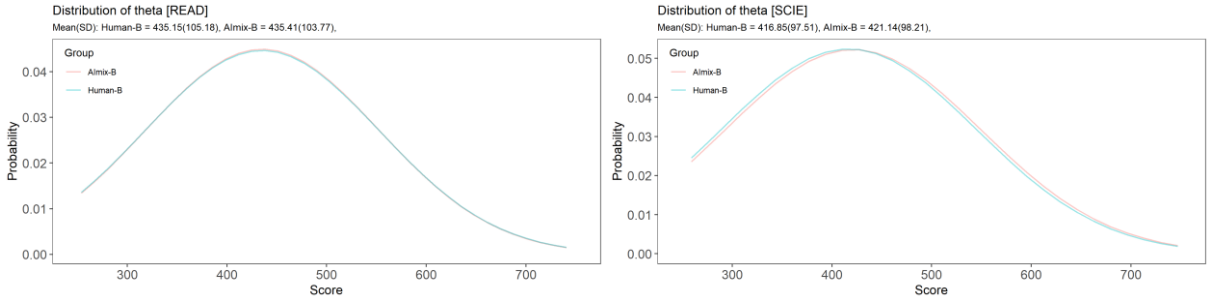


Figure 17: Score distributions of Human-C1 and AImix-C1

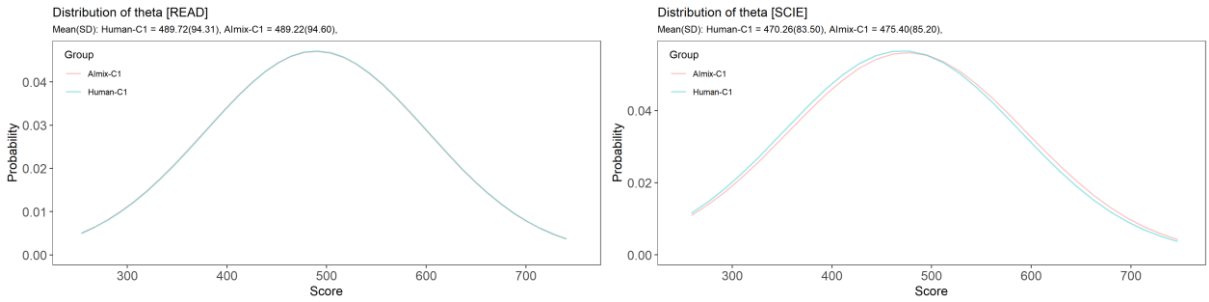
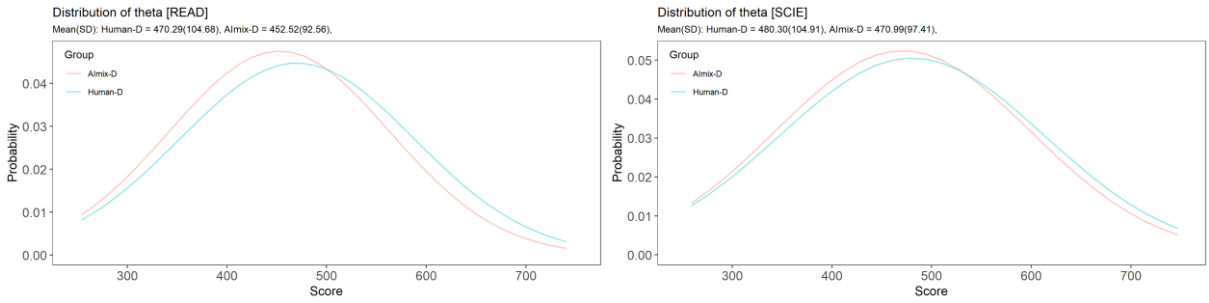


Figure 18: Score distributions of Human-D and AImix-D



4. Scaling student proficiency with probabilistic score data

4.1. Modelling the certainty of predictions in score estimation

Predicted AI scores have some degree of certainty. In this study, the term "*degree of certainty*" refers to a probabilistic score, ranging from 0 to 1. This section describes the scaling model based on probabilistic score data and its psychometric properties. Definitions of certainty can vary, from the probability of membership to a category, to the average predicted score based on multiple patterns of algorithmic and/or parameter settings. The latter mitigates the critics of the arbitrariness of choices of prediction algorithms and parameter settings of the prediction models. In this study, the probability of membership to a category is employed in a probabilistic score (PS) model.

Here, let x_{ij} be the score x of student $i(= 1, \dots, N_g)$ to item j. In some response formats, such as multiple-choice, x_{ij} can be perfectly reproduced, but this is not possible in constructed responses. Therefore, even though the prediction of x_{ij} is treated as equivalent across response formats, the nature of what is predicted is not necessarily the same.

