ACTA DIDACTICA NAPOCENSIA
Volume 19, Number 1, 2026 – DOI: 10.24193/adn.19.1.6

RE-EVALUATING TIMSS ATTITUDE SCALES: VALIDITY AND FACTOR STRUCTURE WITIN A SOUTH AFRICAN CONTEXT

Marien Alet GRAHAM

Abstract: This study investigates the validity and underlying factor structure of four Trends in International Mathematics and Science Study (TIMSS) 2019 student attitude scales — Like Learning Mathematics, Like Learning Science, Confident in Mathematics, and Confident in Science — within the South African Grade 5 context. Using secondary data from the TIMSS 2019 cycle, exploratory factor analysis was conducted to evaluate whether the one-factor structures proposed by TIMSS hold true in this setting. Results reveal that, contrary to the original one-factor design, two-factor structures emerged for all four scales, with negatively worded items forming distinct sub-scales. This finding raises important questions about the appropriateness of mixed-worded scales for younger learners, particularly in contexts like South Africa, where linguistic diversity and educational disparities might influence comprehension. These results highlight the need for careful validation of international large-scale assessment (ILSA) tools in diverse settings to ensure accurate measurement and meaningful interpretation of student attitudes. Recommendations are provided for adapting TIMSS scales to better reflect the unique characteristics of developing nations.

Key words: international large-scale assessment (ILSA); mathematics; science; Trends in International Mathematics and Science Study (TIMSS) 2019; validity

1. Introduction

For decades, researchers have relied on data from international large-scale assessments (ILSAs), such as the Trends in International Mathematics and Science Study (TIMSS), to make sense of the status of their educational system (e.g., TIMSS 2003 (Marchis, 2009), TIMSS 2007 (Ciascai, 2009); TIMSS 2011 (Ciascai & Dulamă, 2013), TIMSS 2015 (Kartianom & Retnawati, 2018), TIMSS 2019 (Chen, 2022)). TIMSS started in 1995 and has monitored trends in maths and science (M&S) achievement every four years with the most recent (TIMSS 2023) being the eighth such assessment (International Association for the Evaluation of Educational Achievement [IEA], 2023). TIMSS collects data at Grade 4 and Grade 8 level; however, South African participates at Grade 5 and Grade 9 level, because of South Africa's overall low performance in previous rounds of TIMSS (Reddy et al., 2015). The focus of the current study is Grade 5 level, as research in the early grades is crucial in educational research as it lays the foundation for lifelong learning, addresses developmental milestones, and informs strategies to enhance literacy, numeracy, and social-emotional skills during the most formative years of a child's education.

For TIMSS 2019, at Grade 4/5 level, 64 countries participated, with South Africa being amongst the lowest of the 64 countries with an achieved score of 374 and 324 for math and science, respectively, which is well below the international benchmark of 500 points (Mullis et al., 2020). TIMSS sets the low benchmark at 400 points, with a score above 400 indicating that learners acquired basic math and science knowledge, and the South African score is below this low benchmark. It should be noted that, while writing this article and analysing the TIMSS 2019 data, the TIMSS 2023 results were released; however, the TIMSS 2023 data had not yet been released to the public, so we could not make use of the most recent TIMSS cycle (TIMSS 2023) for our analysis. However, the TIMSS 2023 results also portray poor M&S outcomes for Grade 5 South African students as both M&S scores decreased from TIMSS 2019 to TIMSS 2023: maths decreased from 374 to 362, and science decreased from 324 to 308 (von Davier et al., 2024). Thus, the TIMSS results are showing devasting results for South Africa, and one has to

Received January 2025.

Cite as: Graham, M. A. (2026) Re-evaluating TIMSS attitude scales: Validity and factor structure within a South African context. *Acta Didactica Napocensia,* 19(1), 76-86, https://doi.org/10.24193/adn.19.1.6

question whether this ILSA (TIMSS) takes into account the unique characteristics of South Africa. While ILSAs aim to provide a benchmark for global educational performance, their limitations — such as cultural bias, neglect of socio-economic disparities, overemphasis on rankings — raise critical questions about their relevance for developing nations (Johansson, 2020; Walker et al., 2023). These criticisms surrounding ILSAs have been around for many years (Popham, 1999), and, moreover, many researchers have identified threats to the validity of ILSAs (Baker & LeTendre, 2005; Berliner, 2020). This study aims to investigate the validity and underlying factor structure of TIMSS 2019 scales measuring Grade 4/5 students' attitudes. The contribution of this study is threefold. First, it provides an empirical re-evaluation of the factor structure of TIMSS 2019 attitude scales in a low-performing and linguistically diverse context. Second, it examines whether mixed-worded items introduce method effects that distort the intended unidimensional structure. Third, it offers practical implications for the adaptation of ILSA instruments in developing countries.

