

## RESEP WORKING PAPER

Department of  
Economics,  
Stellenbosch  
University

Working Paper No.

**09/25**

**NOVEMBER  
2025**

This paper was  
produced as part of the  
MILAPS project, funded  
by Optima

# Do survey items function equivalently in diverse contexts?

## Evidence from TIMSS and PIRLS in South Africa

### Authors

Glen Takalani,  
Debra Shepherd &  
Heleen Hofmeyr

# ***Do Survey Items Function Equivalently in Diverse Contexts? Evidence from TIMSS and PIRLS in South Africa***

Glen Takalani, Debra Shepherd and Heleen Hofmeyr

## **Abstract**

The Trends in Mathematics and Science Study (TIMSS) and Progress in Reading and Literacy Study (PIRLS) provide important data on learners' academic self-perceptions and motivational beliefs, which are constructs known to influence students' educational outcomes. However, a critical methodological challenge emerges when comparing these measured constructs across socioeconomic groups. Observed score differences in the constructs could reflect genuine differences in underlying self-perceptions and motivation, or simply that groups interpret survey questions differently because items hold different meanings for students from different backgrounds. The study uses multigroup confirmatory factor analysis to investigate measurement invariance of self-concept, intrinsic motivation, extrinsic motivation, and affective engagement across socioeconomic groups using TIMSS 2019 (Grades 5 and 9 mathematics) and PIRLS 2021 (Grade 4 reading) South African data, comparing learners from the poorest 60% of schools (Q1-3) with those from Q4 and Q5 schools. Results reveal substantial measurement non-invariance, most pronounced between Q1-3 and Q5 groups. Failures at metric and scalar levels particularly affect mathematics self-concept and extrinsic motivation. Analysis of the PIRLS data reveals that language of instruction also moderates measurement equivalence. These findings highlight the need for caution when interpreting group differences in learner-reported self-perception and motivational constructs, especially in highly unequal societies such as South Africa.

## 1. Introduction

### 1.1. Background

Educational achievement is an important determinant of a developing country's international competitiveness and social development (Demir & Gelbal, 2023). International assessments like TIMSS and PIRLS, coordinated by the International Association for the Evaluation of Educational Achievement (IEA), provide critical data for policymakers. Beyond overall achievement, these studies offer insights into performance disparities across demographic groups defined by gender, socioeconomic status, and region (Mullis & Martin, 2013). Analysing this disaggregated data allows policymakers to identify specific needs, evaluate educational equity, and design targeted interventions to improve the system for all learners, making these surveys vital tools for national development.

The student questionnaires in TIMSS and PIRLS capture data on students' self-perceptions, motivation, and engagement, such as academic self-concept, intrinsic and extrinsic motivational beliefs, and school belonging, which are known to influence academic achievement (Wang & Eccles, 2013; Mullis, et al., 2016). Comparing the measures of these self-perceptions and motivational beliefs across groups is common in educational research. A central challenge, however, is that the measurement of these skills can vary significantly across social groups, developmental stages, and gender (Wurster, 2022; Demir & Gelbal, 2023). This creates a fundamental puzzle, namely that when we observe group differences in these constructs, this may reflect genuine disparities in academic self-perceptions and motivational beliefs or could be artefacts of a measurement tool that functions differently across groups. This distinction is critical, as a score difference could indicate a true divergence in the underlying trait (e.g., students in one group genuinely have lower mathematics self-concept) or stem from measurement bias, where the survey items themselves are interpreted and responded to differently due to varying cultural norms, linguistic frameworks, or even life experiences (Wurster, 2022; Demir & Gelbal, 2023; Abulela, et al., 2024). Consequently, to ensure valid comparisons, survey items must be tested for measurement invariance or equivalence. Without this step, our understanding of true group differences in academic self-perceptions and motivational beliefs remains limited.

Measurement invariance is a critical prerequisite for valid cross-group comparisons of learner's self-perceptions and motivational belief constructs in educational research. It confirms that individuals with the same level of an underlying trait have the same probability of achieving a particular score on a survey item, regardless of group membership (Schmitt & Kuljanin, 2008). It assesses the psychometric equivalence of a construct across groups, demonstrating the construct has the same meaning for different populations (Widaman & Reise, 1997; Putnick & Bornstein, 2016). As membership in sociocultural groups can affect how an underlying trait is conceptualized, establishing measurement invariance is a required condition for comparing group means and construct associations (Bryne & van De Vijver, 2010; Wurster, 2022). Comparisons without

established invariance can lead to inaccurate results, invalidating meaningful evaluations. Establishing invariance is essential to ensure observed differences stem from genuine trait differences rather than differing response tendencies (Vandenberg & Lance, 2000).

To perform measurement invariance testing, a theoretical model specifying the relationship between the latent variables (the unobserved construct, e.g., self-concept) and observed variables (survey items) is developed. Two commonly used traditional methodological approaches for measurement invariance testing are Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) (Reise, et al., 1993; Raju, et al., 2002). IRT posits a probabilistic non-linear relationship between a person's level of a latent trait and their response to a specific item. IRT models estimate the probability of a particular response based on item properties: the difficulty parameter (how much of the trait is needed to endorse the item) and the discrimination parameter (how well it distinguishes between individuals with different levels of the trait). Measurement invariance in IRT means these item parameters are consistent across groups (Reise, et al., 1993). This is assessed using Differential Item Functioning (DIF) tests. DIF indicates that individuals from different groups with the same level of the underlying trait have different probabilities of selecting a given response, violating measurement invariance and compromising the validity of group comparisons.

The second method, CFA, assumes a linear relationship between the latent variable and the observed survey items. Each item is modelled as a linear combination of its latent factor plus a unique error term, typically loading onto a single factor for construct clarity. The standard procedure for invariance testing within this framework is Multigroup Confirmatory Factor Analysis (MGCFA), which involves fitting the same CFA model simultaneously across different groups, such as socioeconomic groups, to test for equivalence (Wurster, 2022). This method allows for a structured, hierarchical examination of whether the measurement instrument performs in the same way for all participants, which is essential for drawing fair and accurate conclusions from the data collected in diverse contexts.

MGCFA tests measurement equivalence through a sequential hierarchy of increasingly restrictive models. The first level, configural invariance, tests whether the same factor structure holds across groups, indicating a shared conceptual understanding of the construct (Vandenberg & Lance, 2000). The second level, metric invariance (also weak invariance), tests whether the factor loadings (strength of item-construct relationships) are equal, ensuring items are equally strong indicators for all groups and allowing comparison of relationships like correlations or regression coefficients (Vandenberg & Lance, 2000; Wurster, 2022; Demir & Gelbal, 2023; Ding, et al., 2023). The third level, scalar invariance (also strong invariance), tests whether item intercepts (the baseline or starting values of observed survey items) are equal, ensuring individuals with the same true trait level have the same expected score. This is essential for validly comparing latent mean scores, confirming that the differences reflect genuine disparities in learners' actual academic self-perceptions and motivational beliefs (Wurster, 2022; Demir & Gelbal, 2023; Ding, et al., 2023). However, achieving full invariance at every level is often challenging. When it is not achieved, researchers can investigate partial invariance (Bryne, et al., 1989; Steenkamp

& Baumgartner, 1998), an approach that is adopted in this present study. Partial invariance acknowledges some survey items may function differently across groups due to contextual factors unrelated to the underlying trait, but meaningful group comparisons can proceed if the core meaning of the construct is preserved through a subset of invariant items (Steenkamp & Baumgartner, 1998).

## 1.2. Research objectives

The primary aim of this present study is to evaluate the measurement invariance of learner-reported academic self-perceptions, motivational beliefs, and engagement across socioeconomic groups using data from large-scale educational assessments in South Africa. Specifically, the study examines key learner-reported constructs at three critical grade levels: (1) reading self-concept, intrinsic motivation, and sense of school belonging among Grade 4 learners in PIRLS 2021, and (2) mathematics self-concept, intrinsic motivation, extrinsic motivation, and sense of school belonging among Grade 5 and Grade 9 learners in TIMSS 2019. These grade levels represent important developmental and decision-making junctures in South Africa. Grade 4 captures early literacy development when language of instruction effects may be most pronounced, Grade 5 represents the transition to upper primary schooling where subject-specific self-perceptions solidify, and Grade 9 precedes the transition into the FET phase (Grade 10-12) of the South African education system, when important subject-choice decisions are made.

Previous research that has investigated these learner self-perceptions and motivational constructs in international assessments has either focused on single items, failed to test measurement invariance, or examined cross-national rather than within-country socioeconomic comparisons. This current study addresses these gaps by conducting measurement invariance testing using MGCFA to investigate whether learners from different socioeconomic contexts conceptualize these constructs through the same factor structure (configural invariance), interpret survey items with the same strength of association to underlying constructs (metric invariance), and use rating scales with equivalent reference points (scalar invariance). The study specifically makes comparisons between learners from the poorest 60% of schools (Quintiles 1-3) with those from more advantaged schools (Quintiles 4 and 5). For PIRLS reading constructs, the analysis further investigates how language of instruction at home moderates these measurement properties by comparing learners who do and do not speak the language of the test at home.

Establishing measurement invariance is important before meaningful comparisons of the self-concept and motivational constructs can be made across socioeconomic groups. If full invariance cannot be achieved, this signals that some survey items in the survey function differently for disadvantaged learners compared to learners from wealthier backgrounds, potentially because aspects like "doing well in mathematics," "finding reading interesting," or "feeling proud of school" hold different meanings or importance depending on learners' educational contexts and lived experiences. Such findings would indicate that policy interventions for boosting learner's academic performance should

account for how disadvantage shapes both the measurement and the underlying reality of academic self-perceptions, which in themselves are important determinants of educational outcomes. This research therefore contributes to understanding not only whether disadvantaged South African learners report different levels of academic motivations, but whether the very instruments used to measure these academic self-perceptions and motivational beliefs capture equivalent psychological processes across the socioeconomic divides that characterize South African society.

## 2. Literature Review

### 2.1. Self-concept

Academic self-concept refers to learners' domain-specific perceptions of their competence and abilities. Research has consistently shown that self-concept predicts academic achievement, cognitive engagement, and educational investment choices across diverse contexts (Marsh & Shavelson, 1985; Bandura, 1997; Marsh, et al., 2019). Rather than reflecting a general sense of academic capability, self-concept operates in a domain-specific manner, with learners' mathematics self-concept and reading self-concept functioning as distinct constructs even when actual abilities in these domains are strongly correlated (Marsh, 1986; Marsh et al., 1988). The Internal/External (I/E) frame of reference model explains this pattern. Learners evaluate their abilities through two simultaneous comparison processes, an external frame comparing their performance to peers within the same domain (social comparison), and an internal frame comparing their achievement across different domains (Möller & Marsh, 2013). These frames of reference are highly sensitive to educational context, which raises critical questions about measurement equivalence across contexts. If disadvantaged learners develop self-concepts through fundamentally different comparison processes and reference points than their wealthier peers, survey items asking about "doing well" or "finding subjects difficult" may capture different realities across socioeconomic groups.

A primary concern stems from well-documented disparities in language exposure and cognitive development associated with SES differences (Hart & Risley, 1995; Hoff, 2003; Heckman, 2006; Cunha & Heckman, 2007; Fernald, et al., 2013). For instance, the seminal work of Hart and Risley (1995) revealed that by age three, children from high-SES families are exposed to approximately 30 million more words than their low-SES counterparts, which may lead to differential comprehension of terminology commonly used in self-concept assessments (Solano-Flores & Li, 2009).

Attributional styles also differ by socioeconomic background in ways that may affect how students interpret self-concept items. Research suggests that learners from higher-SES backgrounds tend to attribute academic success to stable, internal factors such as innate ability (Heckman, 2006),<sup>1</sup> while their lower-SES peers more often associate achievement

---

<sup>1</sup> Heckman (2006) argues that families with more resources can invest in environments which may encourage children to view skills as a natural endowment, meanwhile low-SES contexts foster effort-based attributions to acquired skills due to resource constraints, and this affects educational investment returns.

with effort and external factors (Claro, et al., 2016).<sup>2</sup> This difference in how success is conceptualized may affect responses to items like "I am good at mathematics." If high-SES students interpret this as asking about natural ability, which they may have been socialized to believe they possess, they may rate themselves higher. While, if low-SES students interpret the same item as asking about natural ability but have been socialized to view success as primarily effort-based, they may rate themselves lower even when actual ability levels are similar. However, the direction of these effects on measurement functioning requires empirical testing rather than assumption.

While direct evidence on measurement invariance of self-concept across socioeconomic groups remains limited, cross-national studies may provide some relevant insights because countries differ in both cultural orientations and socioeconomic developmental levels. This type of evidence would have implications for within-country socioeconomic comparisons especially for a highly diverse country like South Africa. For instance, Cicero (2020) examined the Self-Concept Clarity Scale among 2 707 East Asian, Southeast Asian, White, Pacific Islander, and Multiracial undergraduate students at a large U.S. university. The study found configural, metric, and scalar invariance across these racial groups, indicating that the 12-item scale measured the construct equivalently across all subgroups. This evidence suggests that when samples are drawn from similar educational contexts, racial or ethnic differences may not necessarily threaten measurement equivalence. Cross-national studies using large-scale assessment data reveal more complex patterns. Jin et al. (2023) evaluated measurement invariance of a self-concept scale using PISA 2018 data from 42 countries. The analysis supported configural invariance, indicating that learners conceptualise self-concept in similar ways across countries. Metric invariance was achieved for 80.5% of factor loadings (169 of 210 factor loadings), with only seven countries showing non-invariance for the item "I usually manage one way or another."<sup>3</sup> Partial scalar invariance was established with 74.2% of the intercepts (156 of 210), demonstrating equivalence. However, numerous developing economies showed non-invariance intercepts for multiple items.<sup>4</sup> The authors suggest that cultural dimensions, such as differences in the extent to which students are expected to take active roles in learning, and individualism versus collectivism, may explain why certain countries struggled to achieve scalar invariance.

Ding et al. (2023) investigated mathematics self-concept (MSC) and self-efficacy (MSE) constructs by testing the measurement invariance of the constructs using PISA 2003 and 2012 data across 40 countries in each cycle. The initial model fit, which assumed all survey items were measuring a single trait, did not fit the data well. The model showed improvement after accounting for residual covariances, indicating that method effects, such as those caused by similarly or negatively worded item, form part as a significant

---

<sup>2</sup> The findings by Claro et al. (2016) suggest that innate ability beliefs, such as having a fixed mindset mediates poverty's impact on achievement.

<sup>3</sup> The item "I usually manage one way or another" showed metric non-invariance (unequal factor loadings) in Australia, B-S-J-Z (China), Dominican Republic, Montenegro, Moscow Region (Russia), Peru, and Portugal, while achieving metric invariance across the remaining 35 countries in the sample.

<sup>4</sup> Non-invariant intercepts were particularly concentrated in developing economies, including Argentina, Baku (Azerbaijan), B-S-J-Z (China), Georgia, Malaysia, Philippines, Saudi Arabia, and Thailand. These countries have distinct cultural orientations towards collectivism and high power-distance, which suggests that both economic development level and cultural norms shape how students interpret and respond to self-concept items.

aspect of the construct's structure and must be explicitly modelled for accurate measurement. Using the alignment method and multigroup CFA, their findings confirmed configural and metric invariance for both mathematics self-efficacy and mathematics self-concept across all models, enabling comparisons of structural relationships. However, scalar invariance was consistently not achieved, even when analyses were restricted to culturally and educationally similar Nordic countries (Denmark, Finland, Norway, and Sweden). This means that even when countries conceptualise self-concept similarly and items function as equally strong indicators, the baseline response levels differ across the countries.

A key finding from Ding et al.'s (2023) study is the high level of non-invariance found particularly in negatively worded items (e.g., items like "I am just not good at mathematics"), which often introduce method effects and threaten measurement equivalence across diverse populations. Method effects is a systematic variance which is attributable to item wording rather than the underlying trait being measured. Despite repeated model refinements and subgroup analyses, the proportion of non-invariant intercepts for both self-concept and self-efficacy exceeded the 25% threshold that Asparouhov and Muthén (2014) propose as acceptable for approximate invariance. This persistent challenge in achieving scalar invariance points to differences that are driven by factors beyond cultural differences.