In an IRT model, the likelihood function with respect to θ conditioned on the item parameters $\mathbf{\Lambda}_g$ is,

Equation 7

$$f(\boldsymbol{u}_{gi}|\theta, \mathbf{\Lambda}_g) = \prod_{i=1}^{N_g} \prod_{j=1}^{J} \prod_{k=0}^{K_j-1} p_{gjk}(\theta)^{u_{ijk}}$$

where $p_{gjk}(\theta)$ is the item characteristic response function for category k of item j in country/economy g, and u_{ijk} is the binary indicator of student i to category k of item j, such that,

$$u_{ijk} = \begin{cases} 1, & \text{if } x_{ij} = k \\ 0, & \text{otherwise} \end{cases}$$

The likelihood function is the main piece of information, and in most cases, the only piece of information used for estimating student proficiency. Binary data u_{ijk} can be replaced with w_{ijk}, which represents the membership probability of student i to category k of item j. This implies that the level of certainty of AI scoring can be taken into account in the log-likelihood function in the form of the weight of the predicted score; namely,

Equation 8

$$\log f(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}_g) = \sum_{i=1}^{N_g} \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} w_{ijk}\, p_{gjk}(\theta)$$

The parameter of the distribution of θ, $\boldsymbol{\varphi}_g$, can be estimated based on Equation 8 through the expectation-maximisation (EM) algorithm (Dempster, Laird and Rubin, 1977[30]). Further, $\boldsymbol{w}_{ik} = [w_{i1k}, w_{i2k}, \ldots, w_{ijk}]$ can take two different patterns; one is a certainty of the prediction for only the category that the AI predicted. For example, when $K_j = 2$,

Equation 9

$$\boldsymbol{w}_{ik} = \begin{bmatrix} 0 & 0.836 & \cdots & 0.934 \\ 0.901 & 0 & \cdots & 0 \end{bmatrix}$$

which mitigates the impact of prediction on the log-likelihood function. Another option is to compute the membership probability for each category, such as,

Equation 10

$$\boldsymbol{w}_{ik} = \begin{bmatrix} 0.089 & 0.836 & \cdots & 0.934 \\ 0.901 & 0.164 & \cdots & 0.066 \end{bmatrix}$$

Note that in this model $\sum_{k=0}^{K_j-1} w_{ijk} = 1$ holds. Also note that in this study the latter model is employed; however, in future applications the psychometric benefits and limitations of both models should be investigated.

4.2. Distribution of certainty of predictions

This section reports the distributions of the probabilistic scores given by the AI models. Only the results for language-A and language-B are presented, in which the proficiency distributions of these groups were about 50 points lower/higher than the OECD average; thus, wide distributions of w_{ijk} for these particular groups were expected. The histograms of $\boldsymbol{w}_{ik}$ for READ are depicted on the left, and those for SCIE are depicted on the right (Figure 19 and Figure 20). Each histogram reports w_{ijk} (Equation 10) for all i and j, coloured by k, depending on a training data condition.

Since group-A (language-A) received a high proportion of correct response patterns compared to incorrect response patterns, in Figure 19, the probabilistic score patterns for correct responses overlay the majority of response patterns for incorrect responses, particularly in READ. On the other hand, in Figure 20, where group-B (language-B) is reported, the opposite is true, and the probabilistic score patterns for incorrect responses overlay the majority of correct responses. As can be seen in both sets of histograms, when a greater number of human-scored student responses are used in AI training, w_{ijk} converges towards 1.0. Also note that in histograms reporting results based on the largest AI-trained datasets (i.e., $N = 1000$) w_{ijk} converges closely towards 1.0. This implies that the size of AI training datasets affects the level of uncertainty in AI scoring. Although the levels of certainty concentration can vary from group to group and/or model to model, w_{ijk} tends to converge towards 1.0 if the number of AI training responses exceeds 600. Further analysis in section 4 is performed with this data.

