Investigating the Psychometric Properties of Students' Perceived Teaching Practices Scales of PISA: A cross-cultural comparison between Indonesia and the Netherlands

Name: Kalvin Rongre Tandigego

Student number: S5797950

First supervisor: dr. V.E.C. (Letty) Koopman

Second supervisor: dr. J.J. (Jelle) Sijtsema

Word count: 7904

University of Groningen

Faculty of Behavioral and Social Sciences

Master Educational Sciences

Track: Learning in Interaction

Date: 12 January 2025

Abstract

This study investigates the psychometric properties of Students' Perceived Teaching Practices (SPTP) scales from the PISA 2015 dataset, focusing on a cross-cultural comparison between Indonesia and the Netherlands. By employing Mokken Scale Analysis (MSA), a non-parametric item response theory method, this research evaluates the internal structure, scalability, item and scale score, and item difficulty order of these scales across the two countries. Findings indicate that while the SPTP scales exhibit a generally similar structure, significant differences exist in scalability, item and scale score distribution, and item difficulty order, with the Netherlands demonstrating more robust psychometric properties. These disparities highlight the influence of cultural and geographical diversity on students' perceptions of teaching practices. The results emphasize the need for culturally responsive approaches in international educational assessments to ensure valid and equitable comparisons. This study underscores the challenges and considerations essential for adapting educational measurement to reflect diverse educational systems and cultural values, advocating for a more nuanced understanding of cross-cultural variations in student perceptions.

Keywords: PISA 2015, Mokken Scale Analysis, psychometric properties, teaching practices, cross-cultural comparison, Indonesia, Netherlands.

Introduction

Background on Teaching Practices

Teaching practices are central to understand how the teacher's knowledge, skills, and abilities create meaningful educational experiences for students. These practices encompass emotional support, classroom organization and management, and instructional support (Hamre et al., 2013). Emotional support involves creating a positive classroom climate through empathy, encouragement and fostering a sense of belonging among students, which boosts student engagement and motivation (Connel & Wellborn, 1991). Classroom organization and management include structuring the environment, managing transitions, and setting clear expectations. Effective classroom management practices, including clear expectations and proactive strategies, contribute to a positive learning environment and enhance student focus (Oliver et al., 2011). Hamre et al. (2013) defined instructional support as the quality of teacher-student interactions related to academic content, providing feedback, engaging students in meaningful discussions, and improving engagement and learning outcomes (Pianta et al., 2012).

Student ratings have been widely employed in many studies evaluating teaching practices, providing valuable insight into teaching practices. However, the reliability and validity of student ratings can vary across different scales and countries, and can differ at the student level and school level (Aditomo & Köhler, 2020). To enable accurate comparisons, it is important first to understand how they are collected and evaluated. Therefore, investigating the reliability and validity of measurement instruments used to gather student ratings is crucial to ensure that the comparisons made are meaningful and accurate.

Cross-Cultural Variability in Education

Learning is inherently a social process, and it will be meaningful when its processes are integrated with individuals' cultural contexts, as proposed by Vygotsky's Sociocultural theory (Tzuriel, 2021). In other words, learning is a deeply context-dependent process, emphasizing that it can vary across contexts based on the culture in which the individual is situated. Similarly, in his Ecological System theory, Bronfenbrenner (1977) posited culture along with social, economic, and political contexts at the macrosystem level as a broader factor that indirectly yet significantly affects a child's development. The revised version of Bronfenbrenner's Ecological Systems theory by Vélez-Agosto et al. (2017) even emphasizes the role of culture in a child's development. In this view, culture is not just a backdrop but an active force that operates at all levels in shaping how children engage with the environment and perceive their learning experiences.

The perception of teaching practices can vary across different contexts and cultures. Culture influences how people think, perceive and communicate, all of which affect how teachers teach and how students learn, and these cultural norms and practices are transmitted to school by teachers, shaping their teaching practices (Chafi, 2017; Obanya, 2005, as cited in Abdessalam, 2020). What teachers decide to do or not in their classrooms is affected by school cultures like teachers' association with the school policies, tradition, structures, and the interaction among the teachers themselves (Rosenholtz, 1991; Bidwell & Kasarda, 1980; Hargreaves, 1994). For instance, a study by Hofstede (1986) highlights that cultural dimensions, such as power distance, individualism versus collectivism, and uncertainty avoidance, significantly influence how classroom management is perceived and executed. In cultures with high power distance, teachers are often seen as authoritative figures and strict classroom management is expected. In contrast, cultures with lower power distance may

encourage more democratic and participative classroom environments. Generally, high power distance is often found in Asian and African countries, while low power distance is typically seen in Europe, North America, and Oceania.

In teaching practices, cultural differences are pronounced between Western and Eastern countries. Bryer and Beamish (2019) mentioned that Western teaching practices, such as in the Netherlands, prioritize open dialogue and interaction among students to encourage students to engage in discussions to develop critical thinking. Additionally, active participation is strongly emphasized, where students actively participate in their learning through group activities and peer collaboration. Conversely, Fang and Gopinathan (2009) highlighted that in Eastern countries like Indonesia, the hierarchical structure within the classroom reflects broader cultural values where teachers are viewed as authoritative figures who lead the classroom activities and do most of the talking, while students are expected to be passive recipients of knowledge. As a result, students may behave more subduedly, preferring to listen attentively rather than confront or question their teachers directly. However, it is essential to recognize that these observations reflect broad cultural tendencies, and there is significant diversity within both Western and Eastern educational systems.

Geographical composition and linguistic diversity in Indonesia and the Netherlands differ significantly, influencing respective educational strategies and teacher training strategies. The archipelagic nature of Indonesia, with its diverse ethnic groups and languages, poses significant challenges to maintaining consistent and high-quality teacher education across the country and the implementation of new approaches in the national curriculum (Effendi-Hasibuan et al., 2019: Novita, 2022). To address these issues, the Ministry of Education in Indonesia, through Law No. 14/2015, has implemented a certification program for all teachers (De Ree, 2015). A four-year university degree, a teaching certificate, and the

demonstration of professional, pedagogical, personal, and social competences are the general prerequisites for becoming a teacher in Indonesia, according to the law (Law No. 14 of 2005, n.d.).

In contrast, the Netherlands is primarily a single, contiguous piece of land in Western Europe, with Dutch as the dominant language spoken by most of its population. The linguistic diversity in the Netherlands is relatively limited compared to Indonesia. The Dutch education system is known for its proactive approach to changing contexts, crucial in identifying effective teaching practices that align with evolving educational needs (Education Policy Outlook, 2020, p. 213). It adopts a bottom-up approach to address organizational challenges, where project leaders meticulously analyze challenges, propose solutions, and take concrete steps to design and implement educational innovations (Schophuizen, 2020). Dutch teachers actively participate in Continuous Professional Development (CPD) to enhance their skills and adapt to changing contexts. These approaches incorporate culturally responsive pedagogies, respecting local customs and linguistic diversity. With its diverse student population, the Netherlands maintains its robust education system through continuous assessment of teacher effectiveness. To become a teacher in the Netherlands, a bachelor's degree in education or a degree in a specific subject area combined with a teaching qualification is needed. For elementary school teachers, specialized education tailored to that context, known as PABO (Pedagogische Academie voor het Basisonderwijs), is required. Proficiency in Dutch is also essential for most schools, and a certificate of good behavior is also needed (VOG) (Ministerie van Algemene Zaken, 2021).

Indonesia and the Netherlands have been included in some teaching practice comparison studies. A study by André et al. (2020) which investigated student perception of teaching behavior across six countries, including Indonesia and the Netherlands, found that

while the measurement of teaching behavior is adequately invariant across countries, students in Indonesia perceived teaching behavior to be lower compared to students in the Netherlands. The same pattern emerged in the study by Maulana et al. (2020) using the International Comparative Analysis of Learning and Teaching (ICALT) framework highlights the cross-national variations in teaching behavior between several countries including Indonesia and the Netherlands. While the results demonstrated that the ICALT showed full strict invariance in both countries, the Netherlands exhibited higher teaching behavior than Indonesia, underscoring the stark differences in their cultural and educational context. These persistent disparities across studies provide a clear contrast in how teaching behaviors are influenced and perceived, which is crucial for developing teaching strategies that are culturally responsive and effective across diverse educational settings.

Therefore, as previously discussed, the pronounced differences between Indonesia and the Netherlands in their cultural, linguistic, and educational contexts and disparities in some studies provide a compelling basis for this study. These differences uniquely posit Indonesia and the Netherlands in exploring cultural factors that influence students' perceptions of teaching practices, contributing valuable insights into the cross-cultural validity and applicability of educational assessments.

PISA and its Role in Assessing Students' Perceived Teaching Practices

The Organization for Economic Co-operation and Development (OECD) conducts the Program for International Student Assessment (PISA), a global research program that measures the reading, mathematics, and science proficiency of 15-year-old students to assess educational systems, carried out every three years. It has been recognized as a crucial instrument for evaluating educational results and instructional strategies worldwide. The 2015 cycle was the sixth edition of PISA since its beginnings.

The results of PISA offer insights into teaching practices quality through various scales, including Disciplinary Climate in Classes, Inquiry-based Teaching and Learning Practices, Teacher Support in Classes, Teacher-directed Instruction, Perceived Feedback, and Adoption to Instruction. These scales align with Hamre's framework (Hamre et al., 2013), covering emotional support, classroom organization, and instructional support. It is then possible to classify international teaching practices within this model. Specifically, PISA's Teacher Support on Class scale reflects emotional support, while the Discipline Climate in Classes scale measures classroom organization and management. Meanwhile, instructional support is captured through three scales: Adaption to Instruction, Teacher-directed Instruction, Perceived Feedback, and Inquiry-based Teaching and Learning Practices (see Appendix A).

Psychometric Challenges in Cross-Cultural Comparisons

Unlike directly observable physical traits, measuring psychological characteristics, such as emotions, perceptions or behaviors, is arguably more complex. This is because these psychological properties are abstract and not easily measurable through direct observation. Moreover, there is not a single method universally accepted as the way to measure them; instead, psychological measurement often relies on indirect methods, like questionnaires and tests, which can introduce complexity and measurement error (Sijtsma & Van Der Ark, 2020). Other challenges may also arise from different perspectives, such as cultural bias in test design, issues related to language and translation, and construct validity across cultures (Brodin et al., 2010).