2. Literature Review

TIMSS ensures the validity of its scales through a rigorous process that includes the use of a comprehensive assessment framework collaboratively developed with participating countries. This framework encompasses both content and cognitive dimensions. Items are subjected to multiple levels of review and piloting across diverse educational systems to confirm their appropriateness and reliability (Yin & Fishbein, 2020). Although prior research has raised concerns about the use of mixed-worded items in international large-scale assessments, empirical evidence remains limited with regard to younger learners, particularly in linguistically diverse and low-performing contexts such as South Africa. Moreover, existing studies tend to focus on single scales or isolated constructs, offering limited insight into whether similar measurement issues emerge consistently across different attitudinal domains within the same assessment cycle. This study addresses this gap by systematically re-examining the factor structure of four TIMSS 2019 student attitude scales—spanning both mathematics and science—using a unified analytical approach within the South African Grade 5 context. Although the IEA goes through a rigorous process to ensure that its scales are reliable and valid, there are some issues to consider; for instance, low component loadings of some items onto their components for some countries (such as South Africa; Yin & Fishbein, 2020). Other issues include, for example, that studies have shown that questionnaire scales that are mixed-worded, i.e., include both positively and negatively worded items, often suffer from issues like low reliability and more complex latent structures than intended. Recently, Steinmann et al. (2022) have shown this to be the case for TIMSS 2011 and PIRLS 2011 scales, with PIRLS standing for "Progress in International Reading Literacy Study". TIMSS measures students' attitudes about M&S at Grade 4/5 level using four scales (Mullis et al., 2020): "Students Like Learning Mathematics" (p. 25), "Students Like Learning Science" (p. 428), "Students Confident in Mathematics" (p. 433) and "Students Confident in Science" (p. 436); however, all four these scales have mixed-worded items. In addition, for South Africa, the component loadings of these scales are unacceptable. For conciseness, in this article, we'll refer to LikeMath, LikeScience, ConfMath, and ConfScience, respectively, the four scales under consideration. Each scale consists of Likert-scale items measuring the scale, for example, for the LikeMath scale, two examples of items are "I enjoy learning mathematics" and "I wish I did not have to study mathematics" (Mullis et al., 2020, p. 425). The latter would, of course, be reverse-scored, so that all items measure the enjoyment of math, i.e., larger values on the scale indicate higher enjoyment of math. Each scale has scale cut scores:

- < 8.4 "Do Not Like", 8.4-10.2 "Somewhat Like", > 10.2 "Very Much Like" (LikeMath scale; Mullis et al., p. 425)
- < 7.6 "Do Not Like", 7.6-9.7 "Somewhat Like", > 9.7 "Very Much Like" (LikeScience scale; Mullis et al., p. 428)
- < 8.5 "Not Confident", 8.5-10.7 "Somewhat Confident", > 10.7 "Very Confident" (ConfMath scale; Mullis et al., p. 433)
- < 8.2 "Not Confident", 8.2-10.2 "Somewhat Confident", > 10.2 "Very Confident" (ConfMath scale; Mullis et al., p. 436)

At Grade 5, a developmental stage characterised by varying levels of cognitive and linguistic maturity, it is critical to ensure that students understand the items presented to them in attitude scales. This study's focus on the four TIMSS attitude scales — LikeMath, LikeScience, ConfMath, and ConfScience — allows for an exploration of whether the structure of these scales aligns with the developmental and contextual characteristics of young learners in South Africa. The use of mixed-worded items, which include both positively and negatively phrased statements, presents an additional layer of complexity that might challenge young learners' comprehension and response consistency. By focusing on Grade 5 learners, we aim to evaluate the validity of these scales in capturing attitudes accurately within this demographic. Table 1 shows the percentage student responses and average achievement per category for the four scales under consideration.