Figure 19: Distributions of certainty of prediction in language-A

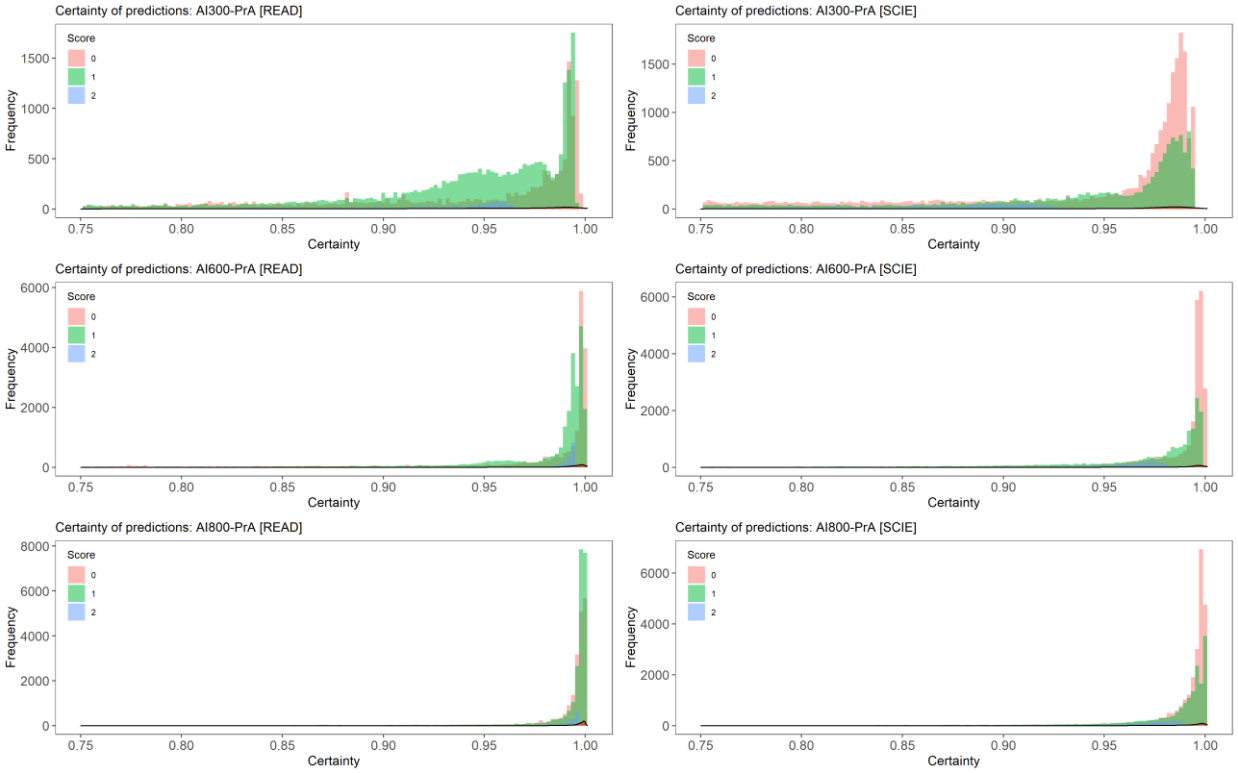
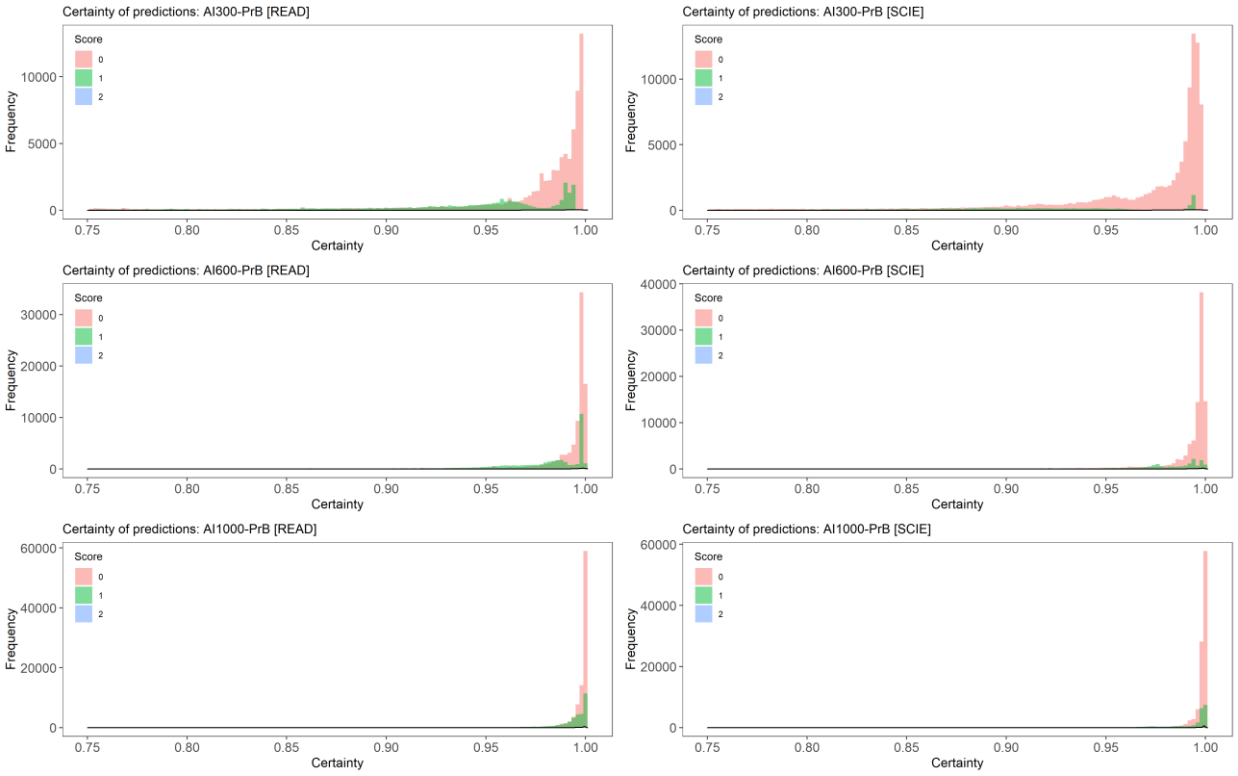


Figure 20: Distributions of certainty of prediction in language-B



4.3. Item functioning by probabilistic score model

Figure 21 and Figure 22 show $\text{Dev}_{(gg')j}(\theta)$ estimated using the probabilistic score (PS) model (Equation 8) of the conditions in language-A and language-B. Figure 21 compares human-scored data to AI-scored data for langauge-A, and Figure 22 compares these results for language-B. The results for READ are depicted on the left, and for SCIE on the right. In all graphs, each line (i.e., $\text{Dev}_{(gg')j}(\theta)$) represents an item. Lines that exceed a deviation of zero indicate that the AI predicted scores that were higher than the human scores and vice versa.