PISA test items are developed in certain countries and might reflect cultural assumptions from those countries. This potentially disadvantages students from different educational values and curricula. Moreover, as an international assessment utilizing

questionnaires, PISA's questionnaires are translated into the languages of the participant countries, such as Indonesia and the Netherlands. Van De Schoot et al. (2012) suggested that the test results cannot automatically be assumed to be valid for the new population. Therefore, it is necessary to investigate whether the test still accurately measures what it is intended to by carefully examining the psychometric properties of the test in the new population.

Potential differences in interpreting particular concepts can add complexity to cross-cultural comparisons. For instance, the concept of “feedback”, measured by PISA in teaching practice questions, may be interpreted differently across countries. In some countries, feedback is perceived positively as constructive comments for improvement. However, due to variations in translation and contextual interpretation, it may be viewed as a personal critique rather than a professional tool for growth. This variation in interpretation can influence how students respond to the questionnaire items and potentially affect the validity of cross-country comparisons.

When comparing teaching practices across such diverse cultural settings, the underlying assumption is that measurement invariance holds – the assumption that the same factors or scales operate similarly across groups (countries) and that the construct measures the same in different countries. This assumption may not hold in cross-cultural contexts where language, cultural norms or educational practices differ significantly. It is crucial in psychometrics and social science research to ensure meaningful and valid comparisons between groups. Without measurement invariance, comparisons between groups could be biased due to the instrument's different functions across groups, leading to misleading conclusions. This concept is essential for accurate and valid psychometric and social science research (Van De Schoot et al., 2015).

Despite PISA's strict measures to reduce cross-cultural bias, it is not possible to assume that its scales are measurement invariant (Aditomo et al., 2020). A valid assessment should account for cultural variations to enhance the accuracy of the instruments and avoid bias for a specific culture. The item response theory (IRT) framework is one strategy to evaluate measurement invariance. PISA has employed parametric item response theory (IRT) analysis across all countries simultaneously and separately for each scale, showing no significant deviations for Indonesia and the Netherlands. However, employing more flexible non-parametric IRT methods, like the Mokken Scale Analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002), may provide deeper insights into the cross-cultural differences in the measurement of teaching practices.

MSA is a nonparametric method well-suited for cross-cultural research, offering greater flexibility than parametric IRT. Unlike IRT, MSA does not impose strict assumptions about the relationship between items and constructs, making it ideal for exploring the fit of theoretical models underlying item responses in different cultural contexts. This method is particularly valuable for assessing the consistency of the internal structure – how items within scales interrelate across Indonesia and the Netherlands and whether these relationships are maintained across diverse educational settings. MSA also assesses the fit of the measurement model to determine whether the scales function equivalently, supporting cross-cultural comparisons. Additionally, it analyzes scale and item scores distributions to explore whether perceptions and behaviors are similar across the two countries. Finally, MSA can pinpoint specific items where the overall difficulty order is violated, providing insights into potential bias and ensuring the fairness and validity of educational assessments. Through these analyses, MSA ensures that cross-cultural teaching practice comparisons are rigorous and meaningful.

An exemplary implementation of MSA is found in the study on the cross-cultural validity of the WHO-5 Well-Being Index and Euthymia Scale (Carrozzino et al., 2022). This study employed MSA to evaluate the psychometric properties of these scales across various cultural contexts. The findings underscored the importance of meticulous scaling when conducting cross-cultural assessments to ensure valid and reliable measurements. This underscores the value of using MSA in this study to investigate whether PISA's teaching practice scales truly measure the same construct in Indonesia and the Netherlands. Besides, this study found that MSA is more adaptable in identifying subtle differences across cultural contexts than traditional parametric methods like Item Response Theory (IRT).

Research Purpose and Questions

Our study aims to investigate the measurement invariance of the Students' Perceived Teaching Practice (SPTP) scales in Indonesia and the Netherlands to determine whether these scales can be used for comparison of perceived teaching practices across Indonesia and the Netherlands.

To make claims about this research goal, the following specific research questions will be answered:

1. Is the internal structure of the scales comparable across Indonesia and the Netherlands?
2. Does the measurement model fit the scales equally well across Indonesia and the Netherlands?
3. Are scale score and item scores similarly distributed across Indonesia and the Netherlands?
4. Are there specific items for which the overall difficulty order is violated in Indonesia and the Netherlands?

Methodology

Procedure

Participants

In this cross-cultural comparison study, we analyzed data from PISA in 2015, focusing on 15-year-old students from Indonesia and the Netherlands. The dataset comprised 6513 students from Indonesia and 5385 students from the Netherlands. According to PISA 2015 Technical Reports (PISA 2015 Database, n.d.-b), participants were selected through a two-stage stratified systematic sampling method. Initially, schools were chosen from a nationwide list based on proportional enrolment of 15-year-olds, ensuring a representative distribution across various geographic and socioeconomic status. Within selected schools, either 42 students were randomly selected or all eligible students were included if fewer than 42 were eligible. A target response rate of 85% for schools and a minimum of 65% of originally sampled schools were necessary for data inclusion, guaranteeing a representative sample of 15-year-old students. Replacement schools were also identified to ensure enough participation. The PISA 2015 Database provided anonymized data, ensuring confidentiality and compliance with ethical standards for educational research (PISA 2015 Database, n.d.-b).

Materials

The study utilized scales from the PISA 2015 questionnaire that assess perceived teaching practices. The international consortium of experts from the Organisation for Economic Co-operation and Development (OECD), along with input from national advisory groups from participating countries, designed the scales in PISA 2015. The design process involved extensive field trials and validation studies to ensure the questions were appropriate and reliable across different contexts. These scales aim to measure students' perceptions of

teaching dimensions in PISA 2015 (OECD, 2017) (see Appendix A for a complete overview of the scales and their items), including:

1. **Disciplinary Climate in Classes** – consists of five items that evaluate how well teachers maintain order, discipline, and a positive learning environment in the classroom.
2. **Inquiry-based Teaching and Learning Practices** – consists of nine items assessing the extent to which teachers encourage students to explore, question, and think critically.
3. **Teacher Support in Classes** – forms in five items to measure the emotional support and encouragement teachers provide to students.
4. **Teacher-directed Instruction** – includes four items that aim to gauge the degree to which teachers lead and guide classroom activities.
5. **Perceived Feedback** – consists of five items examining how effectively teachers provide feedback to students on their performance and progress.
6. **Adaption to Instruction** – consists of three items focusing on teachers' ability to tailor their teaching methods to meet individual student needs.

The questionnaire used the 4-point Likert Scale, with 1 for "Every Lesson," 2 for "Most Lesson," 3 for "Some Lesson," and 4 for "Never or Hardly Ever." This applies to Scales of 1 to 4. However, for Scale 5, the items were reverse coded. Additionally, the Adaption to Instruction scale was excluded from the analysis because the corresponding questions related to this scale were not administered in Indonesia (OECD, 2017).

Data Analysis

Mokken Scale Analysis

In this study, we employed the Mokken Scale Analysis (MSA) using R studio to evaluate the scalability and dimensional structure of teaching practices perceived by students in Indonesia and the Netherlands. MSA, a set of methods based on non-parametric item response theory models, is well-suited for cross-cultural comparisons due to its ability to assess various item and test characteristics (e.g., Palmgren et al., 2018).

Scalability Coefficients

A key element of Mokken scale analysis is the scalability coefficients, which are useful for diagnosing model fit and evaluating discriminatory power (Sijtsma & Molenaar, 2002). There are three types of scalability coefficients:

1. **Item-Pair Scalability Coefficient (H_{ij})**

   H_{ij} assesses the relationship strength between two items, indicating whether high scores on one item predict high scores on another. A high H_{ij} value confirms that the item pair is consistent in measuring the underlying construct, which is crucial for evaluating the pairwise compatibility of scale items.

2. **The Item Scalability Coefficient (H_i)**

   H_i measures how well an individual item correlates with the total score of the rest of the scale, providing insight into an item's integration and significance within the scale. High H_i values suggest strong item contribution to the scale's overall construct.

3. **Total Scale Coefficient (*H*)**

   The *H* coefficient evaluates the overall internal consistency of a scale, reflecting how cohesively all items measure a single latent trait. Higher *H* values indicate a robust, reliable scale.

The relationships among these coefficients are defined as follows:

$$\min(H_{ij}) \leq \min(H_i) \leq H \leq \max(H_i) \leq \max(H_{ij})$$

A set of items is considered a Mokken scale if the following two criteria, are satisfied:

1. $H_{ij} > 0$ for all item pairs,
2. $H_i \geq c$ for all items –, with c being a positive lower bound, commonly set at 0.3.

Automated Item Selection Procedure (AISP)

The Automated Item Selection Procedure (AISP) is a statistical method employed to identify the best set of items that form a Mokken scale while adhering to the criteria of a Mokken scale (Straat et al., 2013; Koopman et al., 2022). The procedure specifically investigates the internal structure of the SPTP scales, focusing on relationship between items and their alignment with the underlying constructs, to determine whether items are consistently grouped to the same scale across varying scalability thresholds in both countries. The AISP provides critical insights into whether the internal structure of SPTP scales of PISA is comparable across Indonesia and the Netherlands, which is crucial for confirming the validity of cross-cultural comparisons made using these scales.

The steps of an AISP are initially selecting the first two items in the scale that meet the criteria of a Mokken scale and have the highest pair scalability coefficient (H_{ij}). Then, it

adds subsequent items as long as the criteria of the Mokken scale are satisfied. This step continues until no more things can be added that meet both Mokken scale criteria or until no more items remain. Then, to create a subsequent scale using just the unselected items, the AISP goes back to the initial step. The AISP comes to an end if there are no more items or item pairings that meet the Mokken scale criterion. The allocation process is based on statistical properties rather than content alone. Therefore, depending on their scalability coefficient, the same items may or may not be allocated to the same scales in different analyses. Ideally, the items within a scale should represent a coherent content domain. However, since the primary focus of AISP is on scalability and measurement properties, the researchers must ensure that content validity is maintained by carefully designing and selecting the items.