Table 1. *Percentage students and average achievement for the four scales under consideration*

		% students	Ave achievement	% students	Ave achievement	% students	Ave achievement	Ave scale score
		Very much like		Somewhat like		Do not like		
LikeMath	SA	46 (1.5)	415 (3.3)	43 (1.1)	345 (3.7)	12 (0.6)	334 (5.4)	10.3 (0.05)
	Intern. Ave	45 (0.2)	520 (0.5)	35 (0.1)	491 (0.6)	20 (0.1)	479 (0.7)	
	Diff	+1	-105	+8	-146	-8	-145	
LikeScience	SA	43 (1.2)	390 (5.2)	45 (0.9)	284 (5.1)	11 (0.5)	264 (6.3)	9.8 (0.05)
	Intern. Ave	52 (0.2)	506 (0.5)	36 (0.1)	478 (0.6)	12 (0.1)	467 (0.9)	
	Diff	-9	-116	+9	-194	-1	-203	
		Very confident		Somewhat confident		Not confident		
ConfMath	SA	17 (0.7)	456 (4.9)	53 (0.5)	371 (3.6)	31 (0.9)	340 (3.7)	9.4 (0.03)
	Intern. Ave	32 (0.1)	545 (0.6)	44 (0.1)	497 (0.5)	23 (0.1)	456 (0.6)	
	Diff	-15	-89	+9	-126	+8	-116	
ConfScience	SA	23 (0.8)	421 (6.3)	47 (0.7)	317 (5.4)	30 (0.7)	274 (5.2)	9.3 (0.03)
	Intern. Ave	38 (0.2)	520 (0.6)	43 (0.1)	486 (0.6)	19 (0.1)	453 (0.7)	
	Diff	-15	-99	+4	-169	+11	-179	

Notes: Standard errors appear in parentheses
Ave = Average, Intern. Ave = International average, Diff = Difference
Values taken from Mullis et al. (2020)

When comparing the percentage of South African students that "do not like" M&S to the international average, the percentage is lower (8% less for math and 1% less for science), which is promising, as fewer South African students dislike M&S when compared to the international average. On the other hand, when comparing the percentage of South African students that are "not confident" in M&S to the international average, the percentages are higher (8% higher for math and 11% higher for science). Thus, although it appears that South African students do not dislike M&S, they are not confident in it compared to an international audience.

When comparing achievement, as expected, South Africa's achievement is much lower than the international average, but the interesting information from Table 1 is that the differences between South Africa's averages and the international averages increase the more students dislike M&S and the more students are not confident in M&S. South African students that don't like math have achieved 145 points less than the international average; for science, this value is 203 points less. South African students not confident in math have achieved 116 points less than the international average; for science, this value is

179 points less. Although the information that these four scales are providing us is interesting, one has to question whether these scales are valid within a South African context. The reason for questioning the validity in these scales, is in the fact that the component loadings per item for South Africa is very low. The loadings, as reported by Yin and Fishbein (2020), for South Africa, are shown in Tables 2 to 5. To give the readers some perspective of how long the loadings are, they are contrasted to those of Singapore, the top-performing country for Grade4/5 in TIMSS 2019 in both M&S (Mullis et al., 2020).

It should be noted that the Cronbach's alpha coefficients (α) for the four scales for South Africa are 0.80 (LikeMath), 0.79 (LikeScience), 0.68 (ConfMath) and 0.66 (ConfScience) (Yin & Fishbein, 2020). The acceptable value for Cronbach's alpha has been debated in the literature, with most researchers having settled on > 0.70. This would mean that the first two coefficients are acceptable, but the latter two are not. This being said, it should be pointed out that α is flawed because it goes up as the number of items on a scale goes up (Hayes & Coutts, 2020); thus, theoretically, if one develops a new scale, one can simply add enough items and your α would be acceptable. Accordingly, α will not be considered further in this article, with our focus being placed on the loadings.

Table 2. *Scale loadings for each item of the LikeMath scale for South Africa*

Item	Loading
"I enjoy learning mathematics"	0.72
"I wish I did not have to study mathematics" (RS)	0.29
"Mathematics is boring" (RS)	0.35
"I learn many interesting things in mathematics"	0.62
"I like mathematics"	0.79
"I like any schoolwork that involves numbers"	0.71
"I like to solve mathematics problems"	0.69
"I look forward to mathematics lessons"	0.70
"Mathematics is one my favorite subjects"	0.73

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.89) and the loadings were taken from Yin and Fishbein (2020, p. 16.93)

From Table 2 it can be seen that the loadings range from 0.29 to 0.79 for South Africa, whereas from Yin and Fishbein (2020, p. 16.93), those for Singapore range from 0.72 to 0.90.