In Figure 21 and Figure 22, some $\text{Dev}_{(gg')j}(\theta)$ results are positively skewed in lower score ranges. Namely, AI scoring with the PS model tended to estimate lower-ability students higher than the human-based IRT model estimated these students. This model characteristic is especially evident in the figures for several items in SCIE. However, this skewness was relatively smaller for datasets based on large training datasets. Furthermore, when the AI training datasets were ≥ 600 responses, this skewness was limited within the population proficiency ranges, which are depicted by the black and grey bands on the x-axes). It is also important to note that conditional expectations of correct responses may be higher in the AI-based PS model because this model awards partial credit for responses that are incorrect.

Furthermore, in some items, $\text{Dev}_{(gg')j}(\theta)$ results in the high score range are negatively skewed. This implies that, in some items at least, the PS model was stricter than human raters for high-performing students who likely answered these items correctly. However, the degree of these AI deviations was generally limited to $\left|\text{Dev}_{(gg')j}(\theta)\right| < 0.12$, particularly in datasets trained with ≥ 600 student responses. Additionally, items with relatively variable certainty tended to show lenient marking in the low-performance range, and strict marking in the high-performance range.

Figure 21: Deviations of item functioning Human-A and AI300-PrA, AI600-PrA, and AI800-PrA

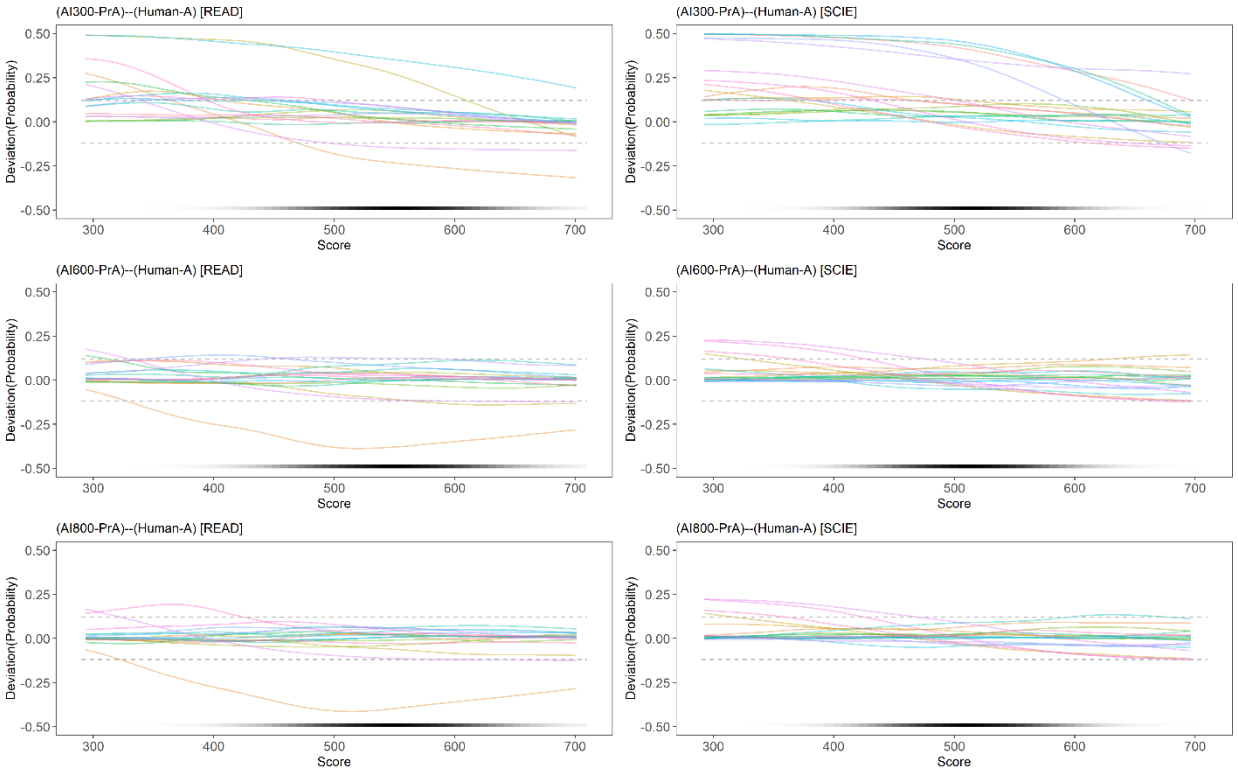
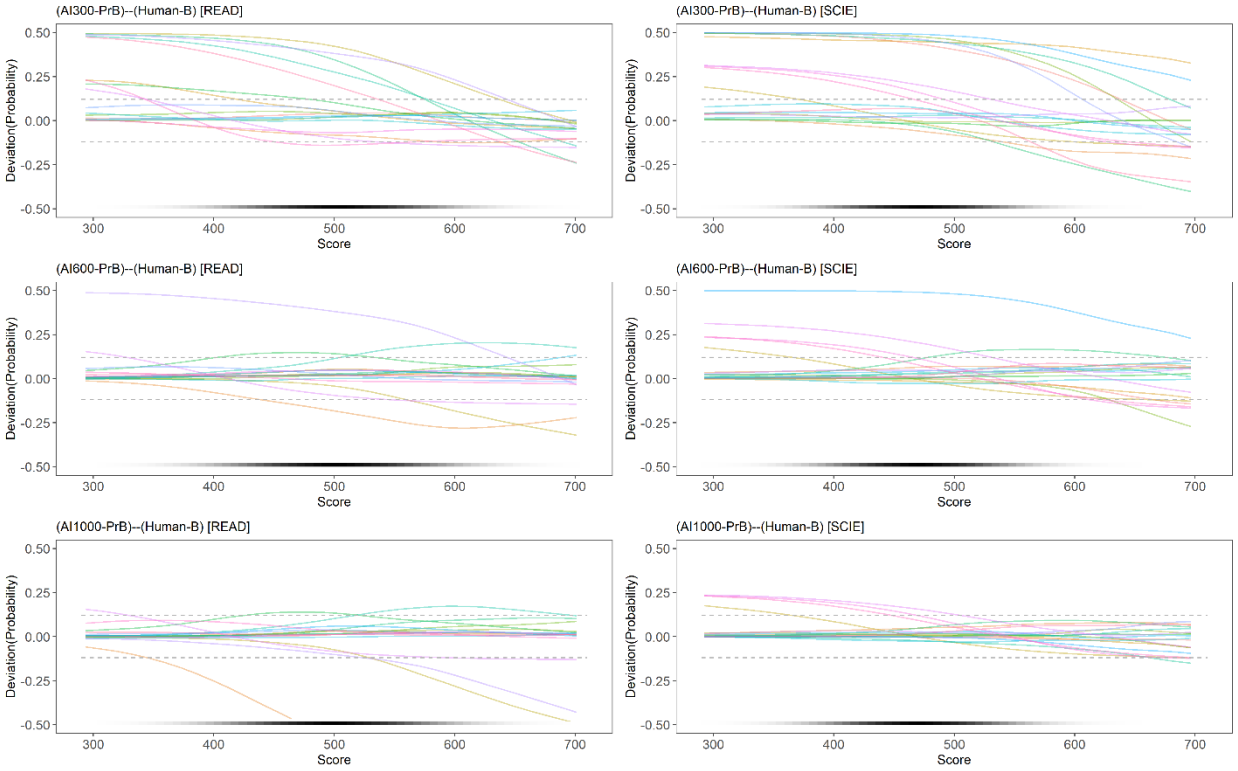