It is possible to repeatedly apply the AISP with increasing lower bound *c* values - a predefined threshold that determines the minimum value of scalability coefficients (e.g., H_{ij}, H_i, and H) necessary for an item to be considered suitable for inclusion in a scale. The requirement for item scalability gets stricter as c increases. In the sample being studied, stricter requirements result in shorter scales with higher discrimination power. It is also advised to use an increasing value of *c* when employing AISP to investigate whether a set of items forms one or more Mokken scales (Hemker et al., 1995).

Analytical Strategy

For the entire analysis, we will combine the measurement invariance procedure by Molenaar and Sijtsma (2000, pp. 88-94) with part of the 10-step MSA procedure by Sijtsma and Van der Ark (2017) (see Appendix B for the complete steps) to evaluate the scalability and structure of the SPTP scales for Indonesia and the Netherlands. The initial step is Step 2 from

the 10-step MSA procedure by Sijtsma and Van der Ark (2017), focusing on removing inadmissible scores and managing missing data.

The group comparison (Step 10) integrates the methodologies from both Molenaar and Sijtsma (2000) and Sijtsma and Van der Ark (2017), particularly focusing on scalability (Step 4), local independence (Step 5), monotonicity (Step 6), and invariant item ordering (Step 7). The whole analysis used the mokken Package in R software (Van der Ark, 2007;2012). The following steps were performed per country.

Step 1: Data Examination

This initial stage includes checking and removing the inadmissible scores and missing data. Given the large sample size, it is possible that the data are not completed. To handle this issue, we will employ listwise deletion, using only complete dataset, under the assumption that the missing data is missing at random. As MSA relies on item-level correlation and scalability coefficients, using only complete data for all items can maintain the integrity of the calculations and results.

Step 2: Scale Identification

The first step with the completed dataset is to check the internal structure of the PISA items by running the Automated Item Selection Procedure (AISP) to get an illustration of how the items perform in both countries along with the possible pairs in one scale at a certain threshold. This step aims to answer the following research question:

1. Is the internal structure of the SPTP scales comparable across Indonesia and the Netherlands?

Step 3: Scalability of the Items and Scales

The scalability will be evaluated based on Loevinger's coefficient, where (H) ≥ 0.3 is the minimum acceptable value, indicating a weak scale; $H \geq 0.4$ indicates a moderate scale; and $H \geq 0.5$ indicates a strong scale. If the model does not fit equally well, where a significant difference in scalability emerges, it could indicate that the construct is not perceived similarly across both countries. However, poor fit in both countries does not necessarily indicate that the fit is equal. Equally poor H coefficients could signal underlying different issues, like varying misinterpretations or item relevance.

Additionally, the 95% confidence intervals of the H for both countries are calculated to determine if there are statistically significant differences in scalability. If the intervals do not overlap, it suggests a statistically significant difference, indicating that the scale may not measure the construct equivalently in both countries. This calculation involves the following formula:

$$H \pm 1.96 \times \text{SE (standard error)}$$

The following research question is answered in this step by employing the guidelines of Mokken Scale Procedure (MSP) provided by Molenaar and Sijtsma (2000):

2. Does the measurement model fit equally well across subgroups (Indonesia and the Netherlands)?

Step 4: Scale and Item Score Distribution

This step involves inspecting the table distribution of the total and individual item scores for both countries. Descriptive statistics like mean and standard deviation are being compared. If the scores are distributed differently, this might suggest that the same scale or items do not

function equivalently across cultural contexts, potentially affecting the validity of cross-comparison. To statistically validate these descriptive insights, the 95% confidence intervals for the means of each item and the total scale are calculated using the formula:

$$\bar{X}_i \pm 1.96 \times \sqrt{\frac{SD}{n}},$$ where n is sample size for each scale.

For this step, we used the guidelines of Mokken Scale Procedure (MSP) provided by Molenaar and Sijtsma (2000) that can help to answer the following research question:

3. Are scale score and item scores similarly distributed across Indonesia and the Netherlands?

Step 5: Item Order within Scales

This step checks whether the relative difficulty of items remains consistent across Indonesia and the Netherlands, a critical aspect of ensuring fair and valid cross-cultural comparison. A violation of monotonicity occurs when an item considered easier in one country becomes harder relative to other items in another country, suggesting that the order of item difficulty is inconsistent between both countries. Monotonicity is a foundational assumption in scaling that an increase in the underlying trait should correspond to an increase in item responses. Violations of this assumption can compromise the fairness of the test and, if severe, may hinder direct comparisons of scores across the two countries, as the items may not measure the construct equivalently. Such discrepancies suggest that the items may be influenced by cultural or educational differences that affect how they are perceived or understood by students in each country.

The guidelines of the Mokken Scale Procedure (MSP) provided by Molenaar and Sijtsma (2000) will be used to evaluate the potential violations and determine the severity level of the violation, as follows:

1. $Crit > 80$, indicates significant model violation, suggesting the need for item removal or revision.
2. $40 \leq Crit \leq 80$, indicates ambiguous evidence of violation, requiring further investigation to assess the impact on the overall measurement.
3. $Crit < 40$, indicates minor violation, typically due to sampling variation and the items are usually retained.

The guidelines also are used in this step to answer the following research question:

4. Are there specific items for which the overall difficulty order is violated in Indonesia and the Netherlands?

Results

Step 1: Data Examination

For the first step, participants with incomplete data were removed from the dataset. After removing the incomplete data, 3646 students from the Netherlands and 5826 from Indonesia remained. This represented a reduction of approximately 32.3% and 10.5% of the initial participants from the Netherlands and Indonesia, respectively. The following step of checking the inadmissible scores for each item for both countries shows that the minimum score for each is one, and the maximum is four for both countries, which means there were no inadmissible scores.

Step 2: Scale Identification

The Automated Item Selection Procedure (AISP) analysis (Appendix C) shows that the original Students' Perceived Teaching Practices (SPTP) scales of PISA are comparable across Indonesia and the Netherlands. The analysis shows that most items are consistently grouped into three primary scales when applying the lower, acceptable scalability threshold of 0.3. Both tables inform that in the first column, all items assigned to the first scale have a value of 1, and for the second scale, have a value of 2 until all items are grouped. However, slight variations are observed: in Indonesia (Table 1), while most items are grouped into three scales at the lower scalability threshold of 0.3, some variation occurs with items extending up to six scales. Meanwhile, in the Netherlands (Table 2), items are predominantly grouped into three scales, with a few extending to a fourth scale. Despite these differences, the overall structure remains fundamentally similar and adequately captures the dimensions of teaching practices in both countries, suggesting that these scales can be relied upon to assess students' perceptions of teaching practices in both countries.

If a lower scalability threshold of 0.3 were accepted, combining certain items into new, broader scales would be possible. For instance, Inquiry-based Teaching and Learning Practices and Teacher Support in Classes could be combined to create a new scale focusing on teacher support and guidance in inquiry-based learning. Similarly, Teacher-directed Instruction and Perceived Feedback items could be merged to form a new scale focused on student engagement through teacher feedback and interaction. However, combining these scales could affect the granularity of measurement, resulting in blurred interpretation of specific teaching practices in each country.

Step 3: Scalability of the Items and Scales

Scale 1: Discipline Climate in Classes

After removing the missing items, the analysis included 6231 students from Indonesia and 4296 students from the Netherlands. The scalability across both countries exceeds the acceptable threshold (0.3), with Indonesia showing moderate scalability at 0.450 (95% CI: 0.434 to 0.466) and the Netherlands demonstrating good scalability at 0.650 (95% CI: 0.632 to 0.668). All items exhibit strong scalability in the Netherlands, with over 0.5, while most items are moderately scalable in Indonesia. For Indonesia, the highest scalability was observed in item 2, with 0.477; in the Netherlands, item 4, with 0.697, was the item with the highest scalability. The order from the highest to the lowest scalability item was similar, especially for item 5, which was the lowest scalability item for Indonesia and the Netherlands, with 0.383 and 0.549, respectively. The confidence intervals of each individual item and the total scale score did not overlap, indicating a statistically significant difference between the two countries. These results suggested that the measurement model did not fit equally well across Indonesia and the Netherlands.

Table 3

Item Mean and Scalability across Indonesia and the Netherlands

Items	$\bar{X}_i$ *(SD)*		H_i *(SE)*	
	IDN	NL	IDN	NL
1	3.01 (0.77)	2.90 (0.73)	0.453 (0.010)	0.626 (0.011)
2	2.81 (0.85)	2.72 (0.73)	0.477 (0.010)	0.687 (0.010)
3	3.04 (0.93)	2.77 (0.78)	0.469 (0.009)	0.655 (0.009)
4	3.25 (0.80)	3.11 (0.74)	0.466 (0.009)	0.697 (0.011)
5	3.31 (0.79)	2.73 (0.80)	0.383 (0.011)	0.549 (0.011)
Total Scale	15.41 (2.99)	14.23 (3.09)	0.450 (0.008)	0.650 (0.009)

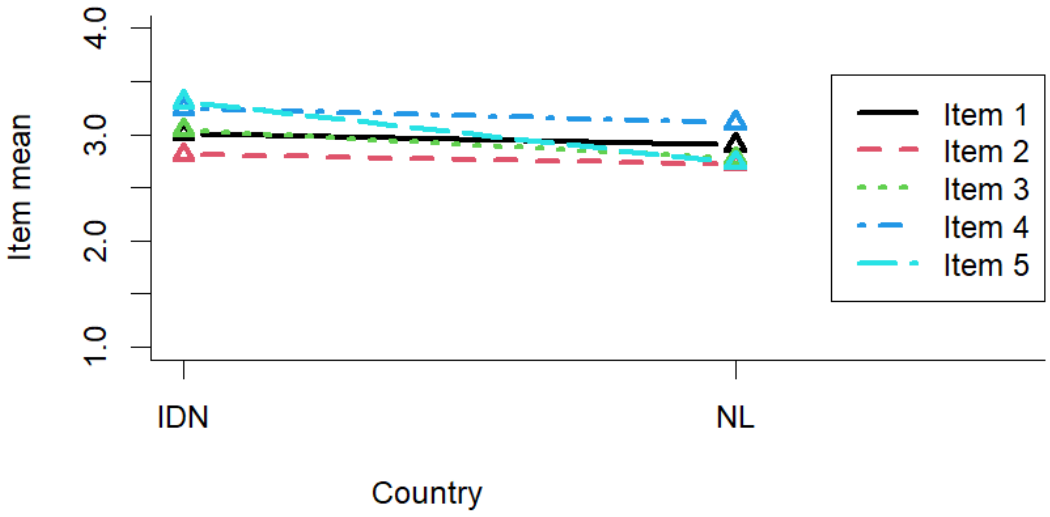
Note. The table above provides the item mean ($\bar{X}_i$) score and scalability (H_i) and their standard deviations (SD) for Indonesia (IDN) and the Netherlands (NL). For confidence interval results, refer to Appendix D.