Table 3. *Scale loadings for each item of the LikeScience scale for South Africa*

Item	Loading
"I enjoy learning science"	0.69
"I wish I did not have to study science" (RS)	0.29
"Science is boring" (RS)	0.35
"I learn many interesting things in science"	0.66
"I like science"	0.78
"I look forward to learning science in school"	0.71
"Science teaches me how things in the world work"	0.68
"I like to do science experiments"	0.69
"Science is one of my favorite subjects"	0.70

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.96) and the loadings were taken from Yin and Fishbein (2020, p. 16.100)

From Table 3 it can be seen that the loadings range from 0.29 to 0.78 for South Africa, whereas from Yin and Fishbein (2020, p. 16.100), those for Singapore range from 0.55 to 0.89.

Table 4. *Scale loadings for each item of the ConfMath scale for South Africa*

Item	Loading
"I usually do well in mathematics"	0.20
"Mathematics is harder for me than for many of my classmates" (RS)	0.69
"I am just not good at mathematics" (RS)	0.70
"I learn things quickly in mathematics"	0.21
"Mathematics makes me nervous" (RS)	0.66
"I am good at working out difficult mathematics problems"	0.16

"My teacher tells me I am good at mathematics"	0.12
"Mathematics is harder for me than any other subject" (RS)	0.74
"Mathematics makes me confused" (RS)	0.69

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.76) and the loadings were taken from Yin and Fishbein (2020, p. 16.80)

From Table 4 it can be seen that the loadings range from 0.12 to 0.74 for South Africa, whereas from Yin and Fishbein (2020, p. 16.80), those for Singapore range from 0.47 to 0.79.

Table 5. *Scale loadings for each item of the ConfScience scale for South Africa*

Item	Loading
"I usually do well in science"	0.21
"Science is harder for me than for many of my classmates" (RS)	0.77
"I am just not good at science" (RS)	0.74
"I learn things quickly in science"	0.29
"My teacher tells me I am good at science"	0.14
"Science is harder for me than any other subject" (RS)	0.79
"Science makes me confused" (RS)	0.69

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.83) and the loadings were taken from Yin and Fishbein (2020, p. 16.86)

From Table 5 it can be seen that the loadings range from 0.14 to 0.79 for South Africa, whereas from Yin and Fishbein (2020, p. 16.86), those for Singapore range from 0.51 to 0.80.

When considering acceptable loadings, various cut-off values for the factor loadings have been advocated in the literature, spanning a range from 0.4 to 0.7, with the prevailing recommendation, however, being in the range of 0.4 to 0.6 (Steenkamp & Maydeu-Olivares, 2023). Additionally, Lambert and Newman (2023) note that a cut-off value of 0.4 is the most commonly employed in practical applications; however, they caution that this is a fairly low standard, with larger loadings being generally better. From the results above, it can be seen that all the loadings for Singapore are above 0.4; however, South Africa has loadings as low as 0.12 which is unacceptable. Therefore, in this article we challenge the acceptance of a one-factor structure for each of the four scales under consideration within a South African setting, and conduct an exploratory factor analysis (EFA) for each scale using the South African Grade 5 data to explore the underlying factor structure of each scale. Although a one-factor scale may be reliable and valid for some countries (e.g., a high-income country like Singapore), it may not hold for others (e.g., a developing country such as South Africa that is very unique in that it is rich in diversity, for example, South Africa has 12 official languages).

3. Methodology

3.1. Research design and participants

A quantitative design with a positivist philosophical stance was utilised. With the positivist paradigm, the researcher focuses on observable and measurable facts, and the researcher is neutral, detached, and independent of what is researched (Maksimovic & Evtimov, 2023). In this study, we are detached because we didn't collect the data or talk to the participants; we just considered the TIMSS 2019 data, which are numbers that were seen and measured. Regarding the research strategy, secondary data analysis was used (Mouton, 2001). For South Africa, the realised sample was 297 schools, 294 M&S teachers, 11903 learners and 11720 parents/guardians at Grade 5 level (Reddy et al., 2020).