Figure 22: Deviations of item functioning Human-B and AI300-PrB, AI600-PrB, and AI1000-PrB



4.4. Score distributions estimated using the probabilistic score model

The estimated score distributions in language-A and language-B data are reported in Figure 23 and Figure 24, respectively. Descriptive statistics of the estimated distributions are also presented in Table 7. Comparing the score distributions estimated based on the PS model to that of AI-scored data using the original model, variance is smaller across the domains and the conditions. In READ, the average scores of the PS model converge to that of human-scored data if the number of training responses reaches 800, while SCIE average scores of the PS model are estimated to be slightly higher than the original model even if the number of training responses reaches 1 000.

Comparing the estimated score distributions of the PS model to the results based on the AI-scored data and a general IRT model shown in Table 6, the estimates of the average score in the small training data condition are improved in both domains; however, the estimated distributions using a PS model and a general IRT model are almost identical when the number of training responses is large. It is aligned with the data behaviour shown in Figure 19 and Figure 20, in which distributions of probabilistic scores are concentrated on 1.0; thus, it is almost equivalent to the predicted binary score (or polytomous score in some items).

Figure 23: Score distributions of Human-A, AI300-PrA, AI600-PrA, and AI1000-PrA

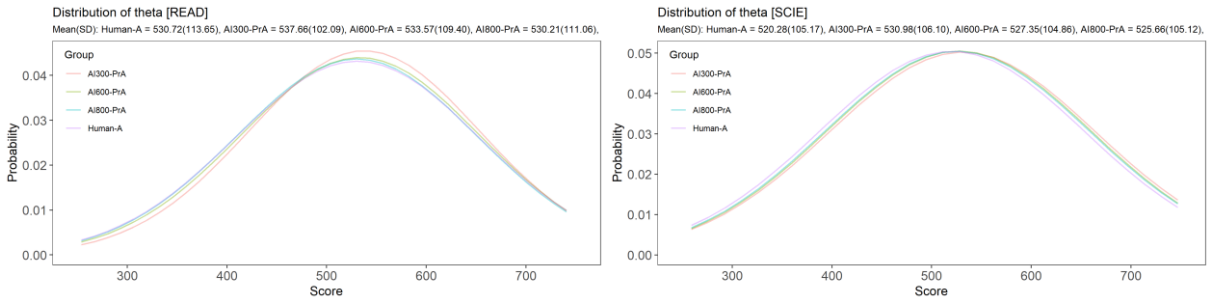


Figure 24: Score distributions of Human-B, AI300-PrB, AI600-PrB, and AI1000-PrB

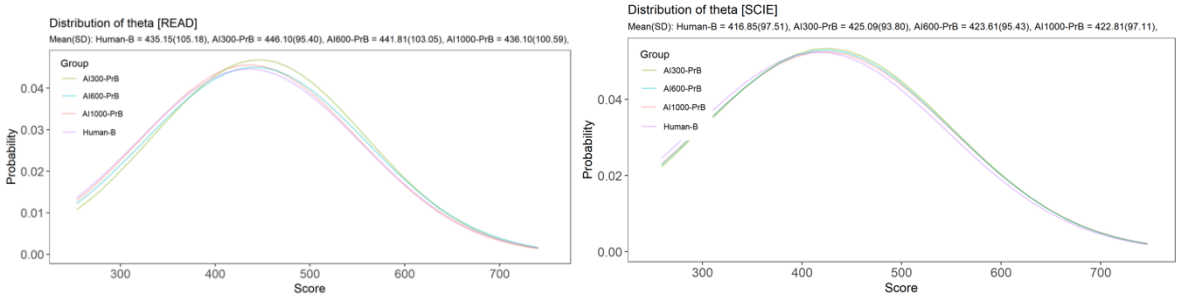


Table 7: Estimated score distributions of READ and SCIE using probabilistic score (PS) model