Based on Table 3, Indonesia consistently exhibits a higher mean score across most items, except item 1, where the item means are relatively similar (IDN: 3.01, NL: 2.90), and the total scale mean is also higher in Indonesia (15.41) compared to the Netherlands (14.23). The significant differences between the two countries were confirmed by the lack of overlap in the confidence interval for each item and the total scale score (IDN: 15.336 to 15.484; NL: 14.138 to 14.322), highlighting significant differences in mean scores. These results indicated that the scale score and item scores were not similarly distributed across the two countries.

The overall difficulty order based on the item means of both countries identified item 4 as the most popular and item 2 as the least popular. In Figure 1, item 5 *(Students don't start working for a long time after)*, which was the most popular in Indonesia and less popular in the Netherlands, appears to violate the overall difficulty order. The *Crit* value of 165 also showed that the violation was statistically significant.

Figure 1

Item Means Order between Indonesia and the Netherlands



Note. IDN = Indonesia, NL = Netherlands. An intersection of the lines (items) indicates a violation.

Scale 2: Inquiry-based Teaching and Learning Practices

After the missing items were removed, the remaining students were 6127 for Indonesia and 3944 for the Netherlands. The scalability across both countries exceeded the acceptable threshold (0.3), with Indonesia exhibiting weak scalability of 0.311 (95% CI: 0.299 to 0.323), and the Netherlands demonstrating moderate scalability of 0.425 (95% CI: 0.407 to 0.443). Scalability was stronger in the Netherlands, with all items showing scores generally above 0.4, while the scalability in Indonesia is generally acceptable. The highest scalability in Indonesia was seen in item 4 at 0.349, while in the Netherlands, items 4 and 9 achieved the highest scalability, each scoring 0.479. Notably, item 8 showed the lowest scalability in Indonesia at 0.285 and relatively low in the Netherlands at 0.385. The non-overlapping confidence interval for each individual item and the total scale score between the two countries confirmed a statistically significant difference, suggesting that measurement invariance is not maintained, where the model fit better in the Netherlands than in Indonesia.

Table 4

Items Mean and Scalability across Indonesia and the Netherlands

Items	$\bar{X}_i$ *(SD)*		H_i *(SE)*	
	IDN	NL	IDN	NL
1	2.10 (0.93)	2.46 (0.87)	0.294 (0.008)	0.371 (0.011)
2	3.21 (0.72)	2.80 (0.83)	0.252 (0.010)	0.344 (0.011)
3	2.36 (0.94)	3.21 (0.83)	0.318 (0.008)	0.478 (0.011)
4	2.55 (0.90)	2.67 (0.82)	0.349 (0.008)	0.479 (0.011)
5	2.18 (0.95)	2.58 (0.85)	0.300 (0.008)	0.410 (0.011)
6	2.99 (0.93)	3.50 (0.77)	0.334 (0.008)	0.438 (0.013)
7	3.31 (0.86)	3.34 (0.80)	0.306 (0.009)	0.467 (0.011)
8	1.93 (0.91)	2.72 (0.88)	0.285 (0.008)	0.385 (0.010)
9	2.83 (0.95)	3.29 (0.83)	0.343 (0.008)	0.479 (0.011)
Total Scale	23.45 (4.79)	26.57 (5.16)	0.311 (0.006)	0.425 (0.009)

Note. The table above provides the item mean ($\bar{X}_i$) score and scalability (H_i) and their standard deviations (SD) for Indonesia (IDN) and the Netherlands (NL). For confidence interval results, refer to Appendix D.

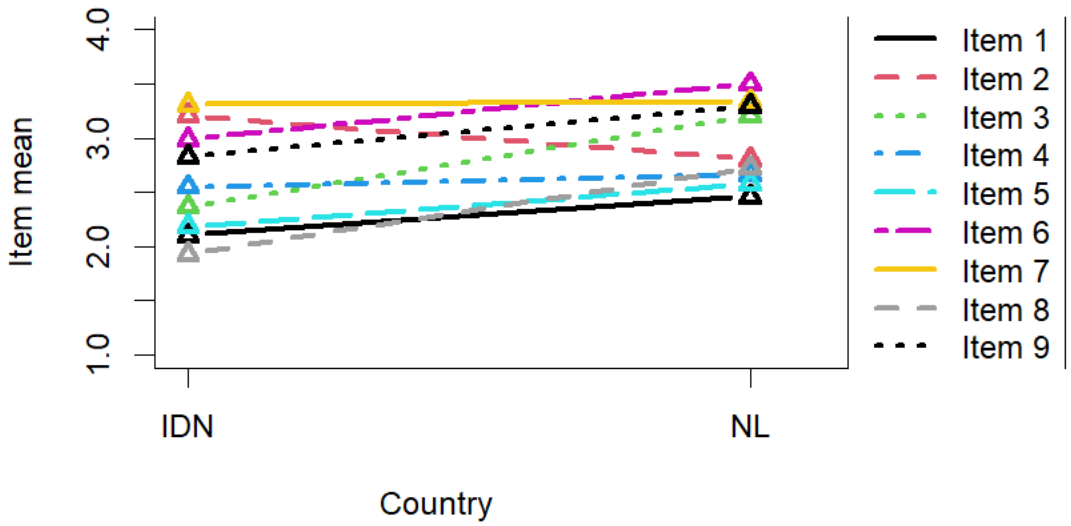
The results in Table 4 shows that the Netherlands generally exhibits higher across most items, except for item 2 (IDN: 3.21, NL: 2.80), compared to Indonesia. There was a 3.11 total scale score gap where the Netherlands (26.57) was higher than Indonesia (23.45). The total scale score and most all items showed no overlap in confidence interval, except item 7 (IDN: 3.288 to 3.332, NL: 3.315 to 3.365) where the confidence interval was slightly overlap, suggesting that the differences in mean score were statistically significant. These results confirmed that the distributions of scale score and item scores across the two countries are not similar.

The overall difficulty order of both countries based on the item means order identified item 7 as the most popular and item 1 as the least popular. Item 2 *(Students spend time in the laboratory doing practical experiments)*, which was more popular in Indonesia and less

popular in the Netherlands (Figure 2), violated the overall difficulty order. The violation was statistically significant, as evident by *Crit* value of 425.

Figure 2

Item Means Order between Indonesia and the Netherlands



Note. IDN = Indonesia, NL = Netherlands. An intersection of the lines (items) indicates a violation.

Scale 3: Teacher Support in Classes

After data removal for the missing items, the analysis 6242 students were from Indonesia, and 4280 are from the Netherlands. Scalability across both countries exceeded the acceptable threshold of 0.3, where Indonesia showed weak scalability of 0.341 (95% CI: 0.325 to 0.357) and the Netherlands with a good scalability at 0.559 (95% CI: 0.543 to 0.575). In Indonesia, the highest scalability was found in item 2 at 0.366, whereas in the Netherlands, item 4 demonstrates the highest scalability at 0.589. Item 1 had the lowest scalability in both countries, with 0.286 for Indonesia and 0.530 for the Netherlands. The absence of overlap in

confidence intervals of each individual item and the total scale score indicated a statistically significant differences in scalability. These findings suggested that measurement invariance is not achieved, where the model well fit in the Netherlands than Indonesia.

Table 5

Item Mean and Scalability across Indonesia and the Netherlands

Items	$\bar{X}_i$ *(SD)*		H_i *(SE)*	
	IDN	NL	IDN	NL
1	2.41 (0.93)	2.24 (0.88)	0.286 (0.010)	0.530 (0.010)
2	2.02 (0.93)	2.09 (0.87)	0.366 (0.009)	0.563 (0.009)
3	1.73 (0.89)	2.61 (0.94)	0.333 (0.009)	0.564 (0.009)
4	1.56 (0.79)	2.24 (0.91)	0.359 (0.010)	0.589 (0.009)
5	1.75 (0.85)	2.35 (0.92)	0.362 (0.009)	0.552 (0.009)
Total Scale	9.46 (2.94)	11.52 (3.65)	0.341 (0.008)	0.559 (0.008)

Note. The table above provides the item mean ($\bar{X}_i$) score and scalability (H_i) and their standard deviations (SD) for Indonesia (IDN) and the Netherlands (NL). For confidence interval results, refer to Appendix D.

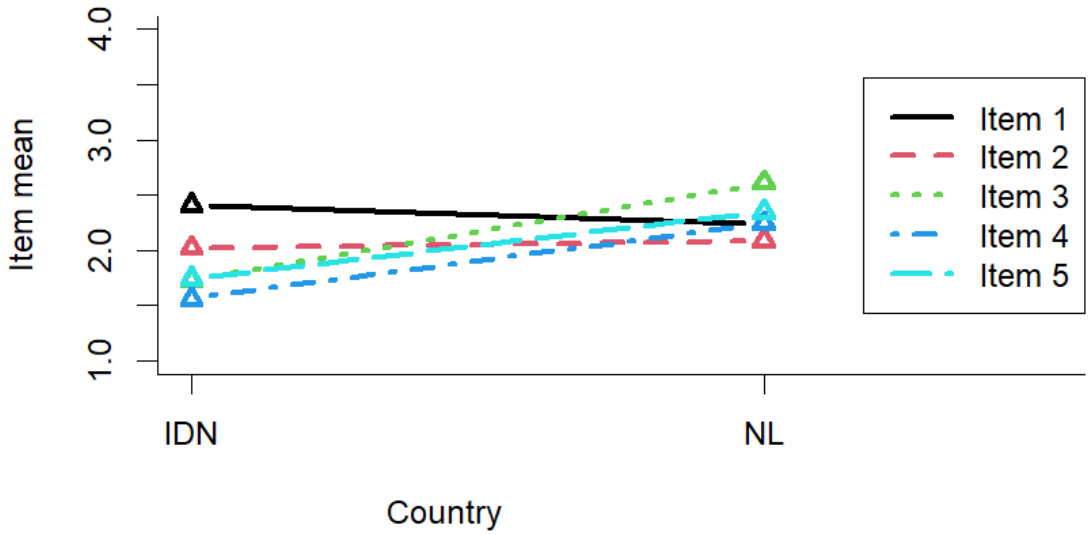
Table 5 shows that the distribution of item mean scores is somewhat higher in the Netherlands, where all items have a mean above 2.0, than in Indonesia, where most items have a mean below than 2.0. The scale scores were also higher in the Netherlands, with 11.52 compared to Indonesia with 9.46. The confidence interval results with no overlap in item scores and scale score indicated that the differences of mean scores were statistically significant, suggesting that the distribution of scale score and item scores across the two countries were not similar.