## 2.2. School belonging

Students' sense of school belonging is defined as the extent to which a student feels accepted, respected, included, and supported within their school's social environment (Goodenow & Grady, 1993). Sense of school belonging has emerged as an important predictor of academic outcomes including achievement and motivation (Goodenow, 1993; Osterman, 2000). This construct, reflects the emotional dimension of students' connection to their educational environment. Research has consistently shown that students who also report stronger feelings of school belonging tend to display better academic motivation, self-esteem, and achievement (Goodenow & Grady, 1993; OECD, 2013; Sirin & Rogers-Sirin, 2004; Wang & Holcombe, 2010).

However, these relationships depend on the social desirability attached to academic achievement across different social groups, specifically socioeconomic groups. Socio-economically advantaged students consistently report greater school belonging than disadvantaged students across diverse educational systems (Isdale, et al., 2017). In PISA 2015, this pattern held in almost every participating education system (OECD, 2017), though this finding assumes measurement equivalence across countries, which is an assumption that subsequent research has shown may not hold (He, et al., 2018; Law, et al., 2022). In the South African context specifically, research shows that SES significantly predicts learners' school belonging (Isdale, et al., 2017). Learners from more affluent backgrounds tend to attend schools with adequate resources (Spaull, 2013; Isdale, et al., 2017), greater safety, and environments that encourage collaboration among peers and teachers, which are factors considered to foster stronger sense of belonging (Masitsa, 2008; Martinot, et al., 2022). Meanwhile, learners in under-resourced schools

characterized by infrastructure deficits, safety concerns, and limited opportunities for positive student-teacher relationships may develop fundamentally different emotional connections to their school environments.

The socioeconomic patterns of school belonging raises important questions about measurement equivalence. Cross-national evidence on measurement invariance of the school belonging constructs reveals varying challenges in establishing equivalence across diverse contexts, especially those that would be relevant for a diverse country context such as South Africa. For example, using Grade 8 data from the 2015 TIMSS and PISA data, He et al. (2019) found that the sense of school belonging scale achieved only metric invariance in TIMSS across the 29 countries participating in both assessments. In contrast, the PISA data failed to meet the metric invariance criterion, which suggests that for both assessments, comparisons of the mean score of the measured constructs would be invalid due to the failure in testing beyond metric invariance. Similarly, Abubakar et al. (2015) investigated the Psychological Sense of School Membership (PSSM; Goodenow, 1993) across two non-Western countries (Kenya and Indonesia) and two Western countries (Netherlands and Spain). The authors reported that the original unidimensional factor structure proposed by Goodenow (1993) did not adequately fit across these contexts. Multiple models were tested adjusting for the impact of negative worded items and also proposing a two-factor model that distinguished between positively and negatively worded items, but the model still did not fit the data well. Examining measurement invariance of school connectedness, a measure closely related to school belonging, Law et al. (2022) confirmed configural and metric invariance of the the 5-item school connectedness construct across Canadian, Chinese, and Tanzanian adolescents. Partial scalar invariance could only be established for all the comparisons, which means that students with the same true level of school connectedness used different reference points when rating themselves on items about school enjoyment and school safety.

### **2.3. Intrinsic and extrinsic motivation**

Within the Expectancy-Value Theory framework (EVT; Eccles, et al., 1983; Wigfield & Eccles, 2000, 2002), learner task values represent the perceived worth or importance learners assign to academic activities. The TIMSS and PIRLS assessment frameworks contain two constructs that form part of this framework, namely intrinsic motivation and extrinsic motivation. Intrinsic motivation refers to engaging in a specific academic domain because it is inherently interesting, enjoyable, or satisfying, without external benefit or pressures (Eccles, et al., 1983; Ryan & Deci, 2009). Extrinsic motivation refers to engaging in an activity to obtain a separable outcome (Eccles, et al., 1983), such as a better career outcome or further educational opportunities (Nagengast & Marsh, 2014). When learners value mathematics primarily for its extrinsic benefit rather than for intrinsic enjoyment, learning becomes instrumental, as a means to an end rather than an end in itself (Ryan & Deci, 2016). Both motivational constructs have been shown to predict academic achievement, subject selection, and persistence across different educational contexts (Wigfield & Cambria, 2010).

Socioeconomic status shapes how students develop motivational beliefs through distinct socialization pathways. Students from lower-SES often prioritize the extrinsic benefits of education, such as escaping poverty, but may simultaneously doubt the attainability of these outcomes given structural barriers they observe in their communities (Destin, et al., 2019). Similarly for intrinsic motivation development, learners from higher-SES backgrounds are exposed to "concerted cultivation" parenting (Lareau, 2003), which provides them with resource-intensive enrichment experiences that nurture interest and enjoyment in specific academic domains (Bodovski & Farkas, 2008; Simpkins, et al., 2012). In the South African context, where socioeconomic disparities intersect with cultural differences (Branson, et al., 2024), these differences in socialization processes can be exacerbated by disparities in school resources, curriculum coverage, and exposure to career role models.

Research on measurement invariance of motivational constructs across socioeconomic and cultural contexts remains limited. To our knowledge, only one study by Liou & Lin (2021) has examined measurement invariance of motivational beliefs across cultural contexts. Using PISA data from 2015, the authors investigate measurement invariance among adolescents from Taiwan, Australia, and the United States. After accounting for the effects of the negatively worded items, the authors could only establish configural and metric invariance. This means that students from Taiwan, Australia, and the United States with the same level of academic motivation used different reference points when responding to PISA survey items about motivation, making direct comparisons of mean scores of motivation levels invalid.

### 3. Data and methodology

#### 3.1. Data

Conducted every four years and five years since 1995 and 2001, respectively, the TIMSS and the PIRLS are international studies developed by the IEA and managed by the TIMSS and PIRLS International Study Centre. TIMSS aims to provide a detailed picture of performance in mathematics and science across countries and over time, using nationally representative samples of Grade 4 and Grade 8 learners and their schools. PIRLS aims to evaluate reading literacy among Grade 4 students worldwide, capturing how well students can understand and interpret written texts. The study is designed to assess both reading comprehension and the influence of school and home environments on students' literacy development. This study will make use of the TIMSS conducted in 2019 and PIRLS conducted in 2021.

The sampling technique for TIMSS and PIRLS is a two-stage stratified cluster sampling method design, which ensures that the sample selected is representative of the population of learners in South Africa. In the first stage, schools are selected using stratification, where the schools are grouped based on key characteristics and selected within each stratum using probability-proportional-to-size sampling. The second stage samples the classes and learners from each school, where a single intact Grade 4 class in PIRLS and Grade 9 or Grade 5 class in TIMSS are randomly chosen. Every learner from

each selected class participates in the assessment, such that the learners themselves are clustered within their schools and classrooms.

In the 2019 wave of TIMSS, achievement and contextual data in South Africa was collected from 20 829 Grade 9 learners taught in 519 schools by 543 mathematics and science teachers. This represents a larger sample compared to previous rounds, as the Gauteng and Western Cape provinces participated as 'benchmarking participants'. This meant that whilst the usual 30 schools were sampled from the remaining seven provinces, 150 schools each were sampled from Gauteng and the Western Cape. In the Grade 5 sample, data was collected from 11 891 learners taught in 297 schools by 297 mathematics and science teachers. In PIRLS 2021 for South Africa 12 426 Grade 4 learners were sampled from 321 schools. Gauteng, KwaZulu-Natal, and the Eastern Cape had the largest representation in the study due to their higher population densities. Sample weights are provided in the PIRLS and TIMSS data so that data from each province in the sample makes an appropriately sized contribution to the overall national performance.

Contextual and background data are collected through questionnaires administered to teachers, learners, and schools. For the TIMSS 2019, South Africa administered the paper-based assessment and questionnaires in English and Afrikaans, following a translation process overseen by the TIMSS and PIRLS International Study Center and verified by IEA Amsterdam in collaboration with capstan Linguistic Quality Control. The translation process required the participation of qualified translators with knowledge of both English and the target language, experience with the country's cultural context, and expertise in mathematics and science education. Translators ensured that the texts in the questionnaires maintained the same register, correct grammar, equivalent qualifiers and modifiers, and appropriate adaption of idiomatic expressions without adding and removing information. After translations, the instruments are reviewed to document quality and comparability to international versions. Any inconsistencies identified were directed back to the National Research Coordinator for clarification and correction.

For PIRLS 2021, South Africa adapted the reading assessment and questionnaires into all 11 official South African languages for Grade 4 learners. The translation process evolved into a five-step procedure, which involved forward translation from English to the 10 other official languages, and then back-translation from the target language to English. Language specialists then revised both translated versions of the instruments, and consulted with the Department of Basic Education and language experts, who assisted in the reconciliation of discrepancies against the source text, with all changes documented in the National Adaption Forms. This process ensured that every language version underwent extensive quality checks before printing, addressing the complexity of preserving text equivalence, item difficulty, and comparability across languages with different linguistic structures (Roux, et al., 2022; DBE, 2023). Students completed questionnaires in the language of the reading test they took, which for Grade 4 learners is also the language of learning and teaching (LOLT).

### 3.2. Measures

The survey items that measure learners' underlying academic self-perceptions and motivations are obtained from the TIMSS and PIRLS student questionnaire. All items were coded on a four-point Likert scale, with 1 indicating "Agree a lot" and 4 indicating "Disagree a lot". For the purposes of this study, reverse-scoring was adopted so that higher values represent more favourable perceptions and attitudes.

#### 3.2.1. Mathematics and reading self-concept

Mathematics self-concept is measured using seven items for TIMSS for both Grades 5 and 9 learners, and six items for PIRLS reading self-concept, as shown in Table 1. Three of the seven items are negatively worded statements (e.g. "Mathematics is not one of my strengths") in the TIMSS student questionnaire, and were reverse-scored to match the positive statements. Four of the six items in PIRLS were negatively worded statements, and were reverse-scored as well (e.g., "Reading is harder for me than any other subject").

**Table 1:** Self-concept questionnaire items

Mathematics Self-Concept Grade 5 and 9	Reading Self-Concept Grade 4
1) I usually do well in mathematics	1) I usually do well in reading
2) Mathematics is more difficult for me than for many of my classmates*	2) Reading is easy for me
3) Mathematics is not one of my strengths*	3) I have trouble reading stories with difficult words*
4) I learn things quickly in mathematics	4) Reading is harder for me than for many of my classmates*
5) I am good at working out difficult mathematics problems	5) Reading is harder for me than any other subject*
6) My teacher tells me I am good at mathematics	6) I am just not good at reading *
7) Mathematics is harder for me than any other subject*	

*Note:* \* refers to reverse-scored negatively-worded items.

#### 3.2.2. School belonging

School belonging (SB) is measured using five items in the TIMSS survey for Grades 5 and 9 learners, and six items in the PIRLS survey for the Grade 4 learners. The PIRLS survey includes one additional item not found in TIMSS, which asks learners if they have friends at school. A full list of the items is provided in Table 2.

**Table 2:** School belonging questionnaire items

TIMSS school belonging Grade 5 and 9	PIRLS school belonging Grade 4
1) I like being in school	1) I like being in school
2) I feel safe when I am at school	2) I feel safe when I am at school
3) I feel like I belong at this school	3) I feel like I belong at this school
4) Teachers at my school are fair to me	4) Teachers at my school are fair to me
5) I am proud to go to this school	5) I am proud to go to this school
	6) I have friends at this school

### 3.2.3. Intrinsic motivation

Intrinsic motivation is measured using nine items in the TIMSS instrument for both Grade 5 and Grade 9 students. Two of the items are reverse-scored because they are negatively worded items (e.g. "Mathematics is boring"). Eight items were used to measure interest and enjoyment in engaging in reading activities, with only one item being a negatively worded statement (e.g., "I think reading is boring").

**Table 3:** Intrinsic motivation questionnaire items

Mathematics Intrinsic Motivation Grade 5 and 9	Reading Intrinsic Motivation Grade 4
1) I enjoy learning mathematics	1) I like talking about what I read with other people
2) I wish I did not have to study mathematics*	2) I would be happy if someone gave me a book as a present
3) Mathematics is boring*	3) I think reading is boring*
4) I learn many interesting things in mathematics	4) I would like to have more time for reading
5) I like mathematics	5) I enjoy reading
6) I like any schoolwork that involves numbers	6) I learn a lot from reading
7) I like to solve mathematics problems	7) I like to read things that make me think
8) I look forward to mathematics class	8) I like it when a book helps me imagine other worlds
9) Mathematics is one of my favourite subjects	

Note: \* refers to reverse-scored negatively-worded items.

### 3.2.4. Extrinsic motivation

Extrinsic motivation is captured through nine items in the Grade 9 TIMSS data. This motivational construct is only captured in the Grade 9 TIMSS data. None of the items for this construct were negatively worded statements. The items making up this construct are shown in Table 4.

**Table 4:** Extrinsic motivation questionnaire items

---

<b>Mathematics Extrinsic Motivation Grade 9</b>
1) I think learning mathematics will help me in my daily life
2) I need mathematics to learn other school subjects
3) I need to do well in mathematics to get into the university of my choice
4) I need to do well in mathematics to get the job I want
5) I would like a job that involves using mathematics
6) It is important to learn about mathematics to get ahead in the world
7) Learning mathematics will give me more job opportunities when I am an adult
8) My parents think that it is important that I do well in mathematics
9) It is important to do well in mathematics

---

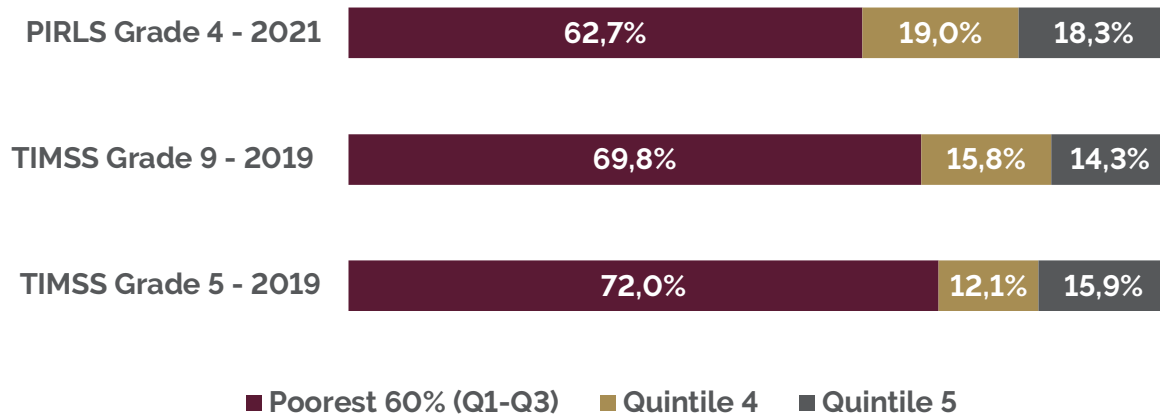
### 3.2.5. Socio-economic status

The South African education system functions as a bimodal system which serves two groups of learners in the system but operating under one umbrella. The one system serves a group of learners that come from 20% of the wealthiest households (Hofmeyr, 2022), who attend schools that are mostly functional and perform at the standard of schools in high-income countries (Spaull & Taylor, 2012). The second part of the system serves the other majority 80% of households, attending schools that are under-resourced and perform poorly even at the standard of an upper middle-income country (Spaull & Taylor, 2012; Hofmeyr, 2022). Therefore, in the present study the SES variable is split-up to reflect the two separate education systems in which measurement comparisons of the self-perception and motivational constructs are made.

The socioeconomic status variable is calculated by the DBE as a poverty index for each public school according to the poverty of the community around the school, as well as certain infrastructural factors such as access to electricity and running water (Reddy, et al., 2020). The public schools are categorised into five equal different school socioeconomic quintiles with quintile 1 to quintile 3 (Q1-3) representing the poorest 60 percent (%) of schools which are the most under-resourced schools and are typically no-fee charging schools. While quintile 4 (Q4) and quintile 5 (Q5) are allowed to charge fees,

and Q5 schools are some the most well-resourced schools in South Africa. The weighted distribution of learners across the different wealth groups is shown in Figure 1 below.

**Figure 1:** Weighted distribution of learners across different SES groups



*Note:* The distribution shown is based on weighted survey estimates, not simple headcounts. Each student's response is multiplied by their survey weight (*totwgt*), which accounts for the stratified sampling design used in TIMSS/PIRLS. This ensures the results are representative of all Grade4/5/9 learners in South Africa. Percentages reflect the weighted proportion of learners in each school wealth quintile within the national population.