Condition	READ (Human)	READ (AI)	Difference (AI-Human)	SCIE (Human)	SCIE (AI)	Difference (AI-Human)
AI300-PrA	530.7(113.6)	537.7(102.1)	+7.0(−11.5)	520.3(105.2)	531.0(106.1)	+10.7(+0.9)
AI600-PrA		533.6(109.4)	+2.9(−4.2)		527.4(104.9)	+7.1(−0.3)
AI800-PrA		530.2(111.1)	−0.5(−2.5)		525.7(105.1)	+5.4(−0.1)
AI300-PrB	435.2(105.2)	446.1(95.4)	+10.9(−10.2)	416.9(97.5)	425.1(93.8)	+8.2(−3.6)
AI600-PrB		441.8(103.1)	+6.6(+1.4)		423.6(95.4)	+6.7(−2.1)
AI1000-PrB		436.1(100.6)	+0.9(−4.6)		422.8(97.1)	+5.9(−0.4)

5. Discussions

5.1. Can AI scoring be used in practice?

This study evaluated the accuracy of AI scoring using XLM-R deep learning technology. In the READ domain, when 600-800 responses from a language group were used in the AI training phase, the estimated score distribution by the AI-scored data was converged with the human-based IRT model distribution, even though a small number of items showed minor differential item functioning (DIF) by scoring modes (i.e., human scoring and AI scoring). This reveals that if the AI training dataset is large enough (i.e., exceeds 1 000 responses per item) and these responses include a wide range of response patterns, AI scoring performs similarly to human scoring, according to the measure of DIF used in this study (see Equation 3). In practice, AI scorings can be an option for constructed-response items in educational assessments, as long as the items do not show scoring-mode DIF. This delimitation prevents any unwanted bias in proficiency distributions.

In SCIE, the AI tended to overestimate scores for some items across conditions, although the size of these overestimates was small. This overestimation may have been caused by the nature of the 19 constructed-response SCIE items, the nature of the AI model employed in this study, or both. Alternatively, it may have been caused by the inconsistent human marking practices used to train the models. Further investigation is needed to clarify this point. Nevertheless, most of the SCIE items marked by AI functioned equivalently to human scoring, especially when the training datasets were relatively large (i.e., $N = 1000$). It is possible to mitigate this kind of bias by excluding any problematic items from AI scoring. Overall, this study has demonstrated that AI scoring can be used in practice as long as the careful psychometric analysis is performed for each item prior to model implementation. Consequently, this study provides strong evidence that accurate AI models can be successfully implemented for constructed-response items in large-scale assessments. The method described in this study is practical and provides accurate feedback on model function at the item level and across all proficiencies.

The index defined in Equation 3 should be used in the psychometric evaluation of the equivalence of AI and human scoring. The descriptive indices $\kappa_{gj}^{(H)}$(Equation 1) and $\kappa_{gj}^{(AI)}$(Equation 2) do not provide sufficient information to judge the equivalence of AI and human scoring. It is important to note here that a specific scoring-mode DIF threshold was not discussed in detail in this study, even though a threshold of 0.12 was used as a general reference, in accordance with the threshold used for human judgements in cognitive items in PISA across countries/economies (OECD, 2017[32]).

This study also demonstrated that language did not seem to be a factor that impacted the accuracy of AI models. As shown in Table 6, regardless of the training data language, the accuracies of the AI models in both READ and SCIE were closely comparable to that of human raters when the AI models were trained using 800-1 000 human-scored responses.

In the most conservative application of this scoring technology, AI scoring could be applied for quality control purposes, whereby human-scored responses are compared with AI-predicted scores, and where human remarking is recommended where discrepancies arise. This procedure can also be used to verify the stability of scoring standards between testing cycles, which is important given that one of the main interests in ILSAs is long-term trends. Indeed, any change in scoring standards introduces unexpected score biases and should be avoided.

It is obvious that cross-lingual language models will be continuously improved. Although XLM-R, which contains 550 million parameters, is used in this study, new cross-lingual pre-trained models are already published after the release of XLM-R. For example, Switch Transformer (Fedus, Zoph and Shazeer, 2021[33]) uses 1.6 trillion parameters and has shown better performances than the model that had been developed so far in various tasks frequently used in the NLP domain. It is natural to think that together with the improvements in the computation environment, cross-lingual language models will leap in future. From a test-management point of view, it is important to set a sustainable model-updating procedure for pre-trained language models, especially in ILSAs, so that scores scored by different models are comparable among testing cycles. Together with continuous refinement procedures for training data, model implementation and replacement rules should be designed before applying AI scoring for educational assessments.