Item means established the overall difficulty order which indicated item 4 as the most popular and item 1 as the least popular. In Figure 3, item 3 *(The teacher helps students with their learning)* appeared to be the most popular in the Netherlands and less popular in

Indonesia, demonstrating the violation of the overall difficulty order with a significant *Crit* value of 410.

Figure 3

Item Means Order between Indonesia and the Netherlands



Note. IDN = Indonesia, NL = Netherlands. An intersection of the lines (items) indicates a violation.

Scale 4: Teacher-directed Instruction

After the removal of missing items, the analysis included 6246 students from Indonesia and 4087 students from the Netherlands. Both countries exhibited the scalability above the acceptable threshold (0.3), where Indonesia indicates a weak scalability of 0.372 (95% CI: 0.356 to 0.388) and the Netherlands with a moderate scalability of 0.418 (95% CI: 0.396 to 0.440). In Indonesia, the highest scalability was observed in item 3 at 0.400 (95% CI: 0.382 to 0.418), compared to the Netherlands where item 4 was the highest at 0.465 (95% CI: 0.443

to 0.487). Notably, item 2 and item 3 showed overlapping scalability intervals between the two countries, with Indonesia at 0.343 (95% CI: 0.344 to 0.379) for item 2 and 0.400 (95% CI: 0.382 to 0.418) for item 3, and the Netherlands at 0.354 (95% CI: 0.327 to 0.382) for item 2 and 0.403 (95% CI: 0.379 to 0.426) for item 3. The absence of overlap in the confidence intervals for the total scale scores, alongside the partial overlaps in individual item scores, highlighted a statistically significant difference in scalability, indicating that the model does not maintain measurement invariance, where it fit better in the Netherlands than Indonesia.

Table 6

Item Mean and Scalability across Indonesia and the Netherlands

Items	$\bar{X}_i$ *(SD)*		H_i *(SE)*	
	IDN	NL	IDN	NL
1	2.44 (0.89)	2.27 (0.90)	0.338 (0.010)	0.446 (0.011)
2	2.46 (0.86)	1.89 (0.83)	0.361 (0.009)	0.354(0.014)
3	2.45 (0.92)	2.77 (0.87)	0.400 (0.009)	0.403 (0.012)
4	2.33 (0.90)	2.23 (0.85)	0.389 (0.009)	0.465 (0.011)
Total Scale	9.69 (2.57)	9.16 (2.50)	0.372 (0.008)	0.418 (0.011)

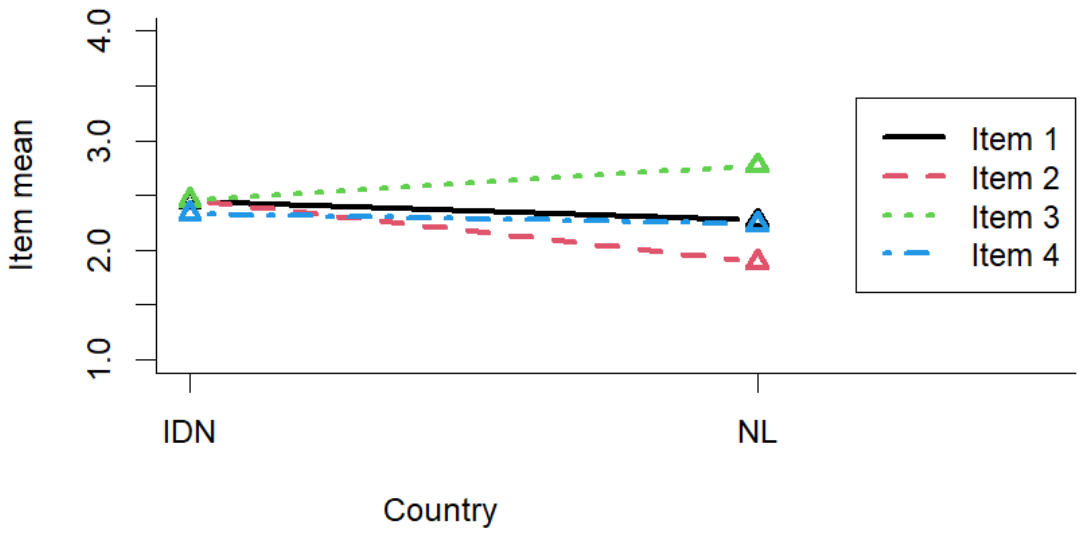
Note. The table above provides the item mean ($\bar{X}_i$) score and scalability (H_i) and their standard deviations (SD) for Indonesia (IDN) and the Netherlands (NL). For confidence interval results, refer to Appendix D.

Table 6 shows the distribution of item mean scores is slightly higher in Indonesia for most items, except for item 3, where the Netherlands scores higher (IDN: 2.45, NL: 2.77). Furthermore, the total scale score in Indonesia was somewhat higher than the Netherlands with only 0.53 points gap. The confidence interval showing no overlap for both item mean scores and scale score, confirmed that the differences are statistically significant. This indicates that the distribution of item mean scores and scale score across the two countries were not similar.

The overall difficulty order based on the item means was item 3 as the most popular and item 2 as the least. In Figure 4, item 2 *(A whole class discussion takes place with the teacher)* is most popular in Indonesia and least popular in the Netherlands, indicating violation to the overall difficulty order. The *Crit* value of 172 confirmed the violation was statistically significant.

Figure 4

Item Means Order between Indonesia and the Netherlands



Note. IDN = Indonesia, NL = Netherlands. An intersection of the lines (items) indicates a violation.

Scale 5: Perceived Feedback

The remaining data after removing the missing items was 6256 students for Indonesia and 4100 students for the Netherlands. Scalability across countries surpassed the acceptable threshold (0.3), where Indonesia exhibited moderate scalability at 0.492 (95% CI: 0.476 to 0.508) and the Netherlands with strong scalability of 0.695 (95% CI: 0.679 to 0.711).

Interestingly, scalability in the Netherlands was consistently strong with all items showing scalability scores above 0.6. In contrast, Indonesia showed a range from moderate to good scalability, with item scores ranging from 0.429 to 0.538. The item with the lowest scalability for both countries was item 1, with 0.429 in Indonesia and 0.613 in the Netherlands. The highest scalability in Indonesia was observed in item 3 at 0.538, and in the Netherlands, item 4 displays the highest scalability at 0.726. There was no overlap in confidence intervals for each individual item scores and the total scale scores between the two countries confirmed a statistically significant difference. These findings suggest that measurement invariance was not maintained, where it fit better in the Netherlands than in Indonesia.

Table 7

Item Mean and Scalability across Indonesia and the Netherlands

Items	$\bar{X}_i$ *(SD)*		H_i *(SE)*	
	IDN	NL	IDN	NL
1	1.96 (0.85)	2.03 (0.77)	0.429 (0.010)	0.613 (0.011)
2	1.94 (0.87)	1.88 (0.81)	0.473 (0.010)	0.741 (0.009)
3	2.22 (0.91)	2.00 (0.83)	0.538 (0.008)	0.712 (0.008)
4	2.57 (0.93)	2.03 (0.82)	0.525 (0.009)	0.726 (0.008)
5	2.65 (0.94)	2.03 (0.83)	0.494 (0.009)	0.686 (0.009)
Total Scale	11.33 (3.34)	9.96 (3.47)	0.492 (0.008)	0.695 (0.008)

Note. The table above provides the item mean ($\bar{X}_i$) score and scalability (H_i) and their standard deviations (SD) for Indonesia (IDN) and the Netherlands (NL). For confidence interval results, refer to Appendix D.

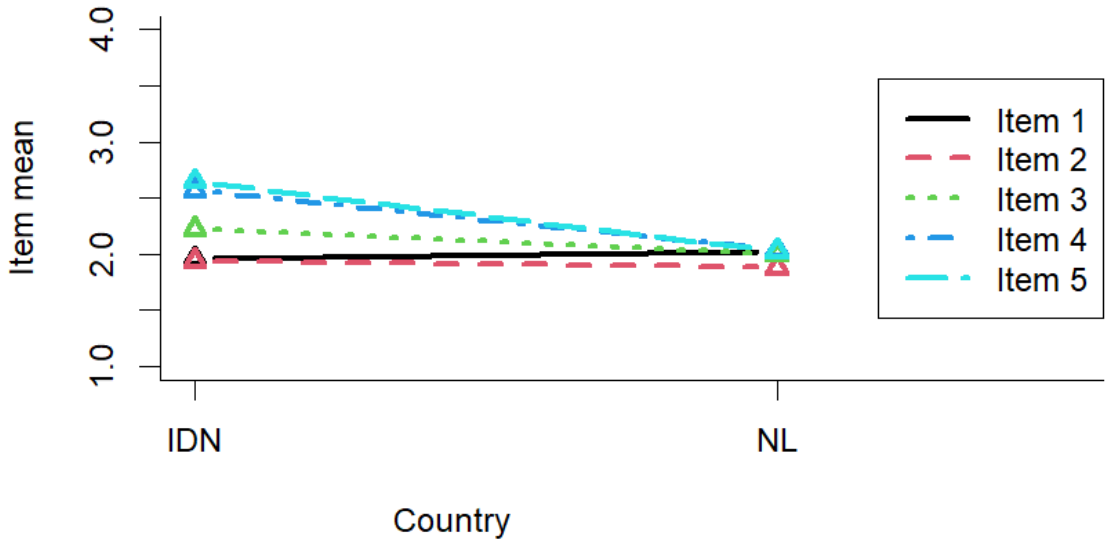
In Table 7, the distribution of item mean score is somewhat higher in Indonesia, where most items have a mean score above 2.0, compared to the Netherlands, where most items hover around 2.0. Similarly, the total scale score also showed that Indonesia has a slightly higher at 11.33 as opposed to the Netherlands at 9.96. These differences were

statistically significant, as evident by confidence interval calculation, which showed no overlap for both item scores and scale score. Consequently, the distribution of scale score and item scores across the two countries were not similar.