### 3.3. Methodology

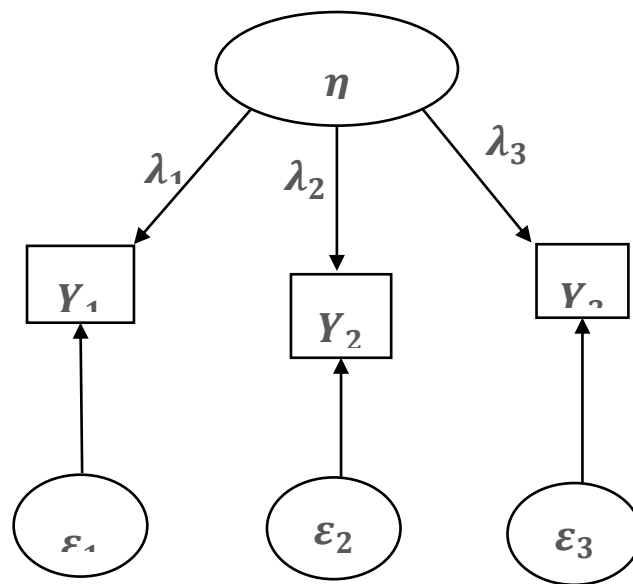
#### 3.3.1. The single-group reflective measurement model

A single-group reflective measurement model is developed for the latent variables discussed in Section 3.3. This approach is foundational to Structural Equation Modelling (SEM) and is defined by its specific causal direction: latent variables are theorized to cause the variations in the observed indicators (Bollen & Lennox, 1991; Chen, 2007). This contrasts with a formative model, where indicators are combined to form the latent variable. Since learner responses to self-perception and motivational questionnaire items require them to reflect on their underlying beliefs and motivations, a reflective measurement model is appropriate for the purposes of this study. The reflective measurement model will be used to test for measurement invariance across groups, a process described by Horn and McArdle (1992) as the test of "whether or not, under different conditions of observing and studying a phenomenon, measurement observations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). In this study, "different conditions" refers to the different socioeconomic contexts, which represent different educational environments, and peer reference groups for developing self-perceptions. MI ensures that comparisons across these SES groups are valid, as the measures must reflect the same constructs consistently across diverse populations (Putnick & Bornstein, 2016).

To clarify, Figure 2 presents a single-group measurement model for a latent variable ( $\eta$ ), such as mathematics or reading self-concept, measured by three hypothetical items,  $Y_1$

to  $Y_3$ . These items represent learners' reported perceptions of their capabilities in mathematics or reading. Each item is hypothesized to be associated with the latent variable through the factor loadings  $\lambda_1$  to  $\lambda_3$ . In reflective models, loadings reflect the strength of association between the latent variable and each item, thereby demonstrating how well each item "reflects" the underlying construct (Bollen & Lennox, 1991). Each item has a corresponding measurement error term ( $\varepsilon_1$  to  $\varepsilon_3$ ), which captures variability not explained by the latent variable (Kline, 2016). By incorporating measurement error, we can achieve a more reliable assessment of the latent variable.

**Figure 1:** Single-group Reflective Measurement Model



### 3.3.2. Multiple group confirmatory factor analysis (MGCFAs)

To examine MI, we follow the multi-group factor analysis model below which also follows Figure 2 above:

$$Y_{ipg} = V_{pg} + \Lambda_{pg}\eta_{ig} + \varepsilon_{ipg} \quad (1)$$

In formalizing the MGCFAs  $p = 1, \dots, P$  represents the number of observed items,  $g = 1, \dots, G$  refers to the number of groups to be compared, and  $i = 1, \dots, N_g$  where  $N_g$  denotes the number of individual observations in group  $g$ . Define  $Y_{ipg}$  as the  $P \times 1$  vector of observed item scores for individual  $i$  in group  $g$ . In addition, let  $\Lambda_{pg}$  represent the  $P \times 1$  vector of factor loadings for group  $g$ ,  $V_{pg}$  is the  $P \times 1$  vector of intercepts, and  $\eta_{ig}$ <sup>5</sup> as the latent variable for individual  $i$  in group  $g$ . The measurement error term ( $\varepsilon_{ipg}$ ) follows a normal

<sup>5</sup>Without loss of generality, I restrict the model to a single latent variable  $\eta$  for illustration purposes

distribution,  $N(0, \Theta_{pg})$ , with  $\Theta_{pg} = \text{Cov}(\varepsilon_{pg}, \varepsilon'_{pg})$  being the group-specific error covariance matrix. This leads to the measurement model formulation above in equation 1 (Eq. (1)).

The respective mean and covariance structure for the latent variable are specified as follows (MACS; Little, 1997):

$$\mu_g = V_g + \Lambda_g \alpha_s \quad (2)$$

$$\text{Cov}(Y_g, Y'_g) = \Lambda_s \varphi_s \Lambda'_s + \Theta_{g'} \quad (3)$$

where  $\mu_g$  represents the observed mean of the observed item scores  $Y_g$ , while  $\alpha_s$  and  $\varphi_s$  denote the latent mean and variance of  $\eta_g$ , respectively, which is also assumed to follow a normal distribution:  $\eta_g \sim N(\alpha_g, \varphi_g)$ .

Following established literature (e.g., Meredith, 1993; Leitgob et al., 2023), measurement invariance testing involves a sequence of nested models with progressively stricter constraints, with each level allowing for different types of important comparisons. The first and least restrictive level is configural invariance. This tests whether the basic factor structure is identical across groups. Configural invariance examines whether items like "I usually do well in mathematics," and "mathematics is not one of my strengths" all load onto the same mathematics self-concept construct for both Q1-3 and Q5 learners. Establishing configural invariance would indicate that the groups conceptualize the construct in the same way, but does not allow for quantitative comparisons of relationships or means. Model fit is assessed using three indices including the Comparative Fit Index (CFI), which measures the improvement in fit of the hypothesized model compared to a baseline model. The second index is the Root Mean Square Error of Approximation (RMSEA), which measures how well the model approximates the population data, assuming the model is correct, while penalizing model complexity. The third statistic index used is the Standardized Root Mean Square Residual (SRMR), which measures the average difference between observed and model-predicted correlations. Acceptable fit is indicated by  $\text{CFI} \geq 0.95$ <sup>6</sup>,  $\text{RMSEA} \leq 0.08$ <sup>7</sup>, and  $\text{SRMR} \leq 0.08$ <sup>8</sup> (Cheung & Rensvold, 2002; Chen, 2007; Wurster, 2022).

The second level, metric invariance (or weak invariance), introduces equality constraints on the factor loadings ( $\lambda$  in Figure 2 or in the vector  $\Lambda_{pg}$  (Eq. 1)) across groups. This level of the test checks whether the item "I usually do well in mathematics" has the same strength of association with the underlying mathematics self-concept construct for learners from the poorest 60% of schools as it does for Q5 learners. If metric invariance holds, it justifies comparing relationships between constructs (such as correlations or regression coefficients) across groups. Metric invariance is supported if the model fit does not deteriorate significantly when factor loadings are constrained to equality. The model "deterioration" refers to the change in fit indices when moving from the configural, where

<sup>6</sup> A fit index above the threshold (0.95) means the hypothesized model shows a 95% improvement over the baseline model, and below the threshold the model is generally considered unacceptable.

<sup>7</sup> Fitness value below the threshold (0.08) indicates the model is a relatively plausible representation of the population data.

<sup>8</sup> A smaller value than the threshold (0.08) indicates that the discrepancy between the observed and predicted correlations are very small.

loadings are freely estimated for each group, to the metric model where loading are constrained equal across groups. Negligible deterioration is indicated by  $\Delta CFI \leq -0.010$ ,  $\Delta RMSEA \leq 0.015$ , and  $\Delta SRMR \leq 0.030$ , where delta ( $\Delta$ ) represents the difference between the metric and configural model fit values (Cheung & Rensvold, 2002; Chen, 2007).

**Table 5:** Levels of measurement invariance

Invariance level	What it implies	Group comparisons allowed	How the invariance level may be assessed
<b>Configural Invariance</b>	The same items measuring the same constructs across groups	None	$CFI \geq 0.950$ , $RMSEA \leq 0.08$ , $SRMR \leq 0.08$
<b>Metric Invariance</b>	The same items have the same factor loadings across groups (at least two equal factor loadings for partial invariance)	Unstandardized associations (covariances, unstandardized regression coefficients with other theoretical constructs of interest)	$\Delta CFI \leq 0.010$ , $\Delta RMSEA \leq 0.015$ , $\Delta SRMR \leq 0.030$
<b>Scalar Invariance</b>	The same items have the same factor loadings and intercepts across groups (at least two items with equal factor loadings and intercepts for partial invariance)	Unstandardized associations and latent means	$\Delta CFI \leq 0.010$ , $\Delta RMSEA \leq 0.015$ , $\Delta SRMR \leq 0.015$
<b>Strict Invariance</b>	The same items have the same factor loadings, intercepts, and error variances across groups	Unstandardized associations and latent means	$\Delta CFI \leq 0.010$ , $\Delta RMSEA \leq 0.015$ , $\Delta SRMR \leq 0.010$

Source: (Leitgob, et al., 2023)

The third level, scalar invariance (or strong invariance), adds equality constraints on the item intercepts ( $\nu_g$  in the vector of intercept  $V_{pg}$  (Eq. 1)). Scalar invariance tests whether learners from different groups with the same level of the latent trait would provide the same mean response to items. If two learners, one from Q1-3 schools and one from Q5

schools, both have identical mathematics self-concept, scalar invariance ensures they would have the same expected response to "I usually do well in mathematics." Without scalar invariance, observed score differences could reflect different reference points or rating scale usage rather than true differences in the underlying trait. Establishing scalar invariance is therefore a prerequisite for valid comparisons of the latent means across groups. The fit of the scalar model is compared to the metric model, with nonsignificant deterioration indicated by  $\Delta\text{CFI} \leq -0.010$ ,  $\Delta\text{RMSEA} \leq 0.015$ , and  $\Delta\text{SRMR} \leq 0.015$  (Cheung & Rensvold, 2002; Chen, 2007).

The final and most stringent level is strict invariance, which requires the residual error variances ( $\varepsilon_i$  in Figure 2 and in the vector of errors  $\varepsilon_{ipg}$  (Eq. 1)) to be equal across groups. This ensures that the measurement precision is equivalent for each item across groups. Strict invariance testing would require that the item "I usually do well in mathematics" has the same reliability for learners from Q1-3 schools and Q5 learners. While not always necessary for latent mean comparisons, strict invariance provides the most robust test for such comparisons by ensuring that group differences are not confounded by differences in measurement reliability (Widaman & Reise, 1997; Vandenberg & Lance, 2000). The fit of the strict model is compared to the scalar model using the same change-in-fit thresholds:  $\Delta\text{CFI} \leq -0.010$ ,  $\Delta\text{RMSEA} \leq 0.015$ , and  $\Delta\text{SRMR} \leq 0.010$ . In practice, overall fit is also considered, with thresholds such as  $\text{CFI} \geq 0.950$  being desirable (Rutkowski & Svetina, 2014). Among the change-in-fit indices,  $\Delta\text{CFI}$  is often prioritized for its robustness in detecting non-invariance. This sequential testing procedure ensures that any observed group differences can be attributed to substantive differences in self-perceptions and motivational beliefs rather than measurement artefacts.

Analysis of the CFA and MGCFA was performed using the R software (R Core Team, 2020) with the package lavaan 0.6-7 (Rosseel, 2012). The package semTools and semPaths often used along with lavaan were also used for CFA and MI testing.

#### 4. Internal Consistency Results

The reliability of each construct that is described in section 3.2. above is determined by calculating the Cronbach's alpha value that measures the internal consistency of respondents' answers. For example, a learner agreeing that they enjoy mathematics would be expected to disagree that they find mathematics to be boring. Reliability scores measure the degree to which the items measure the latent constructs (Hair, et al., 2006). An alpha value of 0.7 or greater is taken as suitable (Cronbach & Meehl, 1995; Hair, et al., 2006). The Cronbach estimated scores for the measured constructs is presented in Table 6 below.

The internal consistency of the motivational and self-perception constructs is presented in Table 6 below, and shows varying results across grade levels and socioeconomic contexts. For TIMSS Grade 9, most of the latent constructs demonstrate acceptable reliability ( $\alpha_{Gr9} > 0.7$ ) across quintiles, with intrinsic motivation showing the highest overall

reliability ( $\alpha_{Gr9} = 0.896$ ) and particularly strong consistency in Q5 schools ( $\alpha_{Gr9} = 0.924$ ). Mathematics self-concept reliability scores ranges from  $\alpha_{Gr9} = 0.709$  in Q1 to  $\alpha_{Gr9} = 0.869$  in Q5, suggesting that the self-perception construct may be measured with greater precision among the more advantaged students. Affective engagement shows more modest reliability ( $\alpha_{Gr9} = 0.656$  to  $\alpha_{Gr9} = 0.750$ ), with the lowest values observed in Q1 and Q2 ( $\alpha_{Gr9} = 0.656$ ), potentially indicating that items measuring school belonging and safety function less consistently for learners in the poorest schools. For TIMSS Grade 5 learners, a similar pattern emerges, with the intrinsic motivation construct maintaining strong reliability ( $\alpha_{Gr5} = 0.791$  overall), while the mathematics self-concept construct shows considerable variation from  $\alpha_{Gr5} = 0.646$  in Q1 to  $\alpha_{Gr5} = 0.788$  in Q5.

**Table 6:** Cronbach's alpha scores of self-perception and motivation measures by school quintiles

Measure	School SES quintile					
	1	2	3	4	5	All
<b>TIMSS Gr 9</b>						
MSC	0.709	0.714	0.725	0.773	0.869	0.771
AE	0.656	0.656	0.675	0.726	0.750	0.699
IM	0.875	0.872	0.882	0.895	0.924	0.896
EM	0.827	0.830	0.825	0.817	0.868	0.834
<b>TIMSS Gr 5</b>						
MSC	0.646	0.655	0.644	0.725	0.788	0.661
AE	0.686	0.683	0.670	0.619	0.722	0.678
IM	0.760	0.760	0.774	0.807	0.874	0.791
<b>PIRLS Gr 4</b>						
RSC	0.629	0.584	0.594	0.573	0.616	0.595
AE	0.684	0.690	0.723	0.748	0.782	0.724
IM	0.774	0.711	0.683	0.714	0.724	0.722

Source: Author's own calculations from TIMSS (2019) and PIRLS (2021)

Note: MSC = mathematics self-concept; AE = affective engagement; IM = intrinsic motivation; extrinsic motivation; RSC = reading self-concept

The PIRLS Grade 4 reading motivational and self-perception constructs demonstrate comparable patterns. Reading self-concept reliability ranges from  $\alpha_{Gr4} = 0.573$  in Q4 to  $\alpha_{Gr4} = 0.629$  in Q1, with notably lower overall reliability ( $\alpha = 0.595$ ) than the mathematics self-

concept construct, while intrinsic motivation shows moderate reliability across quintiles ( $\alpha_{Gr4} = 0.683$  to  $\alpha_{Gr4} = 0.774$ ). The consistent pattern of lower alpha coefficients in lower quintile schools across all constructs and grade levels suggests that these constructs may be less internally consistent for disadvantaged learners and younger students. This potentially reflects genuine differences in how these students conceptualize and respond to items measuring self-perceptions and motivational beliefs. This pattern underscores the importance of formally testing measurement invariance across socioeconomic groups for the different constructs at each grade levels rather than assuming construct equivalence.

## 5. TIMSS Grade 5 Analytical Results

### 5.1. Overall reflective measurement models for Grade 5 constructs

#### 5.1.1. Analysis of school belonging reflective measurement model

This section presents the empirical findings structured to establish a foundation for valid group comparisons for Grade 5 learners across different SES groups. The analysis begins by examining the single-group reflective measurement models for learners' sense of school belonging, mathematics self-concept, and intrinsic motivation. This initial step is fundamental, as it verifies that each theoretical construct is adequately operationalized by its indicators within the overall sample prior to more complex multi-group analysis. A well-specified measurement model is a prerequisite for testing measurement invariance. Subsequently, the results of the Multi-Group Confirmatory Factor Analysis (MGCFA) models are detailed to evaluate measurement invariance across groups (see Table 7 to 9).

The model for school belonging demonstrates excellent fit. The fit indices, including a CFI of 0.998 and a Tucker-Lewis Index (TLI) of 0.996, indicate a near-perfect alignment between the hypothesized model and the observed data. This level of fit suggests that the pattern of responses learners give to the school belonging items are almost exactly what we would expect if the single concept of "sense of school belonging" truly exists and causes their responses.