3.2. Data collection, instruments and quality assurance

The data collection for TIMSS 2019 in South Africa took place in October 2018 (Human Sciences Research Council [HSRC], 2021). Readers who are interested can read Cotter et al. (2020) and LaRoche et al. (2020) for a full explanation of the steps that were taken to create the TIMSS 2019 achievement tools and ensure that the tests were consistent and accurate. To ensure quality assurance, we verified that the data satisfied the assumptions required for the selected statistical methods.

3.3. Data analysis

An EFA using principal axis factoring as the extraction method and promax rotation was conducted to explore the underlying factor structures of the four scales under investigation. We used promax rotation, as it allows for extracted components/factors/sub-scales to be correlated, which was theoretically justified given the conceptual relatedness of the items within each scale. Factor retention was guided by multiple criteria. Specifically, both the Kaiser criterion (eigenvalues > 1) and inspection of scree plots were used to determine the appropriate number of factors; for example, for the LikeMath scale, which consists of nine items, a two-factor solution was supported. As the items on the LikeMath scale are narrowly linked, we used promax rotation which allows for sub-scales to be correlated (which, theoretically, is the case). When conducting an EFA, the Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity were considered to investigate whether the data is suitable for dimension reduction (i.e., factor analysis). The KMO value must be above 0.5 (Schreiber, 2021), which indicates that the data is suitable for factor analysis. The p-value of Bartlett's test of sphericity must be statistically significant ($p > 0.05$), which indicates that sufficient correlations exist among the items to proceed with an EFA (Hair et al., 2019). Communalities and component loadings were considered to make decisions on whether items should be dropped. Güvendir and Özkan (2022) recommends that any item with a communality less than 0.2 be removed. Some literature advocate for removing items with communalities less than 0.3 (Holm et al., 2019) or even those less than 0.4 (Eaton et al., 2019). Although there were some low communalities in the Results Section, low communalities are acceptable for large sample sizes (around > 300) (Güvendir & Özkan, 2022), which is met for the TIMSS dataset. Accordingly, no items are removed based on communalities. Factor loadings were interpreted using a cut-off value of 0.4, as commonly recommended in applied research (Lambert & Newman, 2023).

4. Results

As mentioned earlier, EFA was conducted to explore the underlying factor structures of the four scales under investigation. We start off with the LikeMath scale. The KMO value (= 0.879) is acceptable and the *p*-value of Bartlett's test of sphericity is statistically significant ($\chi^2(36) = 25{,}364.4$, $p < 0.001$). Using the eigenvalue > 1 rule, two factors were extracted explaining 56.5% of the variance. The communalities were considered which range from 0.389 to 0.738 and no items were dropped based on the values of the communalities. Next, the item loadings are investigated to see whether any of the items should be dropped; the highest loading per item is indicated in bold (Table 6).

Table 6. *EFA component loadings for the LikeMatch scale*

	Component	
Item	1	2
"I like mathematics"	**0.799**	-0.012
"Mathematics is one my favorite subjects"	**0.747**	0.011
"I like any schoolwork that involves numbers"	**0.727**	0.026
"I enjoy learning mathematics"	**0.724**	0.003
"I like to solve mathematics problems"	**0.716**	0.014
"I look forward to mathematics lessons"	**0.708**	0.002
"I learn many interesting things in mathematics"	**0.608**	-0.053
"I wish I did not have to study mathematics" (RS)	0.042	**0.868**
"Mathematics is boring" (RS)	-0.048	**0.832**

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.89)
RS = Reverse-scored

Since all loadings in Table 6 are above 0.4, all items were retained and we have established that the one-factor structure for the LikeMath scale, as TIMSS 2019 has it, does not hold for a South African setting and that, rather, a two-factor structure has emerged. Next, we consider the LikeScience scale.

The KMO value (= 0.871) is acceptable and the *p*-value of Bartlett's test of sphericity is statistically significant ($\chi^2(36) = 24{,}440.9, p < 0.001$). Two factors were extracted explaining 56.4% of the variance. Next, the communalities were considered which range from 0.439 to 0.756 and no items were removed based on communalities. Next, the item loadings are investigated later to see whether any of the items should be dropped; the highest loading per item is indicated in bold (Table 7).