The overall difficulty order based on the item means was item 5 as the most popular and item 1 as the least popular. In Figure 5, item 5 *(The teacher advises me on how to reach my learning goals)*, is more popular in Indonesia and less popular in the Netherlands, initially indicated a violation. However, the *Crit* value of 0 for all items confirmed no evidence of violation of the overall difficulty order.

Figure 5

Item Means Order between Indonesia and the Netherlands



Note. IDN = Indonesia, NL = Netherlands. An intersection of the lines (items) indicates a violation.

Discussion

This study revealed that using PISA 2015 scores on students' perceived teaching practices for direct comparison between Indonesia and the Netherlands could be problematic due to some conditions. While the internal structure of the scales is generally comparable, there are differences in scalability and item ordering across these countries. The scales demonstrate variability in measuring constructs, which could skew comparisons. A further complication is the inconsistency in item difficulty order, which suggests that the same items may not be interpreted similarly in each cultural context. These results indicate that the raw scores from PISA 2015 may not provide equitable or meaningful comparisons. Adjustments like scale modification might be necessary to ensure that comparisons are valid and account for these substantial contextual differences.

Aligning with the teaching practices framework by Hamre et al. (2013), it becomes clear why specific SPTP scales of PISA, such as Inquiry-based Teaching and Learning Practices and Teacher Support in Classes, as well as Teacher-directed Instruction and Perceived Feedback, exhibited overlap at lower thresholds. These overlaps suggest that these elements are often experienced as part of broader, integrated instructional practices, where students perceive inquiry-based activities as part of the supportive and encouraging environment the teacher provides that blends engagement and emotional support. Additionally, feedback is a critical component of instructional practices, and students may experience teacher-directed activities that are inseparable from the feedback they obtain, combining aspects of clarity, guidance, and scaffolding. Despite the possibility of combining some of the scales, the decision to maintain the PISA scale was supported by several important reasons. The individual PISA scales offered more measurement precision since they could provide more detailed insights into specific teaching practices, whereas merging

them would reduce the ability to track the subtle variations between countries. Furthermore, if the individual PISA scales were retained, comparing how teaching practices are perceived in each country would be easier. Considering the cultural and educational differences between Indonesia and the Netherlands, maintaining them allows for a clearer understanding of how these perceptions differ, supporting the cross-cultural comparisons.

The measurement invariance results from this study underscore the profound influence of cultural context on the perception of teaching practices, providing support for Vygotsky's Sociocultural theory (Tzuriel, 2021) and Bronfenbrenner's Ecological System theory (1977), which emphasize the role of culture in shaping learning experiences, as discussed in the introduction. The variability in scalability between Indonesia and the Netherlands can be attributed to different societal norms, educational traditions, and expectations that reflect how educational measurements developed within one cultural setting may not be entirely applicable in another. The stronger scalability in the Netherlands suggests that the scales are more resonant with the characteristics of the Dutch educational system. In contrast, the moderate to weak scalability in Indonesia indicates potential misalignment with its educational and cultural contexts.

The observed differences in item and scale scores distribution across Indonesia and the Netherlands further clearly demonstrate how cultural contexts significantly impact student perceptions of teaching practices. These are not mere statistical variations but reflect fundamental cultural influences that uniquely shape educational experiences and evaluations. In Indonesia, higher mean scores across many scales indicate a cultural orientation that values hierarchical and collective structures. This cultural orientation likely leads to greater respect for authority figures, like teachers, influencing students to positively view and rate aspects of emotional and classroom management. This aligns with the societal norms where teachers are

respected as key community figures. Conversely, in the Netherlands, the emphasis on individualism and critical thinking is reflected in higher scores for scales related to open dialogue and student autonomy. Dutch students are nurtured to question and analyze from an early age, fostering an educational environment that prizes independence and critical engagement, which students will likely rate highly. These findings align with Hofstede's (1986) cultural dimensions theory, particularly the contrast between high and low power distance.

The notable deviations in overall item difficulty order and order across Indonesia and the Netherlands, align closely with the concerns about psychometric challenges. These challenges include the potential for cultural bias in test design, issues related to language and translation, and the fundamental difficulty of ensuring construct validity across diverse cultural settings. The observed discrepancies in how certain items are ranked in difficulty between Indonesian and Dutch students can be attributed to these varying cultural norms and educational expectations, which may influence students' understanding and responses to assessment items. For instance, what is deemed a straightforward item in one cultural setting might be interpreted very differently in another due to variations in teaching styles, student-teacher dynamics, or the educational focus emphasized within different schooling systems.

Strength and Limitation

This study employs Mokken Scale Analysis (MSA) to provide nuanced insights into how cultural factors influence student perceptions. MSA can capture the subtler variations in data that parametric methods might overlook due to its flexibility in handling ordinal data without the assumption of a normal distribution. Furthermore, the study advances the understanding of scalability and the structure of educational measurement scales across different populations.

However, this study's findings are also shaped by several limitations. While flexible, the nonparametric nature of MSA limits the assessment of complex forms of measurement invariance, such as factorial invariance, which might be better evaluated through multi-group Confirmatory Factor Analysis within a Structural Equation Modelling framework. Additionally, the exclusion of the Adaption of Instruction scale may overlook important pedagogical aspects, reducing the scope of analyzed teaching practices. The study also faces potential biases from the stratified sampling method and the handling of missing data, which could influence the representativeness of the results. Likewise, the development of PISA instruments often reflects educational norms, practices, and values that are more prevalent in Western contexts and can lead to scales that do not fully capture or resonate with students' educational experiences or value systems in non-Western cultures, such as Indonesia.

Conclusion

This study revealed significant differences in the psychometric properties of the SPTP scales of PISA 2015, which appear to be influenced by cultural and contextual factors. Among the scales, Perceived Feedback and Disciplinary Climate in Classes demonstrated sufficient consistency and scalability, making them more suitable for cross-cultural comparisons. In contrast, Teacher Support and Inquiry-based Teaching scales showed greater cultural sensitivity, limiting them for valid comparison without adaptation. Teacher-directed Instruction exhibited moderate scalability but still showed variability in item performance across both countries. The findings also underscore the importance of integrating culturally sensitive approaches into educational assessments to ensure fairness in cross-cultural comparison. In contrast, culturally sensitive scales require careful adaptation through revision or contextualization to align with diverse educational norms. Further research should focus on refining cross-cultural data to better account for these disparities. Investigating hybrid

scales that merge universally teaching concepts with culturally specific items and longitudinal studies on the evolution of teaching perceptions could offer deeper insights into how cultural contexts influence teaching practice assessments.

References

Abdessallam, K., El Ghouati, A., & Nakkam, J. (2020). The Impact of Culture on Teaching: A study of the impact of teachers' cultural beliefs and practices on students' motivation. In Innovative Space of Scientific Research Journals, *International Journal of Innovation and Scientific Research* (Vols. 279–287).

Aditomo, A., & Köhler, C. (2020). Do student ratings provide reliable and valid information about teaching quality at the school level? Evaluating measures of science teaching in PISA 2015. *Educational Assessment Evaluation and Accountability*, *32*(3), 275–310. https://doi.org/10.1007/s11092-020-09328-6

André, S., Maulana, R., Helms-Lorenz, M., Telli, S., Chun, S., Fernández-García, C., De Jager, T., Irnidayanti, Y., Inda-Caro, M., Lee, O., Safrina, R., Coetzee, T., & Jeon, M. (2020). Student perceptions in measuring teaching behavior across six countries: A Multi-Group Confirmatory Factor Analysis Approach to Measurement invariance. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00273

Bidwell, C. E., & Kasarda, J. D. (1980). Conceptualizing and measuring the effects of school and schooling. *American Journal of Education*, *88*(4), 401–430. https://doi.org/10.1086/443540

Brodin, U., Fors, U., & Laksov, K. B. (2010). The application of Item Response Theory on a teaching strategy profile questionnaire. *BMC Medical Education*, *10*(1). https://doi.org/10.1186/1472-6920-10-14

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531. https://doi.org/10.1037/0003-066x.32.7.513

Bryer, F., & Beamish, W. (2019). Western perspectives on teaching, learning, and behaviour. In *Advancing inclusive and special education in the Asia-Pacific* (pp. 3–23). https://doi.org/10.1007/978-981-13-7177-6_1

Carrozzino, D., Christensen, K. S., Patierno, C., Woźniewicz, A., Møller, S. B., Arendt, I. T., Zhang, Y., Yuan, Y., Sasaki, N., Nishi, D., Montiel, C. B., Ceccatelli, S., Mansueto, G., & Cosci, F. (2022). Cross-cultural validity of the WHO-5 Well-Being Index and Euthymia Scale: A clinimetric analysis. *Journal of Affective Disorders*, *311*, 276–283. https://doi.org/10.1016/j.jad.2022.05.111

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. *In M. R. Gunnar & L. A. Sroufe (Eds.), Self process and development: The Minnesota symposia on child development (Vol. 23, pp. 43–77).* Hillsdale, NJ: Erlbaum.