The strength of the relationship between each item and the latent school belonging construct is quantified by its standardized factor loadings<sup>9</sup> as shown along the arrows in Figure 3 below. The items "I am proud to go to this school" (item 5:  $\lambda = 0.65$ ) and "I feel like I belong at this school" (item 3:  $\lambda = 0.60$ ) demonstrate the strongest loadings, indicating they are strong manifestations of the construct school belonging. The items "I feel safe when I am at school" (item 2:  $\lambda = 0.55$ ) and "I like being in school" (item 1:  $\lambda = 0.59$ ) also show

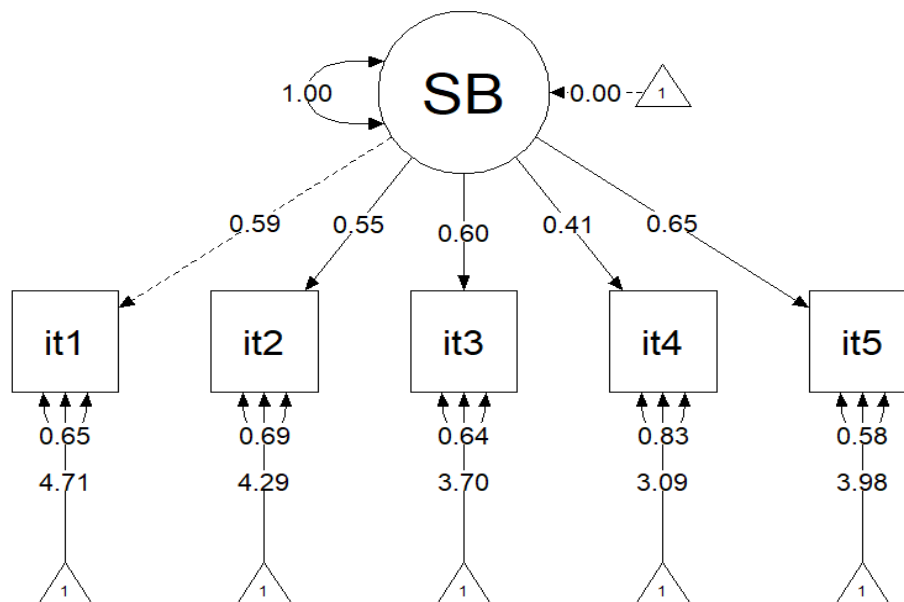
---

<sup>9</sup> Research by Tabachnick & Fidell (2007) follow that of Comrey & Lee (1992) in suggesting stringent cut-offs for factor loadings based on different sample sizes, and when the items have different frequency distributions. The authors suggested cut-offs going from 0.32 (poor), 0.45 (fair), 0.55 (good), 0.63 (very good) or 0.71 (excellent).

moderate to strong relationships. The comparatively lower loading for item related to teacher fairness, "Teachers at my school are fair to me" (item 4:  $\lambda = 0.41$ ), suggests that while perceptions of teacher engagement is related to feelings of school belonging, this specific indicator may capture a greater proportion of unique variance not attributable to the common latent factor. The high model fit indices confirm that, collectively, these items form a coherent and well-defined measure of the school belonging construct, thereby justifying its use in subsequent invariance testing.

The values below the Item boxes in the figure below are the measurement error values and the Intercept values of each of the Items measuring the latent construct, respectively. The Intercept values represent the expected score of the item when the associated latent construct, school belonging, is equal to zero. The positive item scores indicate that learners report some baseline level of each of the items associated with school belonging. Particularly, the learners rate themselves higher on item 1 ( $v = 4.71$ ) and item 2 ( $v = 4.29$ ).

**Figure 3:** Grade 5 school belonging measurement model



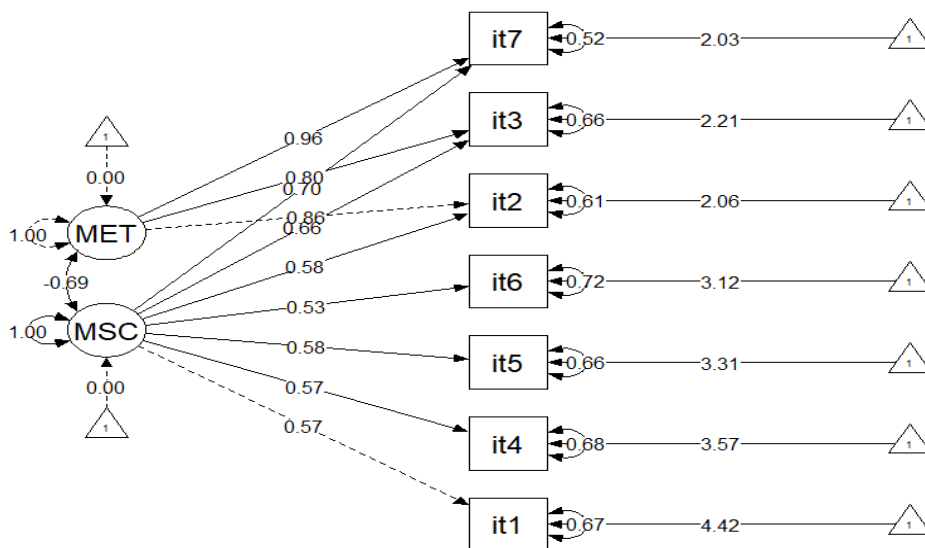
*Note:* "SB" = School belonging; "it" = item; "item 1" = I like being in school; "item 2" = I feel safe when I am at school; "item 3" = I feel like I belong at this school; "item 4" = Teachers at my school are fair to me; "item 5" = I am proud to go to this school.

### 5.1.2. Analysis of self-concept reflective measurement model

The initial measurement model for Mathematics Self-Concept (MSC) demonstrates inadequate fit, with indices falling far below or above acceptable thresholds (CFI = 0.533, TLI = 0.300, RMSEA = 0.177, SRMR = 0.116). This poor fit is a common issue in items containing both positively and negatively worded items, as respondents may process

them differently, creating a systematic "method effect" beyond the intended latent construct. This phenomenon, often attributed to acquiescence bias<sup>10</sup> or careless responding, introduces shared variance that contaminates the measurement model (Podsakoff, et al., 2003; Yang, et al., 2012). To account for this, the model was adjusted by incorporating a method factor for the negatively worded items as illustrated in Figure 4 below. This was operationally achieved by allowing the measurement errors of the negatively worded items to correlate with each other and allowing them to load onto a single, orthogonal method factor, thereby isolating the variance specific to their wording. Following this adjustment, the model fit shows a significant improvement, achieving excellent levels (CFI = 0.991, TLI = 0.983, RMSEA = 0.024, SRMR = 0.015). These results indicate that the adjusted model provides a good representation of the data. Within this well-fitting model, the indicators for the mathematics self-concept construct exhibit moderate to strong standardized factor loadings, ranging from 0.535 to 0.964 (see Figure 4). The strong negative correlation ( $r = -0.69$ ) between the mathematics self-concept factor and the method factor is expected, confirming that the method factor successfully captures the shared variance attributable to the negative wording, which is inversely related to the positive self-concept trait.

**Figure 4:** Grade 5 mathematics self-concept measurement model with method effects adjustment



*Note:* "MSC" = Mathematics self-concept; "MET" = Method factor; "it" = item; "item 1" = I usually do well in mathematics; "item 2" = Mathematics is more difficult for me than for many of my classmates; "item 3" = Mathematics is not one of my strengths; "item 4" = I learn things quickly in mathematics; "item 5" = I am good

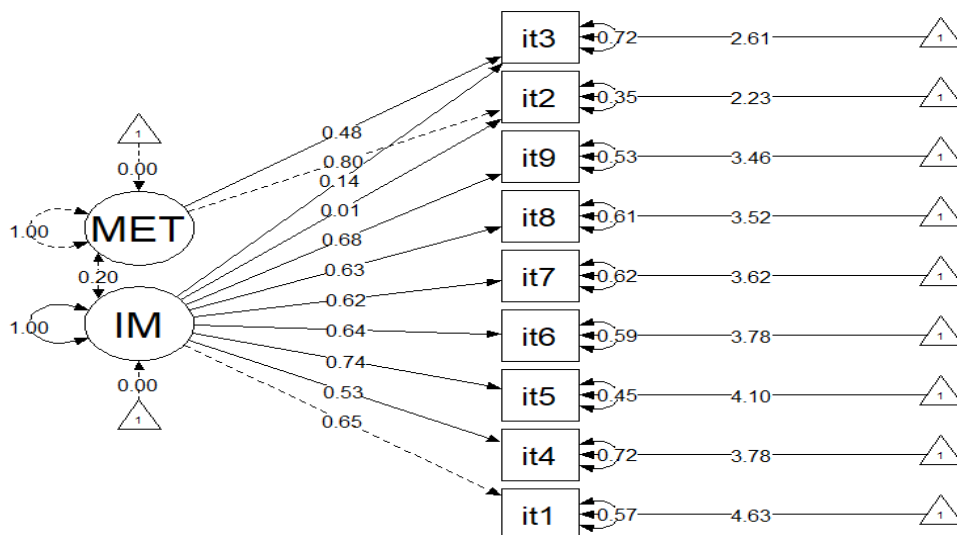
<sup>10</sup> Also known as agreement bias, this is a category of response bias common to survey research in which respondents tend to select a positive response option or indicate a positive connotation disproportionately more frequently.

at working out difficult mathematics problems; "item 6" = My teacher tells me I am good at mathematics; "item 7" = Mathematics is harder for me than any other subject.

### 5.1.3. Analysis of Intrinsic motivation reflective measurement model

The CFA for the Intrinsic motivation (IM) construct reveals that, after adjusting for method effects associated with negatively worded items, the model achieves an excellent fit (CFI = 0.990, TLI = 0.986, RMSEA = 0.030, SRMR = 0.013). Which is an indication of an accurate measurement of the construct through the nine Items as shown in Figure 5 below. The analysis provides clear insight into how each item functions as an indicator. The strongest direct manifestation of the construct is the Item, "I like mathematics" (item 5:  $\lambda = 0.748$ ), which shows a strong, positive relationship with intrinsic motivation. As expected, the contrapositive items, "I wish I did not have to study mathematics" (item 2) and "Mathematics is boring" (item 3), initially demonstrate weak loadings. After accounting for the shared method variance of their negative wording, these items align well with the latent factor. The results further highlight differences in indicator strength. The latent construct explains a substantial portion of the variance in the contrapositive item "I wish I did not have to study mathematics", meaning a student's level of intrinsic motivation is a primary driver of their response to the item. In contrast, the construct explains less variance in the statement "I learn many interesting things in mathematics", suggesting that while related, this item may also tap into external factors affecting learner's response to these items related to intrinsic motivation.

**Figure 5:** Grade 5 mathematics intrinsic motivation measurement model with method effects adjustments



Note: "IM" = Intrinsic motivation; "MET" = Method factor; "it" = item; "item 1" = I enjoy learning mathematics; "item 2" = I wish I did not have to study mathematics; "item 3" = Mathematics is boring; "item 4" = I learn many

interesting things in mathematics; "item 5" = I like mathematics; "item 6" = I like any schoolwork that involves numbers; "item 7" = I like to solve mathematics problems; "item 8" = I look forward to mathematics class; "item 9" = Mathematics is one of my favourite subjects.

Establishing well-fitting reflective measurement models for school belonging, mathematics self-concept, and intrinsic motivation provides the necessary foundation for the subsequent analysis. The preceding analysis confirms that the hypothesized factor structure is an appropriate representation of the data. The results validate the basic model structure before testing whether this measurement structure remains equivalent across groups. Consequently, the research proceeds to the core psychometric objective: testing for measurement invariance to determine if the constructs are measured equivalently for different subgroups, such as those defined by socioeconomic status.

## 5.2. Measurement invariance across school quintiles

### 5.2.1. Grade 5 school belonging

Table 7 presents the MGCFA examining measurement invariance of the mathematics sense of school belonging construct across different school quintile groups of Grade 5 South African learners. Specific comparisons are made between learners from the Poorest 60% of schools (Quintiles 1 to 3) with those from the wealthier quintiles (4th and 5th Quintiles). The analysis follows a systematic approach to testing measurement invariance by progressively constraining model parameters, beginning with configural invariance, which helps to establish that the basic factor structure is identical across groups. Measurement invariance testing then advances through to metric invariance to verify that the strength of the relationship between the items and the latent construct is the same across the groups, which is important for comparing how school belonging relates to other variables. Scalar invariance follows the metric model in the hierarchical testing, checking that the scale's intercepts are equal, allowing for valid comparisons of the latent mean scores between two groups.

The analysis of model fit indices in Table 7 shows that the comparison between learners from the Poorest 60% and Quintile 4 schools demonstrate acceptable fit. The configural invariance model serves as the foundational baseline, confirming that the basic structure of the construct is equivalent across the two groups. Specifically, the model confirms that the items in Table 2 correspond to the latent factor of school belonging. The fit indices (CFI = 0.994, RMSEA = 0.028, SRMR = 0.010) indicate that learners from both wealthier and poorer quintiles conceptualize the sense of school belonging in the same way. The metric invariance model constrains factor loadings to be equal, testing whether each item contributes to the latent construct with the same strength in both groups. The negligible deterioration in fit ( $\Delta\text{CFI} = -0.002$ ,  $\Delta\text{RMSEA} = -0.001$ ) supports metric invariance, meaning that the relationships between the items and the underlying construct are equivalent. Scalar invariance testing further constrains item intercepts to be equal. This tests whether learners from the different quintile groups with the same sense of school belonging would provide the same average score on each item. The minimal change in fit indices ( $\Delta\text{CFI} = -$

0.004,  $\Delta$ RMSEA = 0.002) supports scalar invariance. The test for strict invariance, which requires equal residual variances, revealed a significant decrease in model fit ( $\Delta$ CFI = -0.012). The model revealed that the item "I am proud to go to this school" (*item 5*) functions with different precision or reliability across the groups. To address this, a partial invariance approach was employed, whereby this parameter was freely estimated (i.e., its residual variance was allowed to differ between groups) while all other parameters remained constrained. This modified strict invariance model showed a markedly improved fit (CFI = 0.984), indicating that strict invariance holds for the model with this minor modification.

**Table 7:** MGCFA model fit for Grade 5 school belonging measurement invariance

	Model	df	Fit Indices				Model Comparisons		
			$\chi^2$	CFI	RMSEA	SRMR	$\Delta$ CFI	$\Delta$ RMSEA	$\Delta$ SRMR
Poorest 60% and Quintile 4	Configural	10	47.564	0.994	0.028	0.010			
	Metric	14	63.617	0.992	0.027	0.013	-0.002	-0.001	0.003
	Scalar	18	90.145	0.989	0.029	0.015	-0.004	0.002	0.002
	Strict	23	170.014	0.977	0.036	0.018	-0.012	0.008	0.003
	Strict (partial: item 5)	22	127.373	0.984	0.031	0.017	-0.005	0.003	0.002
Poorest 60% and Quintile 5	Configural	10	24.684	0.998	0.017	0.008			
	Metric	14	31.129	0.997	0.016	0.010	0.000	0.005	0.007
	Scalar	18	256.586	0.965	0.052	0.022	-0.033	0.036	0.012
	Scalar (partial: item 1)	17	98.168	0.988	0.031	0.015	-0.009	0.015	0.005
	Scalar (partial: item 1, 5)	16	69.341	0.992	0.026	0.014	-0.005	0.010	0.004

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For quintile 5 vs poorest 60%, model achieved partial scalar invariance and estimated the strict invariance model with items "I like being in school" (item 1) and "I am proud to go to this school" (item 5) intercepts remaining unconstrained. Model did not converge for strict invariance with sufficient parameters to constraint.

The analysis reveals a similar but more nuanced pattern when comparing learners from the Poorest 60% of schools with those from the most affluent Quintile 5 schools. As with the previous comparison, the measure demonstrates strong configural and metric invariance. This confirms that the fundamental concept of the sense of school belonging

construct is recognized similarly by both groups, and that the survey items relate to this concept with statistically equivalent strength. However, achieving scalar invariance, which is necessary for comparing average scores between groups proved more challenging. A significant deterioration in model fit ( $\Delta\text{CFI} = -0.033$ ) for the full scalar model indicates a systematic bias in how the groups use the response scale for certain items. Specifically, even when learners from the Poorest 60% and Quintile 5 schools possess the same underlying sense of school belonging, they report different average scores on the items "I like being in school" (*item 1*) and "I am proud to go to this school" (*item 5*). This suggests that factors related to school affluence influence how students interpret or respond to these specific questions. To resolve this, a model of partial scalar invariance was tested, in which the intercepts for these two items were freely estimated (i.e., allowed to differ between groups). This adjustment resulted in a well-fitting model ( $\text{CFI} = 0.992$ ,  $\Delta\text{CFI} = -0.005$ ), indicating that after accounting for the unique scaling of these two items, the core measurement of school belonging is equivalent. This partial invariance allows for valid comparisons of the latent means between the groups, provided that the differences in the two non-invariant items are acknowledged.