Table 7. *EFA component loadings for the LikeScience scale*

	Component	
Item	1	2
"I like science"	**0.783**	-0.018
"Science is one of my favorite subjects"	**0.726**	0.041
"I look forward to learning science in school"	**0.725**	0.004
"I enjoy learning science"	**0.722**	0.053
"I like to do science experiments"	**0.704**	-0.023
"Science teaches me how things in the world work"	**0.678**	-0.047
"I learn many interesting things in science"	**0.659**	-0.016
"I wish I did not have to study science" (RS)	0.042	**0.879**
"Science is boring" (RS)	-0.044	**0.847**

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.96)
RS = Reverse-scored

Since all loadings in Table 7 are above 0.4, all items were retained and we have established that the one-factor structure for the LikeScience scale, as TIMSS 2019 has it, does not hold for a South African setting and that, rather, a two-factor structure has emerged. Next, we consider the ConfMath scale.

The KMO value (= 0.790) is acceptable and the *p*-value of Bartlett's test of sphericity is statistically significant ($\chi^2(36) = 16{,}214.6, p < 0.001$). Two factors were extracted explaining 51.6% of the variance. Next, the communalities were considered which range from 0.478 to 0.582 and no items were removed based on communalities. Next, the item loadings are investigated later to see whether any of the items should be dropped; the highest loading per item is indicated in bold (Table 8).

Table 8. *EFA component loadings for the ConfMath scale*

	Component	
Item	1	2
"Mathematics is harder for me than any other subject" (RS)	**0.762**	-0.016
"Mathematics is harder for me than for many of my classmates" (RS)	**0.723**	0.034
"Mathematics makes me confused" (RS)	**0.720**	0.036
"Mathematics makes me nervous" (RS)	**0.691**	0.066
"I am just not good at mathematics" (RS)	**0.686**	-0.123
"I usually do well in mathematics"	-0.027	**0.723**
"I am good at working out difficult mathematics problems"	0.031	**0.717**
"I learn things quickly in mathematics"	-0.039	**0.712**
"My teacher tells me I am good at mathematics"	0.042	**0.707**

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.76)
RS = Reverse-scored

Since all loadings in Table 8 are above 0.4, all items were retained and we have established that the one-factor structure for the ConfMath scale, as TIMSS 2019 has it, does not hold for a South African setting and that, rather, a two-factor structure has emerged. Next, we consider the ConfScience scale.

The KMO value (= 0.738) is acceptable and the *p*-value of Bartlett's test of sphericity is statistically significant ($\chi^2(36) = 13{,}623.8, p < 0.001$). Two factors were extracted explaining 58.6% of the variance. Next, the communalities were considered which range from 0.517 to 0.667 and no items were removed based on communalities. Next, the item loadings are investigated later to see whether any of the items should be dropped; the highest loading per item is indicated in bold (Table 9).

Table 9. *EFA component loadings for the ConfScience scale*

Item	Component	
	1	2
"Science is harder for me than any other subject" (RS)	**0.817**	-0.012
"Science is harder for me than for many of my classmates" (RS)	**0.792**	0.001
"I am just not good at science" (RS)	**0.728**	0.079
"Science makes me confused" (RS)	**0.720**	-0.068
"I usually do well in science"	-0.004	**0.770**
"My teacher tells me I am good at science"	-0.070	**0.762**
"I learn things quickly in science"	0.073	**0.749**

Note. Direct quotes are from Yin and Fishbein (2020, p. 16.83)
RS = Reverse-scored

Since all loadings in Table 9 are above 0.4, all items were retained and we have established that the one-factor structure for the ConfScience scale, as TIMSS 2019 has it, does not hold for a South African setting and that, rather, a two-factor structure has emerged.

5. Discussion

For all four scales under consideration, which, according to TIMSS 2019, has a one-factor structure, we have found a two-factor structure for each one. It is interesting to note that, for each scale, the negatively phrased items formed their own factor (or sub-scale) for each scale under consideration. One plausible explanation is that, in younger populations such as Grade 5 learners, the cognitive processing of negation may be less stable, increasing the likelihood of misunderstanding or inconsistent interpretation of negatively phrased items. This added cognitive load may partly explain why negatively worded items clustered separately across all four attitude scales. This result is consistent with the finding of Steinmann et al. (2022) that has shown that mixed-worded scales often suffer from issues like low reliability and more complex latent structures than intended. The implications of these findings are critical for interpreting TIMSS data in South Africa. While the scales provide insights into student attitudes, their validity may be compromised in contexts where linguistic and educational disparities exist. Moreover, the broader limitations of ILSAs must be considered. As Popham (1999) and Berliner (2020) argued, the overemphasis on rankings and benchmarks in ILSAs often oversimplifies complex educational realities, especially in developing nations. The current study reinforces the need to approach TIMSS results with caution, particularly when interpreting attitudinal data in countries like South Africa, where socio-economic disparities and linguistic diversity play a significant role in shaping educational outcomes.