De Ree, J. (2015). Teacher certification and beyond: An empirical evaluation of the teacher certification program and education quality improvements in Indonesia. In *The World Bank*. The World Bank. https://documents1.worldbank.org/curated/en/129551468196175672/pdf/104599-WP-P102259-PUBLIC-Teacher-Certification-and-beyond-final.pdf

Education Policy outlook. (2020). *Education Policy Outlook*. https://doi.org/10.1787/4cf5b585-en

Effendi-Hasibuan, M. H., Ngatijo, N., & Sulistiyo, U. (2019a). Inquiry-based learning in Indonesia: portraying supports, situational beliefs, and chemistry teachers adoptions. *Journal of Turkish Science Education*, *16*(4), 538–553. https://doi.org/10.36681/tused.2020.6

Fang, Y., & Gopinathan, S. (2009). Teachers and Teaching in Eastern and Western Schools: A Critical Review of Cross-Cultural Comparative Studies. *Springer eBooks*, 557–572. https://doi.org/10.1007/978-0-387-73317-3_36

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through Interactions. *The Elementary School Journal*, *113*(4), 461–487. https://doi.org/10.1086/669616

Hargreaves, A. (2001). *Changing teachers, changing times: Teachers' Work and Culture in the Postmodern Age*. A&C Black.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the Polytomous Mokken IRT model. Applied Psychological Measurement, 19(4), 337–352. https://doi.org/10.1177/014662169501900404

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations*, *10*(3), 301–320. https://doi.org/10.1016/0147-1767(86)90015-5

Koopman, L., Zijlstra, B. J. H., & Van Der Ark, L. A. (2022). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. Quality of Life Research, 31(1), 25–36. https://doi.org/10.1007/s11136-021-02840-2

Law No. 14 of 2005. (n.d.). Database Peraturan | JDIH BPK. https://peraturan.bpk.go.id/Details/40266/uu-no-14-tahun-2005

Maulana, R., André, S., Helms-Lorenz, M., Ko, J., Chun, S., Shahzad, A., Irnidayanti, Y., Lee, O., De Jager, T., Coetzee, T., & Fadhilah, N. (2020). Observed teaching behaviour in secondary education across six countries: measurement invariance and indication of cross-national variations. *School Effectiveness*

and School Improvement, *32*(1), 64–95. https://doi.org/10.1080/09243453.2020.1777170

Ministerie van Algemene Zaken. (2021). *Teaching qualification*. Working in Education | Government.nl. https://www.government.nl/topics/working-in-education/teaching-qualification

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton.

Molenaar, I. W., & Sijtsma, K. (2000). *MPS5 for Windows. A program for Mokken scale analysis for polytomous items*. iec ProGAMMA.

Novita, P. (2022). The quest for teacher education quality in Indonesia: The long and winding road. In M. S. Khine, & Y. Liu (Eds.), *Handbook of Research on Teacher Education* (pp. 651-673). Springer, Singapore. https://doi.org/10.1007/978-981-16-9785-2_32

OECD. (2017). *PISA 2015 Technical Report*. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf

Oliver, R. M., Wehby, J. H., & Reschly, D. J. (2011). Teacher classroom management practices: effects on disruptive or aggressive student behavior. *Campbell Systematic Reviews*, *7*(1), 1–55. https://doi.org/10.4073/csr.2011.4

Palmgren, P. J., Brodin, U., Nilsson, G. H., Watson, R., & Stenfors, T. (2018). Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis – a pragmatic approach. *BMC Medical Education*, *18*(1). https://doi.org/10.1186/s12909-018-1334-8

Pianta, R. C., Hamre, B. K., & Allen, J. P. (2012). Teacher-Student Relationships and Engagement: Conceptualizing, measuring, and improving the capacity of

classroom interactions. In *Springer eBooks* (pp. 365–386). https://doi.org/10.1007/978-1-4614-2018-7_17

PISA 2015 Database. (n.d.-b). OECD. https://www.oecd.org/en/data/datasets/pisa-2015-database.html

Rosenholtz, S. J. (1989). *Teachers' workplace : the social organization of schools*. Longman.

Schophuizen, M., & Kalz, M. (2020). Educational innovation projects in Dutch higher education: bottom-up contextual coping to deal with organizational challenges. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00197-z

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.

Sijtsma, K., & Van Der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158. https://doi.org/10.1111/bmsp.12078

Sijtsma, K., & Van Der Ark, L. A. (2020). Measurement models for psychological attributes. In *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/9780429112447

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*, 72-99.

Tzuriel, D. (2021). The Socio-Cultural Theory of Vygotsky. In *Social interaction in learning and development* (pp. 53–66). https://doi.org/10.1007/978-3-030-75692-5_3

Van De Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01064

Van Der Ark, L. A. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, *20*(11). https://doi.org/10.18637/jss.v020.i11

Van Der Ark, L. A. (2012). New developments in Mokken Scale Analysis INR. *Journal of Statistical Software*, *48*(5). https://doi.org/10.18637/jss.v048.i05

Vélez-Agosto, N. M., Soto-Crespo, J. G., Vizcarrondo-Oppenheimer, M., Vega-Molina, S., & Coll, C. G. (2017). Bronfenbrenner's Bioecological Theory Revision: Moving culture from the macro into the micro. *Perspectives on Psychological Science*, *12*(5), 900–910. https://doi.org/10.1177/1745691617704397

Appendix A
The Codebook of the Scales
(Source: The PISA 2015 Report)

Scale 1: Discipline Climate in Classes

Item	Codebook	Variable *How often does this happen in your (school science) lessons?*
1	ST097Q01TA	Students don’t listen to what the teachers say.
2	ST097Q02TA	There is no noise or disorder.
3	ST097Q03TA	The teacher waits long for students to quiet down.
4	ST097Q04TA	Students cannot work well.
5	ST097Q05TA	Students don't start working for a long time after.

Scale 2: Inquiry-based Teaching and Learning Practices

Item	Codebook	Variable *When learning (school science)?*
1	ST098Q01TA	Students are given opportunities to explain their ideas.
2	ST098Q02TA	Students spend time in the laboratory doing practical experiments.
3	ST098Q03NA	Students are required to argue about science questions.
4	ST098Q05TA	Students are asked to draw conclusions from an experiment they have conducted.
5	ST098Q06TA	The teacher explains (school science) idea can be applied.
6	ST098Q07TA	Students are allowed to design their own experiments.
7	ST098Q08NA	There is a class debate about investigations.
8	ST098Q09TA	The teacher clearly explains relevance (broad science) concepts to our lives.
9	ST098Q10NA	Students are asked to do an investigation to test ideas.

Scale 3: Teacher Support in Classes

Item	Codebook	Variable *How often does this happen in your (school science) lessons?*
1	ST100Q01TA	The teacher shows in every student learning.
2	ST100Q02TA	The teacher gives extra help.
3	ST100Q03TA	The teacher helps students with their learning.
4	ST100Q04TA	The teacher continues teaching until students understand.
5	ST100Q05TA	Teacher gives an opportunity to express opinions.

Scale 4: Teacher-directed Instruction

Item	Codebook	Variable *How often does this happen in (school science)?*
1	ST103Q01NA	The teacher explains scientific ideas.
2	ST103Q03NA	A whole class discussion takes place with the teacher.
3	ST103Q08NA	The teacher discusses our questions.
4	ST103Q11NA	The teacher tells me how I am performing in this course.

Scale 5: Perceived Feedback

Item	Codebook	Variable *How often does this happen in (school science)?*
1	ST104Q01NA	The teacher tells me how I am performing in this course.
2	ST104Q02NA	The teacher gives me feedback on my strengths (school science) subject.
3	ST104Q03NA	The teacher tells me in which area I can still improve.
4	ST104Q04NA	The teacher tells me how I can improve my performance.
5	ST104Q05NA	The teacher advises me on how to reach my learning goals.

Appendix B
The 10-step of Mokken Scale Analysis by Sijtsma and van der Ark (2017), and Steps from Molenaar & Sijtsma, (2000), MSP manual Section 4.8

Data Examination

Step 1. *Recoding.* If necessary, recode the item scores to ensure consistency and meaningful interpretation.

Step 2. *Inadmissible Scores and Missing Data*. Identify and handle inadmissible scores (e.g., out-of-range values). Address missing data appropriately (e.g., impute missing values).

Step 3. *Outliers.* Detect and manage outliers that may affect the analysis.

Scale Identification

Step 4. *Scalability*. Compute scalability coefficients H_i to assess the scalability of the items. Compute total scalability coefficient (Loeveinger's H) to evaluate the strength of the scale.

Step 5. *Local Independence*. Check the assumption of local independence (items are conditionally independent given the latent trait).

Step 6. *Monotonicity*. Investigate whether items exhibit monotonicity (higher scores correspond to higher levels of trait).

Step 7. *Invariant Item Ordering*. Examine the assumption of invariant item ordering (items maintain their order across different levels of trait).

Scale Properties

Step 8. *Reliability.* Assess the reliability of the scale.

Step 9. *Norms*. Establish norms or reference values for interpreting scale scores.

Step 10. *Group Comparison*. Compare scale scores across different groups to identify potential differences.

Steps from Molenaar & Sijtsma, 2000, MSP manual Section 4.8

(The following steps are included in Step 10 (Group Comparison)

a. Does the measurement model fit equally well across subgroups?
b. Are scale score and item scores similarly distributed?

    1. Are scale scores similarly distributed?
    2. Are item means similar?

c. Are there specific items for which the overall difficulty order is violated in one or more subgroups?