### 5.2.2. Grade 5 self-concept

The results in Table 8 demonstrate that the mathematics self-concept construct functions similarly for learners from the Poorest 60% of schools compared to those in Quintile 4. The configural model establishes a strong baseline fit ( $\text{CFI} = 0.992$ ,  $\text{RMSEA} = 0.026$ ,  $\text{SRMR} = 0.014$ ), confirming an equivalent factor structure. Support for metric invariance is clear from the negligible fit deterioration ( $\Delta\text{CFI} = 0.000$ ,  $\Delta\text{RMSEA} = -0.003$ ), indicating the items relate to the latent construct with equal strength across groups. Scalar invariance is also achieved ( $\Delta\text{CFI} = -0.002$ ,  $\Delta\text{RMSEA} = 0.001$ ), confirming that learners with the same level of self-concept provide similar scores on the items. The achievement of strict invariance ( $\Delta\text{CFI} = -0.006$ ,  $\Delta\text{RMSEA} = 0.003$ ) confirms full measurement equivalence, meaning any observed differences in scores can be attributed confidently to true differences in mathematics self-concept.

A more complex picture emerges, however, when comparing learners from the Poorest 60% to the most affluent Quintile 5 schools. While configural and metric invariance hold, the scalar model shows a more pronounced, though still acceptable, deterioration ( $\text{CFI} = 0.979$ ,  $\Delta\text{CFI} = -0.009$ ). The strict invariance model, however, exhibits severe misfit ( $\Delta\text{CFI} = -0.035$ ,  $\Delta\text{RMSEA} = 0.017$ ), indicating significant non-invariance. To identify the source of non-invariance, a partial invariance approach was used. The sequential release of parameters revealed that four items functioned differently between these socioeconomically distant groups. Releasing constraints for the item "I learn things quickly in mathematics" (*item 4*) alone was insufficient ( $\Delta\text{CFI} = -0.025$ ). Acceptable fit ( $\Delta\text{CFI} = -0.007$ ,  $\Delta\text{RMSEA} = 0.004$ ) was only achieved after also freeing the parameters for the items "I am just not good at mathematics" (*item 3*), "I usually do well in mathematics" (*item 1*), and "Mathematics is harder for me than any other subject" (*item 7*). This pattern indicates that perceptions of innate ability, performance, and comparative difficulty are measured with

a lot less precision between the different socioeconomic groups, requiring a model of partial strict invariance to achieve measurement equivalence of the mathematics self-concept construct among Grade 5 learners.

**Table 8:** MGCFA model fit for Grade 5 mathematics self-concept measurement invariance

	Model	Fit Indices					Model Comparisons		
		df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
Poorest 60% and Quintile 4	Configural	22	93.388	0.992	0.026	0.014			
	Metric	30	105.420	0.991	0.023	0.015	0.000	-0.003	0.001
	Scaler	35	127.503	0.989	0.023	0.016	-0.002	0.001	0.001
	Strict	42	183.686	0.983	0.026	0.018	-0.006	0.003	0.002
Poorest 60% and Quintile 5	Configural	22	83.795	0.993	0.024	0.014			
	Metric	30	139.235	0.988	0.027	0.020	-0.005	0.003	0.006
	Scaler	35	222.820	0.979	0.033	0.022	-0.009	0.006	0.002
	Strict	42	552.536	0.944	0.050	0.034	-0.035	0.017	0.011
	Strict (partial: item 4)	41	458.997	0.954	0.045	0.031	-0.025	0.012	0.009
	Strict (partial: item 4, 3)	40	392.953	0.961	0.042	0.029	-0.018	0.009	0.006
	Strict (partial: item 4, 3, 1)	39	326.485	0.968	0.039	0.026	-0.011	0.006	0.004
	Strict (partial: item 4, 3, 1, 7)	38	288.030	0.973	0.036	0.024	-0.007	0.004	0.002

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For quintile 5 vs poorest 60%, model achieved partial strict invariance. Model also achieved partial strict invariance by constraining the error terms for "I usually do well in mathematics" (item 1), "I learn things quickly in mathematics" (item 4), "Mathematics is harder for me than any other subject" (item 7), and "I am just not good at mathematics" (item 3).

### 5.2.3. Grade 5 Intrinsic motivation

Table 9 presents the analysis of measurement invariance for the mathematics intrinsic motivation construct across socioeconomic quintile groups. This analysis tests whether

the construct that is measured by items such as "I enjoy learning mathematics" (*item 1*) and "I like mathematics" (*item 5*; see Table 3) function equivalently for learners from the Poorest 60% of schools compared to their peers in the 4th and 5th socioeconomic school quintiles. The objective is to determine if socioeconomic context influences the measurement properties of intrinsic motivation in mathematics.

**Table 9:** MGCFA model fit for Grade 5 mathematics intrinsic motivation measurement invariance

	Model	Fit Indices					Model Comparisons		
		df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
Poorest 60% and Quintile 4	Configural	52	312.602	0.987	0.032	0.015			
	Metric	61	328.314	0.987	0.030	0.018	0.000	-0.002	0.002
	Scalar	68	340.646	0.986	0.029	0.018	0.000	-0.001	0.000
	Strict	77	514.009	0.978	0.034	0.021	-0.008	0.005	0.003
Poorest 60% and Quintile 5	Configural	52	369.008	0.986	0.035	0.016			
	Metric	61	551.340	0.978	0.040	0.034	-0.008	0.005	0.018
	Scalar	68	785.776	0.968	0.046	0.036	-0.010	0.006	0.003
	Scalar (partial: item 4)	67	622.716	0.975	0.041	0.034	-0.003	0.001	0.001
	Strict (partial: item 5, 1, 9, 6)	72	972.366	0.956	0.050	0.039	-0.016	0.009	0.005

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For quintile 5 vs poorest 60%, model achieved partial scalar invariance, with the item "I learn many interesting things in mathematics" (item 4) freely estimated. Model did not achieve strict invariance. By constraining the error terms for items "I like mathematics" (item 5), "I enjoy learning mathematics" (item 1), "Mathematics is one of my favourite subjects" (item 9), "I like any schoolwork that involves numbers" (item 6) along with intercepts of Item 4 freely estimated at the scalar level, partial strict invariance model does not converge with sufficient parameters constrained.

The measurement invariance analysis, for the comparisons between learners from the Poorest 60% and Quintile 4 schools, demonstrates generally acceptable equivalence at most stages of invariance testing. The configural model establishes adequate baseline fit (CFI = 0.987, RMSEA = 0.032), confirming equivalent factor structure across these two socioeconomic groups. Both metric and scalar invariance are achieved with minimal deterioration in fit indices (metric:  $\Delta CFI$  = 0.000; scalar:  $\Delta CFI$  = 0.000), indicating that factor

loadings and item intercepts function equivalently across these groups. This suggests that students from the Poorest 60% and the Q4 socioeconomic quintile group interpret or respond to mathematics intrinsic motivation items in fundamentally similar ways, meaning that these groups may probably not represent the most extreme socioeconomic contrasts. The progression to strict invariance also confirms that the measurement precision of the items between the two groups is equivalent, and therefore achieving full invariance of the intrinsic motivation construct between learners from the Poorest 60% and Quintile 4 schools.

The comparison between learners from the Poorest 60% and Quintile 5 schools provides strong evidence that larger socioeconomic gaps correspond with poorer measurement invariance. A major finding emerges at the scalar level or the third most stringent level of the MI test, which is necessary for comparing average scores. Significant model deterioration (CFI = 0.968,  $\Delta$ CFI = -0.010) was observed, primarily driven by the item "I learn many interesting things in mathematics" (*item 4*). This scalar non-invariance indicates that learners from different socioeconomic backgrounds interpret this item's scale differently; for the same level of intrinsic motivation, their scored responses appear to be systematically different. The analysis could not establish strict invariance, as constraining error variances for the majority of items caused the model to fail. This fundamental convergence issue further underscores that the measurement instrument itself performs with different consistency and precision across this pronounced socioeconomic divide, challenging the validity of direct comparisons between these groups.

## 6. TIMSS Grade 9 Analytical Results

### 6.1. Overall reflective measurement models for Grade 9 constructs

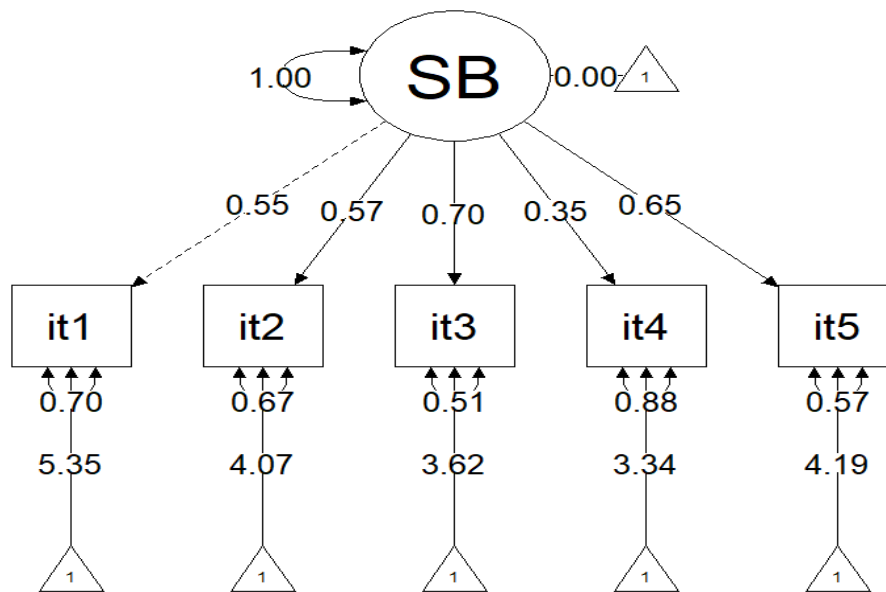
#### 6.1.1. Analysis of school belonging reflective measurement model

This section evaluates the foundational measurement structure of the four key latent constructs by examining their single-group reflective measurement models. The purpose is to verify that the hypothesized constructs— school belonging (SB), Mathematics self-concept (MSC), intrinsic motivation (IM), and extrinsic motivation (EM)—are coherently represented by the observed data before testing for group differences. A well-specified model indicates that the latent variable is a plausible common cause of its indicators, establishing a valid foundation for subsequent analysis. The results reveal distinct levels of measurement precision across the constructs.

The school belonging measurement model demonstrates one of the strongest measurement structures, with excellent fit indices (CFI = 0.996, TLI = 0.992, RMSEA = 0.025, SRMR = 0.009). As shown in Figure 6, the factor loadings range from 0.35 to 0.70, with the item "Teachers at my school are fair to me" (item 4;  $\lambda$  = 0.35) being the weakest indicator. The item with the strongest association to the underlying construct is the item "I feel like I belong at this school" (item 3;  $\lambda$  = 0.70). The other items, measuring feelings of

safety and pride, show moderate to strong relationships with the latent construct, confirming school belonging as a well-defined construct overall. The Item "I like being in school" (*item 1*) appears to be the one item that could mostly be influenced by student response bias or external factors that are not part of the "school belonging" construct being measured. This is because the item has the highest expected score as given by the Intercept ( $\nu = 5.24$ ) in the figure below.

**Figure 6:** Measurement model for school belonging



Note: "SB" = School belonging; "it" = item; "item 1" = I like being in school; "item 2" = I feel safe when I am at school; "item 3" = I feel like I belong at this school; "item 4" = Teachers at my school are fair to me; "item 5" = I am proud to go to this school.

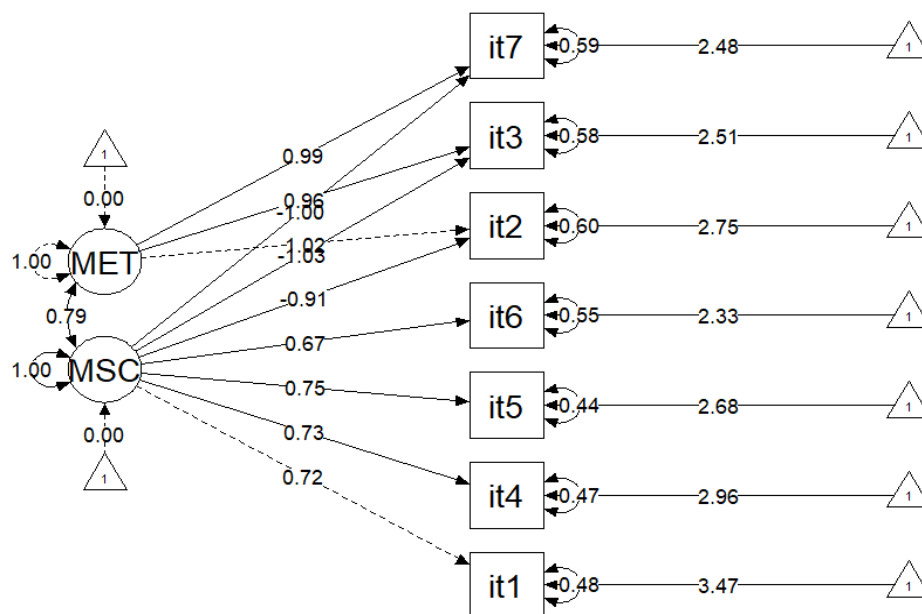
### 6.1.2. Analysis of self-concept reflective measurement model

The mathematics self-concept model exhibits a strong factorial structure after accounting for method effects associated with negatively worded items. Three items reflecting negative self-perceptions (i.e., "Mathematics is more difficult for me than for many of my classmates", "Mathematics is not one of my strengths", "Mathematics is harder for me than any other subject") were modelled to load onto both the main mathematics self-concept factor and a separate orthogonal method factor, as illustrated in Figure 7. This specification allows the model to separate substantive variance in self-concept from method variance introduced by the negative wording format. The revised measurement model demonstrates excellent fit to the data (CFI = 0.995, TLI = 0.990, RMSEA = 0.029, SRMR = 0.009). Factor loadings across most indicators affirm its robustness ( $\lambda = 0.67-1.01$ ), indicating a strong relationship between the items and the underlying latent construct of mathematics self-concept. The substantial negative correlation ( $r = -0.79$ ) between the

mathematics self-concept factor and the method factor validates the presence of method variance and confirms the successful isolation from the substantive construct.

Further examination of the item-level parameters in Figure 7 demonstrate that the intercept values for most of the positively worded items generally have higher expected scores ( $\nu = 2.33 - 3.47$ ) than negatively worded items ( $\nu = 2.08 - 2.34$ ). This result is consistent with the tendency for students to endorse positive self-statements more readily. The residual variances also vary considerably across items, suggesting differential measurement precision. Positively worded items show smaller residual variances, demonstrating more reliable measurement, as a greater proportion of their variance is explained by the latent construct.