5.2. Does multilingual data improve scoring accuracy?

Training the AI models using response data in multiple languages improved the accuracy of AI scoring. The utility of multilingual data was reported in section 3. Table 6 shows that even when the training data was limited to 360-600 responses per language, the AI scoring accuracy was reached at the human-scoring level. As such, the use of multilingual training data may enhance the applicability of AI scoring in ILSAs, in as much as small population countries/economies can also benefit from AI-based scoring.

Multilingual data also provide the opportunity to assess human scoring practices between countries/economies. At present, in the PISA-based Test for Schools, human scoring of constructed-response items is assessed through random spot-checks in each country/economy. In addition, constructed-response item functioning is assessed in the scaling phase; however, this analysis cannot pinpoint problematic scoring practices within specific countries/economies. A multilingual approach to AI model training would allow for the assessment of human scoring practices from an entirely different perspective. Specifically, it would allow for the comparison of human scoring in specific countries/economies with a baseline of AI predictions based on models built using datasets from all participating countries/economies. Discrepancies would indicate the need for the careful cross-checking of human scoring. Thus, in short, an important finding of this study is that multilingual data can be used in AI model training to support the international comparability and targeted intervention of human marking practices.

In ILSAs, the growing number of participating countries/economies has, at times, been considered a negative factor for the reliability and validity of assessments, despite empirical evidence to the contrary (Okubo, 2022[34]). The present study has demonstrated that an increase in the number of participating countries/economies will, instead, probably improve insights into human scoring practices within countries/economies via AI models trained using multilingual constructed-response data.

5.3. Should certainty of prediction be considered when estimating scores?

The more training data is used for AI, the more the certainty concentrates on 1.0, and therefore the impact of the probabilistic score dilutes the impact if there is sufficient training data. The comparison between Table 6 and Table 7 indicates no significant difference between the two models in terms of the point estimates of the mean, although the PS model for the small training data improved the accuracy of the estimation in READ of group-A. On the other hand, there is a clear difference in the estimated variance of the proficiency distributions, where the PS model estimates them smaller than the original IRT model. This is due to the change of shape of the log-likelihood function of each student, where the mode of the log-likelihood is biased towards the centre. Further psychometric properties of the PS model should be investigated with different datasets.

Although the PS model didn't show significant improvement compared to the original IRT model, it still can be considered a valuable model since it gives us an opportunity to mitigate critics of model choices of deep learning models and pre-trained datasets. The arbitrary choice of models and datasets can be avoided by taking an average of the predictions given by the different AI models and the pre-trained datasets, which is important from a validity perspective.

5.4. Future research and development

In this study, cognitive items of the PISA-based Test for Schools (PBTS) were analysed. The apparent next step in this research is to investigate the feasibility of AI scoring for other international large-scale assessments, such as PISA. Given that the current study demonstrated the utility of multilingual data for AI-based predictions in the PBTS, we expect higher accuracy in PISA, which includes even more countries/economies. Furthermore, if historical student responses to PISA trend items are sufficient, high accuracy is expected from the AI models. It is important to note, however, that this study also found that some items cannot be accurately predicted by AI. Therefore, in future practice, it will be essential to identify these kinds of items and ensure that they continue to be scored by humans. It may also be beneficial to investigate the substantive reasons why these items cannot currently be predicted by AI.

Defining procedures for AI scoring is indispensable work if we are to maximise the efficiency in scoring constructed-response items. These AI models should be continuously updated with additional training data that is robust from a psychometric perspective. To achieve this, the workflow of human

scoring practices should be analysed and discussed, and a system that supports greater consistency in human scoring should be carefully designed. Human scoring protocols that foster consistent scoring across all markers, such as the two-stage pairwise model (Humphry and Heldsinger, 2020[35]), may significantly improve the quality and AI training data and, in turn, the predictive accuracy of AI models.

Obviously, AI scoring can be implemented using multi-stage testing (MST) in order to increase the efficiency of score estimation at each stage of calibration and the validity of item assignments. Branching students based on selected-response items may include unexpected errors because of unparameterised item characteristics such as differences in response-format functioning between subgroups, such as gender. In addition to this, using constructed-response items for provisional score estimations will provide us with more options for booklet design.

Finally, in ILSAs, both cognitive items and student responses regarding occupation are coded by human markers. For example, in PISA and PBTS, student responses to parent occupation questions are coded in accordance with the international standard classification of occupations (International Labour Office, 2012[36]), as defined by the International Labour Organisation (ILO). The coding accuracy here does not contribute to the point estimate of country/economy mean and variance of proficiency distributions; nonetheless, many human resources and financial costs are expended for this task alone in each country/economy since it affects student's ESCS index (PISA measure of economic, social and cultural status), which is widely used in various studies. As such, this international standard classification of occupations (ISCO) coding task is the next target that we plan to work on. Automating ISCO coding (Gweon et al., 2017[37]) will reduce financial costs in the future.