Appendix C

The Automated Item Selection Procedure (AISP) results

Table 1

Automated Item Selection Procedure result for Indonesia dataset

Codebook	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
DISCLI.1	1	1	2	2	2	2	2	2	2	2	2
DISCLI.2	1	1	2	2	2	2	2	2	2	2	2
DISCLI.3	3	3	2	2	2	2	2	2	2	2	0
DISCLI.4	1	1	2	2	2	2	2	2	2	2	0
DISCLI.5	3	3	2	2	2	2	2	2	0	0	0
TDTEACH.1	1	1	1	1	1	1	3	0	0	0	0
TDTEACH.2	1	1	1	1	1	1	3	3	3	3	0
TDTEACH.3	1	1	1	1	1	1	3	3	3	3	0
TDTEACH.4	1	1	1	1	1	1	3	3	3	0	0
TEACHSUP.1	2	2	3	3	3	3	5	0	0	0	0
TEACHSUP.2	2	2	3	3	3	4	5	5	5	0	0
TEACHSUP.3	2	2	3	3	3	4	5	5	5	0	0
TEACHSUP.4	2	2	3	3	3	4	5	5	0	0	0
TEACHSUP.5	2	2	3	3	3	3	4	4	4	0	0
PERFEED.1	1	1	1	1	1	1	1	1	1	0	3
PERFEED.2	1	1	1	1	1	1	1	1	1	1	3
PERFEED.3	1	1	1	1	1	1	1	1	1	1	1
PERFEED.4	1	1	1	1	1	1	1	1	1	1	1
PERFEED.5	1	1	1	1	1	1	1	1	1	1	1
IBTEACH.1	2	2	3	3	3	3	4	4	4	4	0
IBTEACH.2	2	2	3	3	3	3	0	0	7	0	0
IBTEACH.3	2	2	3	3	3	3	4	4	4	4	0
IBTEACH.4	2	2	3	3	3	3	4	4	7	0	0
IBTEACH.5	2	2	3	3	3	3	6	7	8	0	0
IBTEACH.6	2	2	3	3	3	3	4	6	6	0	0
IBTEACH.7	2	2	3	3	3	3	4	6	6	0	0
IBTEACH.8	2	2	3	3	3	3	6	7	8	0	0
IBTEACH.9	2	2	3	3	3	3	4	6	6	0	0

Note. Codebook = SCALE NAME.(item number). For example, DISCLI.1 = Disciplinary Climate in Classes item 1, etc.), TDTEACH = Teacher-directed Instruction, TEACHSUP = Teacher Support in Classes, PERFEED = Perceived Feedback, IBTEACH = Inquiry-based Teaching and Learning Practices. Each shade indicates a set of items in the same original scale of PISA.

Table 2

Automated Item Selection Procedure result for Netherlands dataset

Codebook	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
DISCLI.1	1	1	1	2	2	2	2	2	2	2	2
DISCLI.2	1	1	1	2	2	2	2	2	2	2	2
DISCLI.3	1	1	1	2	2	2	2	2	2	2	2
DISCLI.4	0	0	0	2	2	2	2	2	2	2	2
DISCLI.5	1	1	1	2	2	2	2	2	2	2	2
TDTEACH.1	1	1	1	1	1	1	1	5	5	6	0
TDTEACH.2	1	1	1	1	1	1	1	0	0	0	0
TDTEACH.3	1	1	1	1	1	1	1	5	5	6	7
TDTEACH.4	1	1	1	1	1	1	1	1	5	6	7
TEACHSUP.1	2	2	2	3	3	3	4	4	4	4	4
TEACHSUP.2	2	2	2	3	3	3	4	4	4	4	4
TEACHSUP.3	2	2	2	3	3	3	4	4	4	4	4
TEACHSUP.4	2	2	2	3	3	3	4	4	4	4	4
TEACHSUP.5	2	2	2	3	3	3	4	4	4	4	4
PERFEED.1	1	1	1	1	1	1	1	1	1	1	1
PERFEED.2	1	1	1	1	1	1	1	1	1	1	1
PERFEED.3	1	1	1	1	1	1	1	1	1	1	1
PERFEED.4	1	1	1	1	1	1	1	1	1	1	1
PERFEED.5	1	1	1	1	1	1	1	1	1	1	1
IBTEACH.1	2	2	2	3	3	3	3	3	0	0	0
IBTEACH.2	2	2	2	3	3	3	3	3	3	3	5
IBTEACH.3	2	2	2	3	3	3	3	3	3	3	3
IBTEACH.4	2	2	2	3	3	3	3	3	3	3	5
IBTEACH.5	2	2	2	3	3	3	3	3	3	5	6
IBTEACH.6	2	2	2	3	3	3	3	3	3	3	3
IBTEACH.7	2	2	2	3	3	3	3	3	3	3	3
IBTEACH.8	2	2	2	3	3	3	3	3	3	5	6
IBTEACH.9	2	2	2	3	3	3	3	3	3	3	3

Note. Codebook = SCALE NAME.(item number). For example, DISCLI.1 = Disciplinary Climate in Classes item 1, etc.), TDTEACH = Teacher-directed Instruction, TEACHSUP = Teacher Support in Classes, PERFEED = Perceived Feedback, IBTEACH = Inquiry-based Teaching and Learning Practices. Each shade indicates a set of items in the same original scale of PISA.

Appendix D
Confidence Interval Results per Scale

Scale 1: Discipline Climate in Classes

Item	95% CI for Item and Scale Score		95% CI for Scalability	
	IDN	NL	IDN	NL
1	(2.991, 3.029)	(2.878, 2.922)	(0.433, 0.473)	(0.604, 0.648)
2	(2.789, 2.831)	(2.698, 2.742)	(0.457, 0.497)	(0.667, 0.707)
3	(3.017, 3.063)	(2.747, 2.793)	(0.451, 0.487)	(0.637, 0.673)
4	(3.230, 3.270)	(3.088, 3.132)	(0.448, 0.484)	(0.675, 0.719)
5	(3.290, 3.330)	(2.706, 2.754)	(0.361, 0.405)	(0.527, 0.571)
Total Scale	(15.336, 15.484)	(14.138, 14.322)	(0.434, 0.466)	(0.632, 0.668)

Note. The sample size (n) for this scale is 6231 for Indonesia and 4296 for the Netherlands. There is no overlap in scale score and item scores, indicating that the differences are statistically significant.

Note. There is no overlap in item scalability and the total scale scalability, indicating that the differences are statistically significant.

Scale 2: Inquiry-based Teaching and Learning Practices

Item	95% CI for Item and Scale Score		95% CI for Scalability	
	IDN	NL	IDN	NL
1	(2.077, 2.123)	(2.433, 2.487)	(0.278, 0.310)	(0.349, 0.393)
2	(3.192, 3.228)	(2.774, 2.826)	(0.232, 0.272)	(0.322, 0.366)
3	(2.336, 2.384)	(3.184, 3.236)	(0.302, 0.334)	(0.456, 0.500)
4	(2.527, 2.573)	(2.644, 2.696)	(0.333, 0.365)	(0.457, 0.501)
5	(2.156, 2.204)	(2.553, 2.607)	(0.284, 0.316)	(0.388, 0.432)
6	(2.967, 3.013)	(3.476, 3.524)	(0.318, 0.350)	(0.413, 0.463)
7	(3.288, 3.332)	(3.315, 3.365)	(0.288, 0.324)	(0.445, 0.489)
8	(1.907, 1.953)	(2.693, 2.747)	(0.269, 0.301)	(0.365, 0.405)
9	(2.806, 2.854)	(3.264, 3.316)	(0.327, 0.359)	(0.457, 0.501)
Total Scale	(23.330, 23.570)	(26.409, 26.731)	(0.299, 0.323)	(0.407, 0.443)

Note. The sample size (n) for this scale is 6127 for Indonesia and 3944 for the Netherlands. The overlap only occurs in the confidence interval of item 7. All other items and the total score show no overlap, confirming the statistically significant differences.

Note. There is no overlap in item scalability and the total scale scalability, indicating that the differences are statistically significant.

Scale 3: Teacher Support in Classes

Item	95% CI for Item and Scale Score		95% CI for Scalability	
	IDN	NL	IDN	NL
1	(2.387, 2.433)	(2.214, 2.266)	(0.266, 0.306)	(0.510, 0.550)
2	(1.997, 2.043)	(2.064, 2.116)	(0.348, 0.384)	(0.545, 0.581)
3	(1.708, 1.752)	(2.582, 2.638)	(0.315, 0.351)	(0.546, 0.582)
4	(1.540, 1.580)	(2.213, 2.267)	(0.339, 0.379)	(0.571, 0.607)
5	(1.729, 1.771)	(2.322, 2.378)	(0.344, 0.380)	(0.543, 0.570)
Total Scale	(9.387, 9.533)	(11.411, 11.629)	(0.325, 0.357)	(0.543, 0.575)

Note. The sample size (n) for this scale is 6242 for Indonesia and 4280 for the Netherlands. There is no overlap in scale score and item scores, indicating that the differences are statistically significant.

Note. There is no overlap in item scalability and the total scale scalability, indicating that the differences are statistically significant.

Scale 4: Teacher-directed Instruction

Item	95% CI for Item and Scale Score		95% CI for Scalability	
	IDN	NL	IDN	NL
1	(2.418, 2.426)	(2.242, 2.298)	(0.318, 0.358)	(0.424, 0.468)
2	(2.439, 2.481)	(1.865, 1.915)	(0.343, 0.379)	(0.327, 0.381)
3	(2.427, 2.473)	(2.743, 2.797)	(0.382, 0.418)	(0.379, 0.427)
4	(2.308, 2.352)	(2.204, 2.256)	(0.371, 0.407)	(0.443, 0.487)
Total Scale	(9.626, 9.754)	(9.083, 9.237)	(0.356, 0.388)	(0.396, 0.440)

Note. The sample size (n) for this scale is 6246 for Indonesia and 4087 for the Netherlands. There is no overlap in scale score and item scores, indicating that the differences are statistically significant.

Note. There is overlap in item 2 and 3. Item 1 and 4 and the total scale show no overlap, indicating that the differences are statistically significant.

Scale 5: Perceived Feedback

Item	95% CI for Item and Scale Score		95% CI for Scalability	
	IDN	NL	IDN	NL
1	(1.939, 1.981)	(2.006, 2.054)	(0.409, 0.449)	(0.591, 0.635)
2	(1.918, 1.962)	(1.855, 1.905)	(0.453, 0.493)	(0.723, 0.759)
3	(2.197, 2.243)	(1.975, 2.025)	(0.522, 0.554)	(0.696, 0.728)
4	(2.547, 2.593)	(2.005, 2.055)	(0.507, 0.543)	(0.710, 0.742)
5	(2.627, 2.673)	(2.005, 2.055)	(0.476, 0.5122	(0.668, 0.704)
Total Scale	(11.247, 11.413)	(9.854, 10.066)	(0.476, 0.508)	(0.679, 0.711)

Note. The sample size (n) for this scale is 6256 for Indonesia and 4100 for the Netherlands. There is no overlap in scale score and item scores, indicating that the differences are statistically significant.

Note. There is no overlap in item scalability and the total scale scalability, indicating that the differences are statistically significant.