**Figure 7:** Measurement model of MSC with method effects adjustment



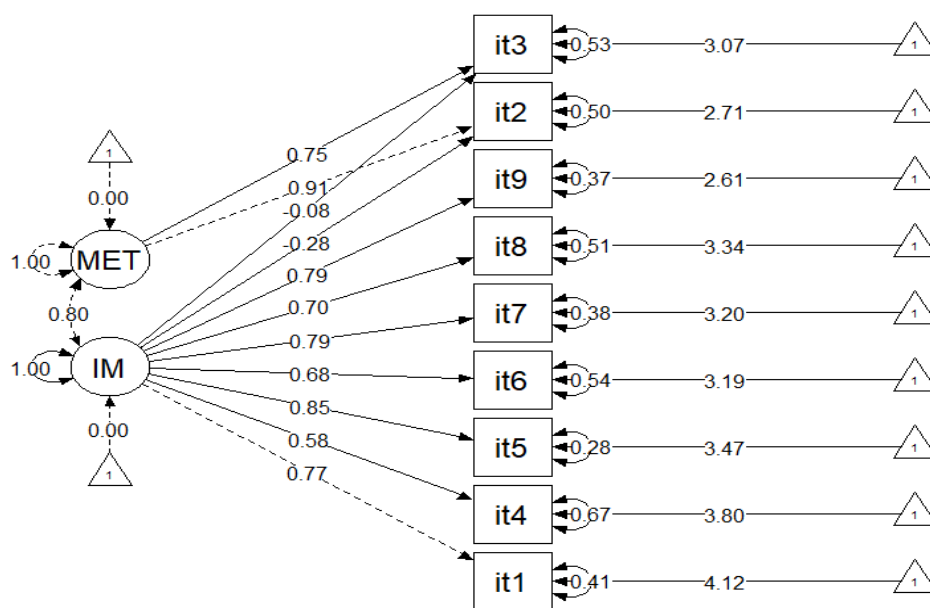
*Note:* "MSC" = Mathematics self-concept; "MET" = Method factor; "it" = item; "item 1" = I usually do well in mathematics; "item 2" = Mathematics is more difficult for me than for many of my classmates; "item 3" = Mathematics is not one of my strengths; "item 4" = I learn things quickly in mathematics; "item 5" = I am good at working out difficult mathematics problems; "item 6" = My teacher tells me I am good at mathematics; "item 7" = Mathematics is harder for me than any other subject.

### 6.1.3. Analysis of intrinsic motivation reflective measurement model

The IV model shows adequate fit to the data (CFI = 0.986, TLI = 0.980, RMSEA = 0.048, SRMR = 0.016), though a bit weaker than the other motivational constructs. As illustrated in Figure 8, factor loadings range from moderate to strong ( $\lambda = 0.58$  to 0.91), with most items loading strongly on the intrinsic motivation construct. The measurement model was adjusted for method effects by including a separate method factor for two negatively

worded Items (i.e., “I wish I did not have to study mathematics “and “Mathematics is boring”), which load on both the main intrinsic motivation factor and the method factor. Notably from the results is the substantial positive correlation ( $r = 0.79$ ) between the intrinsic motivation factor and the method factor, rather than the negative correlation that is typically expected. This result may suggest that the wording effect may actually be aligned with the construct itself, potentially indicating that negative perceptions of mathematics are an important component of low intrinsic motivation rather than merely a measurement artefact.<sup>11</sup>

**Figure 8:** Measurement model of mathematics IM with method effects adjustments



*Note:* "IM" = Intrinsic motivation; "MET" = Method factor; "it" = item; "item 1" = I enjoy learning mathematics; "item 2" = I wish I did not have to study mathematics; "item 3" = Mathematics is boring; "item 4" = I learn many interesting things In mathematics; "item 5" = I like mathematics; "item 6" = I like any schoolwork that Involves numbers; "item 7" = I like to solve mathematics problems; "item 8" = I look forward to mathematics class; "item 9" = Mathematics is one of my favourite subjects.

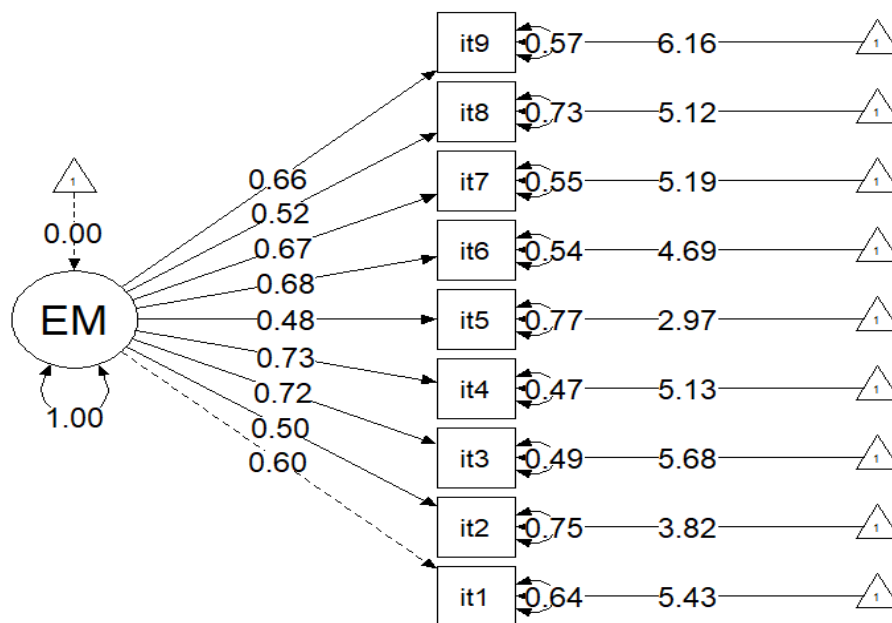
#### 6.1.4. Analysis of extrinsic motivation reflective measurement model

Finally, the Extrinsic motivation (EM) model demonstrates adequate but modest fit (CFI = 0.961, TLI = 0.948, RMSEA = 0.062, SRMR = 0.028). The range of factor loadings ( $\lambda = 0.48$ – $0.73$ ) indicates a moderate to strong link between the items and underlying construct as illustrated in Figure 9 below. This means that these self-reported measures of

<sup>11</sup> This means the data suggests that, for these group of students, agreeing with a negative statement like "I do not like mathematics" is not just a problem with the survey's wording, this may reflect true low motivation in mathematics.

mathematics extrinsic motivation indicate a sufficient manifestation of the actual underlying extrinsic motivation. The item related to specific goals<sup>12,13,14</sup> (e.g., "to get the job I want") show much stronger links as manifestations of the actual underlying construct than those concerning general extrinsic motivation. The goal-oriented Items also show higher baseline endorsement levels through high intercept values ( $\nu = 4.69 - 5.68$ ). This suggests that even students with low underlying extrinsic motivation for mathematics likely recognize the value of mathematics for achieving specific educational or labour market outcomes.

**Figure 9:** Measurement model for mathematics extrinsic motivation



*Note:* "EM" = Extrinsic motivation; "it" = item; "item 1" = I think learning mathematics will help me in my daily life; "item 2" = I need mathematics to learn other school subjects; "item 3" = I need to do well in mathematics to get into the university of my choice; "item 4" = I need to do well in mathematics to get the job I want; "item 5" = I would like a job that involves using mathematics; "item 6" = It is important to learn about mathematics to get ahead in the world; "item 7" = Learning mathematics will give me more job opportunities when I am an adult; "item 8" = My parents think that it is important that I do well in mathematics; "item 9" = It is important to do well in mathematics.

<sup>12</sup> The item related to the specific goal of getting a job the learner wants has the strongest factor loading of  $\lambda = 0.73$ .

<sup>13</sup> The item related to the specific goal of getting into the desired university has the second highest loading  $\lambda = 0.72$ .

<sup>14</sup> The item related to the specific goal of getting ahead in the world has a factor loading of  $\lambda = 0.68$ .

## 6.2. Measurement invariance across school quintiles

### 6.2.1. Grade 9 school belonging

MGCFA was used to assess the measurement invariance of the school belonging construct for Grade 9 learners across school socioeconomic quintile groups. As shown in Table 10, both comparisons of learners from the Poorest 60% versus Quintile 4 and Poorest 60% versus Quintile 5 schools, show a strong support for configural invariance ( $CFI \geq 0.990$ ,  $RMSEA \leq 0.08$ ,  $SRMR \leq 0.08$ ). Establishing configural invariance confirms that the factor structure of the school belonging construct is equivalent across these groups, meaning that the model shows that the same set of observed items are manifestations of the same underlying theoretical construct. Furthermore, metric invariance was also established for both group comparisons ( $\Delta CFI \leq -0.010$ ,  $\Delta RMSEA \leq 0.030$ ,  $\Delta SRMR \leq 0.030$ ), indicating that the strength of the relationship between the individual items and the underlying latent construct is statistically equivalent across the socioeconomic groups.

**Table 10:** MGCFA model fit for Grade 9 school belonging measurement invariance

	Model	Fit Indices					Model Comparisons		
		df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
Poorest 60% and Quintile 4	Configural	10	84.014	0.994	0.030	0.011			
	Metric	14	94.457	0.994	0.027	0.012	-0.001	-0.004	0.002
	Scalar	18	202.928	0.985	0.035	0.019	-0.008	0.009	0.006
	Strict	22	329.027	0.976	0.041	0.030	-0.015	0.012	0.014
	Strict (partial: item 1)	21	176.898	0.988	0.030	0.020	-0.003	0.001	0.004
Poorest 60% and Quintile 5	Configural	10	90.361	0.994	0.032	0.011			
	Metric	14	108.254	0.992	0.029	0.014	-0.001	-0.003	0.004
	Scalar	18	1126.747	0.912	0.088	0.044	-0.081	0.059	0.030
	Scalar (partial: item 1)	17	132.074	0.991	0.029	0.016	-0.002	0.000	0.001
	Strict	22	950.977	0.926	0.073	0.055	-0.065	0.044	0.039
	Strict (partial: Item 1, 4)	20	161.806	0.989	0.030	0.017	-0.002	0.001	0.002

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are estimated with previously freely estimated/unconstrained items. For quintile 4 vs poorest 60%, model achieved partial strict invariance with the error term unconstrained for the item "I like being in school" (item 1). For the quintile 5 vs poorest 60% comparison, the model achieved partial scalar invariance with the intercept for item "I like being in school"

(item 1) freely estimated. In addition to the error term of item "item 1" freed, the error term for the item "Teachers at my school are fair to me" (item 4) was also unconstrained to achieve partial strict invariance.

The results for scalar invariance revealed a clear divergence between the socioeconomic comparisons. For learners from the Poorest 60% versus Quintile 4 schools, scalar invariance was supported. Although the model fit showed a discernible decrease from the metric level ( $\Delta\text{CFI} = -0.008$ ), the final indices nonetheless remained within acceptable thresholds ( $\text{CFI} = 0.985$ ,  $\text{RMSEA} = 0.035$ ). This indicates that the measurement scale possesses an equivalent baseline; students with the same underlying sense of school belonging are expected to provide the same score on the survey items, validating subsequent comparisons of their average engagement levels. In contrast to the Quintile 4 comparison, the comparison between the Poorest 60% and the most affluent Quintile 5 schools initially failed to achieve scalar invariance, with a substantial deterioration in model fit ( $\Delta\text{CFI} = -0.081$ ). This significant misfit indicated that learners from these different backgrounds were not using the response scale for all items in the same way. To address this, a partial scalar invariance model was tested by freeing the intercept for the item "I like being in school" (*item 1*). This modification resulted in a well-fitting model ( $\Delta\text{CFI} = -0.002$ ), establishing partial scalar invariance. This finding confirms that the core construct is measured equivalently only using four out of the five items measuring school belonging. This equivalence allows for latent mean comparisons between these stark socioeconomic groups, accounting for the item that is answered differently between the two groups.

The tests for strict invariance revealed further differentiation between the socioeconomic groups. For the comparison between the Poorest 60% and Quintile 4, the initial strict model showed a notable decline in fit ( $\Delta\text{CFI} = -0.015$ ). However, after releasing constraints on the residual variance of the item "I like being in school" (*item 1*), model fit improved significantly ( $\Delta\text{CFI} = -0.003$ ,  $\Delta\text{RMSEA} = 0.001$ ,  $\Delta\text{SRMR} = 0.004$ ), achieving partial strict invariance. A more pronounced pattern emerged for the comparison between the Poorest 60% and Quintile 5. The initial strict model demonstrated substantial misfit ( $\Delta\text{CFI} = -0.065$ ,  $\Delta\text{RMSEA} = 0.044$ ,  $\Delta\text{SRMR} = 0.039$ ), indicating differences in measurement precision. To establish an acceptable model, it was necessary to free the residual variances for two items: "I like being in school" (*item 1*) and "Teachers at my school are fair to me" (*item 4*). This modification resulted in a well-fitting partial strict invariance model ( $\Delta\text{CFI} = -0.002$ ,  $\Delta\text{RMSEA} = 0.001$ ,  $\Delta\text{SRMR} = 0.002$ ). These results confirm a clear gradient in measurement equivalence. While configural and metric invariance hold fully, to attain scalar and strict invariance, it requires progressively more extensive partial invariance adjustments, particularly for the comparison spanning the widest socioeconomic divide (Poorest 60% vs. Quintile 5). This pattern indicates that the most significant measurement differences lie in the baseline scores (intercepts) and the precision (error variances) of specific items that are freed, reflecting how socioeconomic context shapes the response processes for these indicators.

### 6.2.2. Grade 9 mathematics self-concept

Measurement invariance of the mathematics self-concept construct was examined to determine if it functions equivalently across socioeconomic groups. When comparing learners from the Poorest 60% and Quintile 4 schools, the construct demonstrates full measurement invariance. Excellent model fit at the configural level (CFI = 0.995, RMSEA = 0.028, SRMR = 0.009) confirms both groups recognize the same conceptual framework of self-concept. Critically, the minimal fit deterioration at metric ( $\Delta CFI = 0.000$ ), scalar ( $\Delta CFI = -0.004$ ), and strict ( $\Delta CFI = -0.003$ ) levels establishes complete measurement equivalence. This robust pattern validates direct comparisons of both relationships with other variables and latent mean scores between these groups.

**Table 11:** MGCFA model fit for Grade 9 mathematics self-concept measurement invariance

	Model	Fit Indices					Model Comparisons		
		df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
Poorest 60% and Quintile 4	Configural	22	159.757	0.995	0.028	0.009			
	Metric	30	181.553	0.995	0.025	0.012	0.000	-0.003	0.003
	Scalar	35	287.926	0.991	0.030	0.015	-0.004	0.005	0.003
	Strict	42	386.020	0.988	0.032	0.019	-0.003	0.002	0.004
Poorest 60% and Quintile 5	Configural	22	216.807	0.994	0.033	0.010			
	Metric	30	470.706	0.986	0.043	0.035	-0.009	0.012	0.030
	Metric (partial: item 6)	29	412.542	0.988	0.041	0.032	-0.006	0.007	0.022
	Scalar	34	629.267	0.981	0.047	0.034	-0.007	0.006	0.003
	Strict	41	1037.297	0.968	0.055	0.044	-0.013	0.008	0.009
	Strict (partial: item 6, 4)	39	881.317	0.973	0.052	0.040	-0.008	0.005	0.005

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For quintile 5 vs poorest 60%, model achieved partial metric invariance and estimated the scalar invariance model with item "My teacher tells me I am good at mathematics" (item 6) loadings remaining unconstrained. Model achieved full scalar invariance. Partial strict invariance was achieved with the error term for items "My teacher tells me I am good at mathematics" (item 6) and "I learn things quickly in mathematics" (item 4) freely estimated across groups, with loading for "item 6" still unconstrained.

In contrast, the comparison between the Poorest 60% and the most affluent Quintile 5 schools revealed notable measurement non-invariance. While the basic factor structure was shown to be equivalent (configural CFI = 0.994, RMSEA = 0.033, SRMR = 0.010), metric invariance was compromised, as indicated by a significant change in the SRMR fit index ( $\Delta$ SRMR = 0.030). This signifies that the relationship strength between certain items and the underlying self-concept construct differs between these groups. Partial metric invariance was only achieved after allowing the parameter for the item "My teacher tells me I am good at mathematics" (*item 6*) to vary. This was necessary again at the strict invariance level, where the error variances for this item and an additional item "I learn things quickly in mathematics" (*item 4*) had to be freed again to establish partial strict invariance. This pattern is a potential indicator that external validation (teacher feedback) and perceived learning speed are interpreted or experienced inconsistently across this pronounced socioeconomic divide (poorest 60% vs Q5), challenging the validity of using the constructs in group comparisons.

### 6.2.3. Grade 9 mathematics intrinsic motivation

The measurement invariance analysis for the mathematics intrinsic motivation shows distinctly different outcomes across socioeconomic comparisons. For learners from the Poorest 60% versus Quintile 4 schools, the construct demonstrates full, sequential invariance. The model maintained strong fit from configural invariance (CFI = 0.986, RMSEA = 0.047, SRMR = 0.017) through to strict invariance ( $\Delta$ CFI = 0.000, RMSEA = 0.042, SRMR = 0.032), with negligible changes at each sequential level of testing. This robust pattern confirms that the constructs factor structure, factor loadings, item intercepts, and measurement precision are entirely equivalent between these groups, thereby validating all subsequent comparisons of latent means and structural relationships.