6. Limitaitons

TIMSS is an ILSA and ILSAs have been critiqued throughout the years for many things, including cultural bias (Johansson, 2020; Walker et al., 2023) and placing high-stakes pressure on educators and students (Emler et al., 2019). In fact, over the years, many researchers have identified threats to the validity of ILSAs (Baker & LeTendre, 2005; Berliner, 2020). Klemenčič and Mirazchiyski (2018) highlight that the media often oversimplifies or misrepresents the findings of ILSAs by labelling countries as either "good" or "bad" based on their rankings. This approach overlooks the fact that

rankings are relative to the specific countries participating in each assessment cycle, which can vary over time.

7. Conclusions and recommendations

This study explored the validity and factor structures of the four TIMSS 2019 students' attitude scales within the South African Grade 5 context. The results revealed that the one-factor structures proposed by TIMSS do not hold in this setting, with two-factor structures emerging for all four scales. A consistent pattern was observed where negatively worded items formed distinct sub-scales, suggesting that these items may confuse young learners, particularly in linguistically diverse and resource-constrained contexts like South Africa. These findings raise critical concerns about the appropriateness of mixed-worded scales for measuring student attitudes in ILSAs, especially in developing countries. This study underscores the importance of validating ILSA tools in diverse settings to ensure their reliability and meaningfulness across varying contexts. Recommendations include that stakeholders should consider revising mixed-worded scales to ensure they are cognitively accessible to younger learners, particularly in contexts where language may be a barrier. Further research is needed to investigate the role of positively and negatively worded items in influencing the validity and reliability of attitude scales, particularly in linguistically diverse environments. Finally, once the TIMSS 2023 data is released, we plan to use the South African Grade 5 TIMSS 2023 data to conduct a confirmatory factor analysis on the two-factor structure that we found for all four scales under consideration.

8. Ethics approval

No permission was required to evaluate the TIMSS 2019 data, as the database is publicly accessible on the IEA's website (Fishbein et al., 2021).

References

Baker, D., & LeTendre, G. (2005). *National differences, global similarities: World culture and the future of schooling*. Stanford University Press.

Berliner, D. C. (2020). The implications of understanding that PISA is simply another standardized achievement test. In G. Fan & T. S. Popkewitz (Eds.), *Handbook of education policy studies* (pp. 239-258). Springer. https://doi.org/10.1007/978-981-13-8343-4_13

Chen, X. (2022). The effects of individual-and class-level achievement on attitudes towards mathematics: An analysis of Hong Kong students using TIMSS 2019. *Studies in Educational Evaluation*, 72, Article 101113. https://doi.org/10.1016/j.stueduc.2021.101113

Ciascai, L. (2009). Comparative study on Romanian school science curricula and the curriculum of TIMSS 2007 testing. *Acta Didactica Napocensia*, 2(2), 25-34.

Ciascai, L., & Dulamă, M.-E. (2013). What specific science abilities and skills are Romanian students developing during primary education? A comparison with the abilities tested by the TIMSS 2011 inquiry. *Acta Didactica Napocensia*, 6(4), 29-44.

Cotter, K. E., Centurino, V. A. S., & Mullis, I. V. S. (2020). Developing the TIMSS 2019 mathematics and science achievement instruments. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center.

Eaton, P., Frank, B., Johnson, K., & Willoughby, S. (2019). Comparing exploratory factor models of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, 15(2), Art. # 020133, 11 pages. https://doi.org/10.1103/PhysRevPhysEducRes.15.020133

Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education*, 2(3), 279-296. https://doi.org/10.1177/2096531119878964

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database*. TIMSS & PIRLS International Study Center. Available at https://timss2019.org/international-database/downloads/TIMSS-2019-User-Guide-for-the-International-Database-2nd-Ed.pdf. Accessed 1 December 2024.