References

ACER (2012), *PISA-based Test for Schools: Technical Report*, Australian Council for Educational Research (ACER), https://www.oecd.org/pisa/aboutpisa/PISA-based%20Test%20for%20Schools%20Technical%20Report%20-%20ACER%202012.pdf (accessed on 1 October 2022). [26]

Baker, F. and S. Kim (2004), *Item Response Theory: Parameter Estimation Techniques. 2nd Edition*, CRC Press, Boca Raton. [31]

Bloom, B., T. Hastings and G. Madaus (1971), *Handbook on formative and summative evaluation of student learning*, McGraw-Hill, New York. [5]

Brennan, R. (2001), *Generalizability theory*, Springer-Verlag, New York. [13]

Butler, D. and P. Winne (1995), "Feedback and self-regulated learning: A theoretical synthesis", *Review of Educational Research* 65, pp. 245-281. [6]

Conneau, A. et al. (2019), "Unsupervised Cross-lingual Representation Learning at Scale", *arXiv:1911.02116*, pp. 1-12, https://arxiv.org/abs/1911.02116. [25]

Crooks, T. (1988), "The impact of classroom evaluation practices on students", *Review of Educational Research*, Vol. 58, pp. 438-481. [1]

Dempster, A., N. Laird and D. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *ournal of the Royal Statistical Society, Series B.*, Vol. 39, pp. 1-38. [30]

Devlin, J. et al. (2018), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv:1810.04805*, pp. 1-16, https://arxiv.org/pdf/1810.04805.pdf. [22]

Downing, S. (2003), "Guessing on selected-response examinations", *Medical Education*, Vol. 37, pp. 670-671. [9]

Fedus, W., B. Zoph and N. Shazeer (2021), "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity", pp. 1-40, https://arxiv.org/pdf/2101.03961.pdf. [33]

Frederiksen, J. and A. Collins (1989), "A systems approach to educational testing", *Educational Researcher*, Vol. 18, pp. 27-32. [7]

Gweon, H. et al. (2017), "Three Methods for Occupation Coding Based on Statistical Learning", *Journal of Official Statistics*, Vol. 33, pp. 101–122. [37]

Haladyna, T. (2004), *Developing and validating multiple choice items*, Lawrence Erlbaum, Mahwah, New Jersey. [10]

Halpin, G. and G. Halpin (1982), "Experimental investigations of the effects of study and testing on student learning, retention, and ratings of instruction", *Journal of Educational Psychology*, Vol. 74, pp. 32-38. [3]

Harkins, S. (2001), *Multiple perspectives on the effects of evaluation on performance: Toward an integration.*, Kluwer Publishers, New York. [2]

Hu, J. et al. (2020), "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization", *arXiv:2003.11080*, pp. 1-20, https://arxiv.org/pdf/2003.11080.pdf. [21]

Humphry, S. and S. Heldsinger (2020), "A Two-Stage Method for Obtaining Reliable Teacher Assessments of Writing", *Frontiers in Education*, Vol. 5, pp. 1-10. [35]

International Labour Office (2012), *International Standard Classification of Occupations: ISCO–08*, International Labour Office, https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf. [36]

Jolly, B. (2010), *Written Examinations*, Wiley-Blackwell, London. [12]

Leacock, C. and M. Chodorow (2003), "C-rater: Automated scoring of short-answer questions", *Computers and the Humanities*, Vol. 37, pp. 389-405. [19]

Liu, Y. et al. (2019), "RoBERTa: A Robustly Optimized BERT Pretraining Approach", pp. 1-13, https://arxiv.org/pdf/1907.11692.pdf. [23]

Lord, F. and M. Novick (1968), *Statistical theories of mental test scores*, Addison-Wesley, Menlo Park. [29]

McArthur, C., S. Graham and J. Fitzgerald (eds.) (2006), *Applications of computers in assessment and analysis of writing*, Guilford Publications, New York. [16]

Messick, S. (1989), *Validity*, American Council on Education/Macmillan, New York. [8]

Nungester, R. and P. Duchastel (1982), "Testing versus review: Effects on retention", *Journal of Educational Psychology*, Vol. 74, pp. 18-22. [4]

OECD (2019), *PISA 2018 Assessment and Analytical Framework*, OECD Publishing, Paris. [18]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris. [32]

OECD (2016), *PISA-based Test for Schools Technical Report 2016*, OECD Publishing, Paris. [28]

OECD (n.d.), *PISA 2018 Technical Report*, https://www.oecd.org/pisa/data/pisa2018technicalreport/ (accessed on 1 October 2022). [14]

Okubo, T. (2022), "Theoretical considerations on scaling methodology in PISA", *OECD EDU Working Papers*, Vol. 282, pp. 1-31. [34]

Okubo, T. et al. (2021), "PISA-Based Test for Schools: International Linking Study 2020", *OECD EDU Working Papers*, Vol. 244, pp. 1-37. [27]

Raffel, C. et al. (2020), "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research, arXiv:1910.10683*, Vol. 21, pp. 1-67, https://arxiv.org/pdf/1910.10683.pdf. [24]

Ramesh, D. and S. Sanampudi (2022), "An automated essay scoring systems: a systematic literature review", *Artificial Intelligence Review*, Vol. 55, pp. 2495-2527. [17]

Stalnaker, J. (1951), *The essay type of examination*, George Banta, Menasha. [11]

Vaswani, A. et al. (2017), "Attention Is All You Need", pp. 1-15, https://arxiv.org/pdf/1706.03762.pdf. [20]

Yamamoto, K. et al. (2018), "Development and Implementation of a Machine-Supported Coding System for Constructed-Response Items in PISA", *Psychological Test and Assessment Modeling*, Vol. 60, pp. 145-164. [15]