The comparison between the Poorest 60% and Quintile 5 schools revealed significant measurement non-invariance, indicating the items may function differently across the two groups of students. Although configural invariance was established (CFI = 0.985, RMSEA = 0.049, SRMR = 0.017), metric invariance ( $\Delta$ SRMR = 0.055) was not established, due to the variation in the item "Mathematics is boring" (*item 3*), which appears to relate to the latent construct with different strength in each group according to the model test. Achieving partial scalar invariance ( $\Delta$ CFI = -0.008,  $\Delta$ RMSEA = 0.008,  $\Delta$ SRMR = 0.006) further required freeing intercepts for items measuring enjoyment of mathematics class and work with numbers, indicating that students with the same level of intrinsic motivation provided systematically different baseline responses to these items. Most critically, the strict invariance model failed to converge, suggesting fundamental differences in the reliability of the items. This convergence failure signifies that the measurement instrument itself may lack equivalent precision across these groups, presenting a substantial challenge for valid quantitative comparisons.

**Table 12:** MGCFA model fit for Grade 9 mathematics intrinsic motivation measurement invariance

		Fit Indices					Model Comparisons		
	Model	df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
Poorest 60% and Quintile 4	Configural	52	978.831	0.986	0.047	0.017			
	Metric	61	1045.033	0.985	0.044	0.032	-0.001	-0.003	0.009
	Scalar	68	1135.125	0.983	0.044	0.033	-0.001	-0.001	0.001
	Strict	77	1163.470	0.983	0.042	0.032	0.000	-0.002	0.000
Poorest 60% and Quintile 5	Configural	52	1044.102	0.985	0.049	0.017			
	Metric	61	1367.679	0.981	0.052	0.072	-0.005	0.003	0.055
	Metric (partial: item 3)	61	1267.59	0.983	0.049	0.025	-0.002	-0.001	0.008
	Scalar partial: (item 6, 8)	68	1860.97	0.975	0.056	0.031	-0.008	0.008	0.006

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For quintile 5 vs poorest 60%, model achieved partial metric invariance and estimated the scalar invariance model with item "Mathematics is boring" (item 3) loadings remaining unconstrained. Model also achieved partial scalar invariance with the intercept for the items "I like any schoolwork that involves numbers" (item 6) and "I look forward to mathematics class" (item 8) unconstrained. Item "item 3" loadings remain unconstrained across groups when estimating the partial scalar invariance model.

#### 6.2.4. Grade 9 mathematics extrinsic motivation

The analysis of the extrinsic motivation construct across the Poorest 60% and Quintile 4 schools showed good fit across all invariance levels. Configural invariance (CFI = 0.957, RMSEA = 0.064, SRMR = 0.028) indicated a stable factor structure. Metric invariance ( $\Delta CFI$  = 0.000,  $\Delta RMSEA$  = -0.004,  $\Delta SRMR$  = 0.001) and scalar invariance ( $\Delta CFI$  = -0.003,  $\Delta RMSEA$  = -0.002,  $\Delta SRMR$  = 0.002) confirmed the equivalence of item relationships and latent means, respectively. At strict invariance, some decline in fit ( $\Delta CFI$  = -0.011) was observed, suggesting differences in measurement precision. Adjustments for the specific item, "I need to do well in mathematics to get into the university of my choice" (item 3) improved fit, revealing the differences in the measurement error for learner's perception of mathematics being a tool for future academic opportunities.

For the Poorest 60% vs Quintile 5 comparison, fit indices were less consistent. While configural invariance showed acceptable fit (CFI = 0.947, RMSEA = 0.073, SRMR = 0.031), constraining factor loadings led to declines in the comparative fit index ( $\Delta$ CFI = -0.010), suggesting some variability in response patterns between the groups. Factor loadings on the item related to the learners rating on how important their parents think mathematics is for them were freely estimated to achieve partial metric invariance. Scalar invariance showed more substantial deterioration ( $\Delta$ CFI = -0.026), indicating differences in some item mean scores. The two items showing the most variation relates to the learners' career aspirations of getting a job which would require mathematics (*item 5*), and also the learner's perception of their parent's expectation for their performance in mathematics (*item 8*). Partial scalar invariance was achieved after freeing the intercepts of those two items. Strict invariance again for this model between the Poorest 60% and Quintile 5 groups did not converge, with most items showing high variability in measurement error.

**Table 13:** MGCFA model fit for Grade 9 mathematics extrinsic motivation measurement invariance

	Model	Fit Indices					Model Comparisons		
		df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta$ CFI	$\Delta$ RMSEA	$\Delta$ SRMR
Poorest 60% and Quintile 4	Configural	54	1863.243	0.957	0.064	0.028			
	Metric	62	1888.302	0.957	0.060	0.030	0.000	-0.004	0.001
	Scalar	70	2025.986	0.954	0.058	0.031	-0.003	-0.002	0.002
	Strict	79	2505.664	0.943	0.061	0.037	-0.011	0.003	0.005
	Strict (partial: item 3)	78	2309.409	0.947	0.059	0.034	-0.007	0.001	0.003
Poorest 60% and Quintile 5	Configural	54	2433.068	0.947	0.073	0.031			
	Metric	62	2878.411	0.938	0.074	0.047	-0.010	0.001	0.017
	Metric (partial: item 8)	61	2623.43	0.943	0.071	0.037	-0.004	-0.002	0.007
	Scalar	69	3817.69	0.917	0.081	0.048	-0.026	0.010	0.010
	Scalar (partial: item 8)	68	3341.18	0.928	0.076	0.043	-0.016	0.005	0.006
	Scalar (partial: Item 8, 5)	67	3053.91	0.934	0.073	0.041	-0.009	0.002	0.003

Source: Own calculations from TIMSS (2019)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely

estimated/unconstrained items. For quintile 5 vs poorest 60%, model achieved partial metric invariance and estimated the scalar invariance model with item "My parents think that it is important that I do well in mathematics" (item 8) loadings remaining unconstrained. Model also achieved partial scalar invariance with the addition to the intercept for the item "Item 8," the item "I would like a job that involves using mathematics" (item 5) was also unconstrained to achieve partial scalar invariance. Item "item 8" loadings remain unconstrained across groups when estimating the partial scalar invariance model.

## 7. PIRLS Grade 4 Analytical Results

### 7.1. Measurement Invariance across SES groups and test language speakers

#### 7.1.1. Grade 4 school belonging

This analysis in Table 14 examines whether the school belonging construct for Grade 4 learners' functions equivalently across learners from the Poorest 60% and Quintile 5 schools, with a specific focus on whether they speak the language of the test at home, which is also the language of instruction. For learners who speak the language of the test at home, the items demonstrate strong cross-socioeconomic validity, achieving full configural, metric, and scalar invariance. This indicates that the construct is conceptualized similarly, the items relate to a sense of school belonging with equal strength, and the rating scale is used with the same baseline across affluent and poor schools. Consequently, both relationships with other variables and latent mean scores can be validly compared between these socioeconomic groups. The only deviation was at the strict invariance level ( $\Delta CFI > -0.010$ ), where the item "I have friends at this school" (*item 6*; see Table 2) exhibited differing measurement precision; however, partial strict invariance was achieved by freeing its error variance, preserving the overall robustness of the measurement model.

The findings reveal a more complex picture when language status differs between groups. For non-language speakers, the items demonstrate full measurement invariance across socioeconomic groups, including strict invariance. However, a critical divergence occurs in the cross-group comparison between affluent language speakers and poor non-language speakers, where metric invariance was not established, specifically, fewer than two factor loadings were equivalent across these two cross-groups. The failure to achieve even partial metric invariance ( $\Delta CFI > -0.010$ ) indicates that the items do not relate to the underlying engagement construct with the same strength for these groups, rendering their scores fundamentally incomparable. Conversely, the items demonstrate full measurement invariance in the reverse comparison, between affluent non-language speakers and poor language speakers. This pattern may suggest that the disruptive effect of language disparity is not symmetrical; it primarily undermines measurement when the language disadvantage is concentrated within the lower socioeconomic group, thereby highlighting a specific intersection of socioeconomic and linguistic marginalization.

**Table 14:** MGCFA model fit for Grade 4 school belonging measurement invariance

	Model	Fit Indices					Model Comparisons		
		<i>df</i>	$\chi^2$	<i>CFI</i>	<i>RMSEA</i>	<i>SRMR</i>	$\Delta$ <i>CFI</i>	$\Delta$ <i>RMSEA</i>	$\Delta$ <i>SRMR</i>
<b>Poorest 60% and Quintile 5</b>	<b><i>Within language speakers</i></b>								
	Scalar	28	178.778	0.975	0.038	0.026	-0.007	0.002	0.002
	Strict	34	375.171	0.942	0.052	0.037	-0.032	0.014	0.011
	Strict (partial: item 6)	31	340.280	0.948	0.052	0.036	-0.027	0.014	0.011
	<b><i>Within non-language speakers</i></b>								
	Scalar	28	46.131	0.989	0.026	0.023	-0.005	0.006	0.002
	Strict	34	59.392	0.985	0.028	0.026	-0.004	0.002	0.003
	<b><i>Poorest 60%: Non-language speakers<sup>15</sup></i></b>								
	Configural	18	54.113	0.990	0.035	0.016			
	Metric	23	119.819	0.973	0.050	0.041	-0.017	-0.019	0.025
	Metric (partial: item 6, 5, 3, 2)	24	105.497	0.977	0.045	0.030	-0.013	0.010	0.014
	<b><i>Quintile 5: Non-language speakers<sup>16</sup></i></b>								
	Scalar	28	75.797	0.988	0.024	0.016	0.000	-0.002	0.000
	Strict	34	83.236	0.988	0.022	0.017	0.000	-0.002	0.001

Source: Own calculations from PIRLS (2021)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For the Poorest 60%: Non-language speakers, the model only establishes full configural invariance and could not establish any invariance level beyond metric invariance due to less than two items having equal loadings across the cross-group comparisons with Quintile 5: language speakers.

<sup>15</sup> This panel of the table compares learners from the Poorest 60% schools who do not speak the language of the test at home to learners from Quintile 5 schools who speak the language of the test at home.

<sup>16</sup> This panel compares the reverse comparison of the panel above—learners in Quintile 5 schools who do not speak the language of the test at home to learners from the Poorest 60% of schools who speak the language of the test at home.

### 7.1.2. Grade 4 reading self-concept

Table 15 presents the analysis of the literacy self-concept construct, examining its measurement invariance across different combinations of school affluence and home language. For learners who share the same language status, the items measuring reading self-concept demonstrate strong validity. The model for learners who speak the test language at home achieved scalar invariance, confirming that learners from the Poorest 60% and Quintile 5 schools conceptualize RSC identically. The only deviation was for learners who do not speak the test language at home at the strict invariance level, where the initial model showed a significant deterioration in fit ( $\Delta CFI > -0.010$ ). The item "Reading is easy for me" (*item 2*; see Table 1) showed differential precision, and allowing its error variance to differ established partial strict invariance (final  $\Delta CFI \leq -0.010$ ).

An important difference emerges when language status and school affluence are crossed. The model comparing learners from the Poorest 60% of schools who do not speak the test language at home with learners from Quintile 5 schools who do speak it failed to achieve metric invariance ( $\Delta RMSEA > 0.015$ ), even after freeing factor loadings for four of the six items measuring literacy confidence. Conversely, the reverse comparison achieved full measurement invariance, with the scalar model showing no deterioration ( $\Delta CFI = 0.000$ ,  $\Delta RMSEA = -0.003$ ) and the strict model demonstrating excellent fit ( $\Delta CFI = -0.002$ ,  $\Delta RMSEA = 0.000$ ). This demonstrates that the construct functions identically for these cohorts of students, indicating that the disruptive effect of language disparity is again most likely acute when it coincides with economic disadvantage.

**Table 15:** MGCFA model fit for Grade 4 reading self-concept measurement invariance

		Fit Indices					Model Comparisons		
	Model	df	$\chi^2$	CFI	RMSEA	SRMR	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
<b>Poorest 60% and Quintile 5</b>	<b><i>Within language speakers</i></b>								
	Scalar	22	67.921	0.991	0.024	0.015	-0.001	-0.001	0.001
	Strict	28	98.089	0.986	0.026	0.019	-0.005	0.002	0.004
	<b><i>Within non-language speakers</i></b>								
	Scalar	22	42.392	0.986	0.031	0.026	-0.002	-0.001	0.002
	Strict	28	80.709	0.963	0.045	0.032	-0.023	0.013	0.006
	Strict (partial: item 2)	27	59.972	0.977	0.036	0.021	-0.009	0.005	-0.005

Model	df	$\chi^2$	Fit Indices			Model Comparisons		
			CFI	RMSEA	SRMR	$\Delta$ CFI	$\Delta$ RMSEA	$\Delta$ SRMR
<b>Poorest 60%: Non-language speakers</b>								
Configural								
Metric	18	52.138	0.986	0.034	0.027	-0.014	0.034	0.020
Metric (partial: item 6, 5, 4, 1)	15	27.438	0.995	0.022	0.019	-0.005	0.022	0.012
<b>Quintile 5: Non-language speakers</b>								
Scalar	22	48.051	0.993	0.020	0.014	0.000	-0.003	0.000
Strict	28	62.087	0.991	0.020	0.015	-0.002	0.000	0.001

Source: Own calculations from PIRLS (2021)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For the Poorest 60%: Non-language speakers, the model only achieved configural invariance and could establish any invariance level beyond metric invariance due to less than two items having equal loadings across the cross-group comparisons with Quintile 5: language speakers.

### 7.1.3. Grade 4 reading intrinsic motivation

The analysis of the intrinsic reading motivation construct first examines learners who share the same language status. For learners who speak the test language at home, comparisons between those from the Poorest 60% and Quintile 5 schools demonstrate full measurement invariance through to strict invariance. This confirms the items function identically across the socioeconomic divide for this group. Similarly, for learners who do not speak the test language at home, comparisons between the Poorest 60% and Quintile 5 schools also show strong equivalence, achieving scalar invariance. However, to achieve an acceptable model at the strict invariance level for this group, the error variance for the item "I would like to have more time for reading" (item 4; see Table 3) had to be freely estimated, resulting in a partial strict invariance model.

The analysis reveals a critical and asymmetric pattern when language status and socioeconomic status are crossed. The comparison between learners from the Poorest 60% of schools who do not speak the test language at home and learners from Quintile 5 schools who do speak the language achieved only partial invariance. Partial metric invariance ( $\Delta$ CFI < -0.010,  $\Delta$ RMSEA < 0.015,  $\Delta$ SRMR < 0.030) was established after freeing the factor loadings for the enjoyment items "I think reading is boring" and "I enjoy reading."

Scalar invariance initially failed ( $\Delta CFI = -0.018$ ) and was only achieved partially (final  $\Delta CFI = -0.009$ ) after freeing the intercepts for the item related to the learners perceived benefit of reading "I learn a lot from reading" (i.e., *item 4*) and the socially orientated item "I like talking about what I read with other people" (i.e., *item 1*). Finally, partial strict invariance ( $\Delta CFI < -0.010$ ) required also freeing the error variance for the latter social item and the item "I would be happy if someone gave me a book as a present" (i.e., *item 2*). The reverse comparison testing of the learners from Quintile 5 schools who do not speak the test language versus learners from the Poorest 60% who do speak the test language achieved full measurement invariance through to strict invariance. This wide difference in measurement invariance results may be indicative that the measurement instrument may at times fail to function equivalently when language disadvantage is concentrated within the less affluent group, but works well when the same language difference exists alongside socioeconomic advantage.