Güvendir, M. A., & Özkan, Y. Ö. (2022). Item removal strategies conducted in exploratory factor analysis: A comparative study. *International Journal of Assessment Tools in Education*, 9(1), 165-180. https://doi.org/10.21449/ijate.827950

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed). Cencage Learning.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, 14(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Holm, M., Alvariza, A., Fürst, C. J., Öhlen, J., & Årestedt, K. (2019). Psychometric evaluation of the anticipatory grief scale in a sample of family caregivers in the context of palliative care. *Health and Quality of Life Outcomes*, 17(1), Art. # 42, 11 pages. https://doi.org/10.1186/s12955-019-1110-4

Human Sciences Research Council. (2021). *Trends in International Mathematics and Science*. Available at http://www.hsrc.ac.za/en/departments/ied/timsssa#:~:text=The%20release%20of%20the%20Trends,took%20place%20in%20September%202019. Accessed 1 December 2024.

International Association for the Evaluation of Educational Achievement. (2023). *TIMSS: Trends in International Mathematics and Science Study*. IEA. Retrieved December 12, 2024, from https://timssandpirls.bc.edu/timss-landing.html

Johansson, S. (2020). Analysing the (mis)use and consequences of international large-scale assessments. In J. Zajda (Ed.), *Globalisation, ideology and education reforms: Emerging paradigms* (pp. 13-24). Springer. https://doi.org/10.1007/978-94-024-1743-2_2

Kartianom, K., & Retnawati, H. (2018). Why are their mathematical learning achievements different? Re-analysis TIMSS 2015 data in Indonesia, Japan and Turkey. *International Journal on New Trends in Education and Their Implications*, 9(2), 33-46.

Klemenčič, E., & Mirazchiyski, P. V. (2018). League tables in educational evidence-based policy-making: Can we stop the horse race, please? *Comparative Education*, 54(3):309-324. http://dx.doi.org/10.1080/03050068.2017.1383082

Lambert, L. S., & Newman, D. A. (2023). Construct development and validation in three practical steps: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*, 26(4):574-607. https://doi.org/10.1177/10944281221115374

LaRoche, S., Joncas, M., & Foy, P. (2020). Sample design in TIMSS 2019. In M. O. Martin, M. von Davier, & I. V. S Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center.

Maksimovic, J., & Evtimov, J. (2023). Positivism and post-positivism as the basis of quantitative research in pedagogy. *Research in Pedagogy*, 13(1), 208-218. https://doi.org/10.5937/IstrPed2301208M

Marchis, I. (2009). Comparative analysis of the mathematics problems given at international tests and at the Romanian national tests. *Acta Didactica Napocensia, 2*(2), 141-147.

Mouton, J. (2001). *How to succeed in your Master's and doctoral studies: A South African guide and resource book*. Van Schaik.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. TIMSS & PIRLS International Study Center.

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice*, 18, 13–17.

Reddy, V., Winnaar, L., Juan, A., Arends, F., Harvey, J., Hannan, S., Namome, C., & Zulu, N. (2020). *TIMSS 2019 highlights of South African Grade 5 results in mathematics and science*. Department of Basic Education. https://www.timss-sa.org/download/TIMSS-2019_Grade-5_HSRC_FinalReport.pdf

Reddy, V., Zuze, T. L., Visser, M., Winnaar, L., Juan, A., Prinsloo, C. H., Arends, F., & Rogers, S. (2015). *Beyond benchmarks: What twenty years of TIMSS data tell us about South African education*. Human Sciences Research Council.

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy*, 17(5), 1004-1011. https://doi.org/10.1016/j.sapharm.2020.07

Steenkamp, J. B. E., & Maydeu-Olivares, A. (2023). Unrestricted factor analysis: A powerful alternative to confirmatory factor analysis. *Journal of the Academy of Marketing Science*, 51(1), 86-113. https://doi.org/10.1007/s11747-022-00888-1

Steinmann, I., Sánchez, D., van Laar, S., & Braeken, J. (2022). The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments. *Assessment in Education: Principles, Policy & Practice*, 29(1), 5-26. https://doi.org/10.1080/0969594X.2021.2005302

von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., & Yin, L. (2024). *TIMSS 2023 international results in mathematics and science*. TIMSS & PIRLS International Study Center.

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). Culturally responsive assessment: Provisional principles. *ETS Research Report Series*, 2023(1), 1-24.

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center.

Author

Marien Alet GRAHAM, Department of Mathematics Education, College of Education, University of South Africa, Pretoria (South Africa). E-mail: grahama@unisa.ac.za.