**Table 16:** MGCFA model fit for Grade 4 reading intrinsic motivation measurement invariance

		Fit Indices					Model Comparisons		
Model		<i>df</i>	$\chi^2$	<i>CFI</i>	<i>RMSEA</i>	<i>SRMR</i>	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
<b>Poorest 60% and Quintile 5</b>	<b><i>Within language speakers</i></b>								
	Scalar	53	414.316	0.966	0.043	0.031	-0.004	0.002	0.002
	Strict	61	472.474	0.961	0.043	0.033	-0.005	0.000	0.003
	<b><i>Within non-language speakers</i></b>								
	Scalar	53	123.454	0.977	0.038	0.031	-0.007	0.004	0.002
	Strict	61	169.065	0.965	0.043	0.036	-0.012	0.006	0.005
	Strict (partial: item 4)	60	151.480	0.971	0.040	0.036	-0.007	0.003	0.005
	<b><i>Poorest 60%: Non-language speakers</i></b>								
	Configural	40	173.378	0.975	0.045	0.024			
	Metric	47	269.372	0.958	0.053	0.048	-0.017	0.009	0.024
Metric (partial: item 3, 5)	45	209.126	0.969	0.047	0.034	-0.006	0.002	0.010	
Scalar	51	308.481	0.951	0.055	0.041	-0.018	0.008	0.006	

Model	Fit Indices					Model Comparisons		
	<i>df</i>	$\chi^2$	<i>CFI</i>	<i>RMSEA</i>	<i>SRMR</i>	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$
Scalar (partial: item 1, 6)	49	262.398	0.960	0.051	0.038	-0.009	0.004	0.003
Strict	57	357.781	0.943	0.056	0.049	-0.016	0.005	0.011
Strict (partial: item 1, 2)	55	317.972	0.950	0.053	0.042	-0.009	0.002	0.005
<b>Quintile 5: Non-language speakers</b>								
Configural	40	211.019	0.980	0.038	0.021			
Metric	47	215.416	0.980	0.035	0.022	0.000	-0.003	0.001
Scalar	53	228.957	0.979	0.034	0.022	-0.001	-0.001	0.000
Strict	61	270.475	0.975	0.034	0.023	-0.004	0.001	0.001

Source: Own calculations from PIRLS (2021)

Note: Items in parenthesis are freely estimated at each level of invariance, to achieve partial invariance. Subsequent to achieving partial invariance, the models are then estimated with previously freely estimated/unconstrained items. For the Poorest 60%: Non-language speakers, the model only achieved configural invariance and could establish any invariance level beyond metric invariance due to less than two items having equal loadings across the cross-group comparisons with Quintile 5: language speakers.

## 8. Discussion and conclusion

### 8.1. Summary of main results

The aim of this paper is to validate the measurement of learners' self-perception and motivational belief constructs across socio-economic groups in South African TIMSS and PIRLS data. The analysis of internal consistency, using Cronbach's alpha, reveals consistent patterns across the different socioeconomic status (SES) groups. Firstly, for every construct (e.g., self-concept, intrinsic motivation), reliability is consistently higher in Grade 9 than in Grade 5 or 4. For instance, the intrinsic motivation alpha score is estimated at 0.897 for Grade 9 learners, compared to an estimate of 0.791 for Grade 5, and 0.722 for Grade 4 learners. Secondly, a strong socioeconomic effect is evident; reliability is almost always lowest amongst learners in Quintile 1 schools (lowest SES) and highest in the Quintile 5 subset (highest SES). For instance, learners in Grade 5 present a mathematics self-concept reliability score of 0.646 in Q1 compared to 0.788 in Q5. In Grade 9 the mathematics self-concept reliability score has an estimate of 0.709 in Q1 compared to an estimate of 0.869 in Q5. Overall, the motivational constructs, intrinsic- and extrinsic motivation, are the most reliable constructs with reliability scores generally above the 0.7

threshold (Hair, et al., 2006), while self-concept (MSC/RSC) and school belonging show much more variability, especially among younger learners, and those from low-SES schools.

These patterns can be explained through developmental and contextual lenses. One potential explanation for the higher reliability in older grades is that students' self-perceptions and motivations become more stable and well-defined over time (Kuzucu, et al., 2013). The strong socioeconomic effect observed in the data, where reliability estimates for constructs such as mathematics self-concept and intrinsic motivation consistently increase from the lowest (Quintile 1) to the highest (Quintile 5) SES groups, suggests that the measurement properties of these constructs differ depending on the contextual environment the constructs are measured. This effect may be influenced by school environmental factors that are associated with higher SES, such as better resources or instructional quality, which could provide a setting where students develop clearer and more differentiated self-perceptions and motivations (Mndawe, et al., 2024; Chen, et al., 2025). However, there is a danger in using aggregate reliability scores, which is that the scores mask substantial measurement differences for specific subgroups. For instance, the overall reliability coefficient for the Grade 5 mathematics self-concept construct is 0.661, which is misleading, because the measure performs notably better in higher-SES schools (e.g., Quintile 4,  $\alpha = 0.725$ ) and notably worse in mid-to-lower SES schools (e.g., Quintile 1,  $\alpha = 0.646$ ). These overall scores cannot be directly used to identify "low self-concept," because if the low scores in low-to-mid SES schools are partly due to poor measurement reliability, an attempt to intervene in improving learner's motivation and expectancy beliefs could potentially be misguided. The reliability scores which are below the 0.7 threshold likely reflect measurement artefacts more than it shows that the learners have a low score or a true lack of motivation or expectancy of success. Beyond these Cronbach's alpha estimates, the multi-group confirmatory factor analysis (MGCFA) further substantiates varied measurement functioning across these subgroups for the different constructs. In other words, measurement invariance further shows that we cannot trust a low score truly means a student lacks confidence or motivation, but rather that the survey tool itself may be faulty in certain school contexts and grades.

## 8.2. Limitations

Measurement invariance testing fundamentally depends on correct specification of the underlying factor structure, where misspecification of the baseline model would hold true through all subsequent invariance tests. Since measurement invariance testing proceeds hierarchically, errors at earlier levels compromise the validity of conclusions at later levels. Additionally, decisions about whether invariance has been achieved rely on fit index cut-offs (such as  $\Delta CFI < .01$  or  $\Delta RMSEA < .015$ ) that represent conventions rather than absolute thresholds. When full invariance cannot be achieved, partial invariance models require decisions about which items to free, typically based on identifying those showing the largest changes in fit indices. However, different researchers might reasonably make different choices about which parameters to free, potentially leading to different conclusions about the degree and nature of non-invariance. A fundamental limitation is

that measurement invariance testing identifies where differences exist between groups without explaining why they occur. This is because measurement invariance testing cannot definitively determine whether detected non-invariance reflects genuine psychological differences in how groups experience and conceptualize the constructs or if its measurement artefacts arising from differential item functioning across contexts, or some combination of both. The theoretical interpretations offered in this study, while grounded in existing research remain somewhat speculative, and alternative explanations may be equally plausible.

The PIRLS dataset presents unique challenges because it was translated into all 11 official South African languages. Subtle differences in item meaning across languages could create non-invariance that reflects translation artefacts rather than genuine differences in how language groups experience reading engagement or self-concept. The cross-sectional nature of this study further limits conclusions about how motivational and belief frameworks develop over time. While comparing Grade 5 and Grade 9 cohorts from TIMSS 2019 provides some evidence of stability in non-invariance patterns, these are independent samples rather than longitudinal data following the same learners. Observed differences between Grades could reflect developmental changes in how learners conceptualize motivational constructs, or cohort effects from different educational experiences, or both.

### 8.3. Conclusion

The measurement invariance results have demonstrated that socioeconomic background does not only predict differential educational outcomes among students as research has previously shown (Taylor & Yu, 2009; Shepherd & Van der Berg, 2020; Gutfleisch & Kogan, 2022), but also shapes how students understand and respond to survey items measuring their self-perceptions and motivation. The growing global recognition of learners' self-concept and motivation makes the measurement invariance findings in this study particularly significant. The non-invariance observed across socioeconomic groups provide evidence that students from different backgrounds are navigating distinct educational contexts that shape how they perceive their capabilities, value academic subjects, and experience school. One other critical finding that emerges from the results is the language-mediated effects that compound existing disadvantages for specific groups of learners, particularly those from poorer socioeconomic contexts who also do not speak the language of instruction at home.

From an educational equity perspective, these measurement invariance results suggest important limitations in how we currently understand student experiences across different school environments and demographic groups. The assessment tools used to measure learners' self-perceptions and motivation appear to function differently depending on students' socioeconomic contexts, raising questions about the validity of direct comparisons across these groups. The results reveal that disadvantaged learners may systematically interpret and respond to questions about their academic experiences differently than their advantaged peers. This finding has important implications for how we understand educational inequality. When we assume that all students interpret survey

items about self-concept, motivation, and belonging in the same way, we risk misunderstanding the actual experiences and needs of disadvantaged students. This strongly suggests that simply collecting more student voice data without accounting for measurement non-equivalence may misrepresent learners' true self-perceptions and motivation. When disadvantaged Grade 9 learners, for example, report lower self-perceptions and motivational beliefs than their more advantaged peers, the results show that there is systematic difference in how they respond to and interpret items about "doing well," "finding mathematics useful for careers," or "being good at difficult problems." These interpretations are likely shaped by the learners' constrained educational experiences, different peer reference groups, and limited exposure to certain career pathways. Policy interventions designed to "boost academic motivation" based on these survey responses may therefore be misdirected if they fail to recognize that measurement itself is capturing different realities across socioeconomic contexts.

## References

- Abubakar, A. et al., 2015. Assessing Sense of School Belonging Across Cultural Contexts Using the PSSM: Measurement and Functional Invariance: Measurement and Functional Invariance. *Journal of Psychoeducational Assessment*, 34(4), pp. 380-388.
- Abulela, M. A., Nickodem, K. & Rodriguez, M. C., 2024. *Measurement Invariance of Social and Emotional Learning Measures across Four Administrations: Conventional Fit Statistics Versus RMSEA*, Philadelphia, PA: National Council on Measurement in Education Annual Meeting.
- Altonji, J. G. & Mansfield, R. K., 2011. The Role of Family, School, and Community Characteristics in Inequality in Education and Labor-Market Outcomes. In: G. J. Duncan & R. J. Murnane, eds. *Whither Opportunity?: Rising Inequality, Schools, and Children's Life Chances*. New York: Russell Sage Foundation, pp. 339-358.
- Archer, L., DeWitt, J. & Willis, B., 2014. Adolescent boys' science aspirations: Masculinity, capital, and power. *Journal of Research in Science Teaching*, 51(1), pp. 1-30.
- Asparouhov, T. & Muthén, B., 2014. Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), pp. 495-508.
- Atilgan, M. & Deniz, K. Z., 2023. Investigation of the measurement invariance of affective characteristics related to TIMSS 2019 mathematics achievement by gender. *Journal of Measurement and Evaluation in Education and Psychology*, 14(3), pp. 185-199.
- Bandura, A., 1997. *Self-efficacy: The exercise of control*. New York: W.H. Freeman and Company.
- Behrmann, T., 2018. *Evaluating the Effects of Mother Tongue on Math and Science Instruction*. Monument, CO, USA: ISTES Organization.
- Bodovski, K. & Farkas, G., 2008. "Concerted cultivation" and unequal achievement in elementary school. *Social Science Research*, 37(3), pp. 903-919.
- Bollen, K. & Lennox, R., 1991. Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), pp. 305-314.
- Bourdieu, P., 1986. The Forms of Capital. In: J. G. Richardson, ed. *Handbook of Theory and Research for the Sociology of Education*. New York: Greenwood Press, pp. 241-258.
- Bourdieu, P. & Passeron, J. -C., 1977. *Reproduction in Education, Society and Culture*. New Delhi: Sage Publications.
- Bowles, S. & Gintis, H., 2002. Schooling in capitalist America revisited. *Sociology of Education*, 75(1), pp. 1-18.
- Branson, N. et al., 2024. The socioeconomic dimensions of racial inequality in South Africa: A social space perspective. *The British Journal of Sociology*, 75(4), pp. 613-635.
- Bryne, B. M., Shavelson, R. J. & Muthen, B., 1989. Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin*, 105(3), pp. 456-466.
- Bryne, B. M. & van De Vijver, F. J. R., 2010. Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10(2), pp. 107-132.
- Chen, F. F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), pp. 464-504.

- Chen, M., Liu, S., Wijaya, T. T. & Cao, Y., 2025. Influence of family socioeconomic status on academic buoyancy and adaptability: Mediating effect of parental involvement. *Acta Psychologica*, Volume 253, p. 104753.
- Cheung, G. & Rensvold, R. B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), pp. 233-255.
- Claro, S., Paunesku, D. & Dweck, C. S., 2016. Growth mindset tempers the effects of poverty on academic achievement. *PNAS*, 113(31), pp. 8664-8668.
- Comrey, A. L. & Lee, H. B., 1992. *A first course in factor analysis*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Croizet, J. -C. & Millet, M., 2012. Social class and test performance: From stereotype threat to symbolic violence and vice versa. In: M. Inzlicht & T. Schmader, eds. *Stereotype threat: Theory, process, and application*. Oxford: Oxford University Press, pp. 188-201.
- Cronbach, L. J. & Meehl, P. E., 1995. Construct validity in psychological tests. *Psychological Bulletin*, 52(4), pp. 281-302.
- Cunha, F. & Heckman, J. J., 2007. The technology of skill formation. *American Economic Review*, 97(2), pp. 31-47.
- Demir, M. C. & Gelbal, S., 2023. An Examination of TIMSS 2015 Science Affective Factors with Regard to Gender and Regions. *Bartın University Journal of Faculty of Education*, 12(3), pp. 579-593.
- Department of Basic Education (DBE), 2023. *PIRLS 2021: South African Main report*, Pretoria: Department of Basic Education.
- Désert, M., Préaux, M. & Jund, R., 2009. So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. *European Journal of Psychology of Education*, Volume 24, pp. 207-218.
- Destin, M. et al., 2019. Do Student Mindsets Differ by Socioeconomic Status and Explain Disparities in Academic Achievement in the United States?. *AERA Open*, 5(3).
- Ding, Y., Hansen, K. Y. & Klapp, A., 2023. Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychology of Education*, Volume 38, pp. 709-732.
- Ding, Y., Yang Hansen, K. & Klapp, A., 2023. Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychological Education*, Volume 38, pp. 709-732.
- Doyle, L., Easterbrook, M. J. & Harris, P. R., 2023. Roles of socioeconomic status, ethnicity and teacher beliefs in academic grading. *The British journal of educational psychology*, 93(1), pp. 91-112.
- Eccles, J. et al., 1983. Expectancies, values and academic behaviors. In: J. T. Spence, ed. *Achievement and Achievement Motives*. San Francisco, Calif: W. H. Freeman, pp. 75-146.
- Eccles, J. S. & Wigfield, A., 2020. From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, Volume 61, p. 101859.
- Eriksson, K., Helenius, O. & Ryve, A., 2019. Using TIMSS items to evaluate the effectiveness of different instructional practices. *Instructional Science*, Volume 47, pp. 1-18.

- Eriksson, K., Lindvall, J., Helenius, O. & Ryve, A., 2021. Socioeconomic status as a multidimensional predictor of student achievement in 77 societies. *Frontiers in Education*, 6(731634).
- Eser, D. C., 2021. Investigation of Measurement Invariance According to Home Resources: TIMSS 2015 Mathematical Affective Characteristics Questionnaire. *International Journal of Assessment Tools in Education*, 8(3), pp. 633-648.
- Fernald, A., Marchman, V. A. & Weisleder, A., 2013. SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, 16(2), pp. 234-248.
- Fleisch, B., 2008. *Primary Education in Crisis: Why South African Schoolchildren Underachieve in reading and mathematics*. Cape Town: Juta & Co..
- Gerber, C., Mans-Kemp, N. & Schlechter, A. F., 2019. Investigating the moderating effect of student engagement on academic performance. *Acta Academica*, 45(4), pp. 256-274.
- Girard, C. et al., 2021. The relation between home numeracy practices and a variety of math skills in elementary school children. *PLOS ONE*, 16(9), p. e0255400.
- Goodenow, C., 1993. The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools*, 30(1), pp. 79-90.
- Goodenow, C. & Grady, K., 1993. The relationship of school belonging and friends' values to academic motivation among urban adolescent students. *The Journal of Experimental Education*, 62(1), pp. 60-71.
- Goudeau, S. & Croizet, J. -C., 2017. Hidden advantages and disadvantages of social class: How classroom settings reproduce social inequality by staging unfair comparison. *Psychological Science*, 28(2), pp. 162-170.
- Gutfleisch, T. & Kogan, I., 2022. Parental occupation and students' STEM achievements by gender and ethnic origin: Evidence from Germany. *Research in Social Stratification and Mobility*, Volume 82.
- Haberman, M., 2010. The Pedagogy of Poverty versus Good Teaching. *Phi Delta Kappan*, 92(2), pp. 81-87.
- Hair, J. F., Black, W., Babin, B. J. & Anderson, R. E., 2006. *Multivariate Data Analysis. Technometrics*, 31(3).
- Harrison, L. A., Stevens, C. M., Monty, A. N. & Coakley, C. A., 2006. The consequences of stereotype threat on the academic performance of White and non-White lower income college students. *Social Psychology of Education*, Volume 9, pp. 341-357.
- Hart, B. & Risley, T. R., 1995. *Meaningful differences in the everyday experience of young American children*. Baltimore, Maryland: Paul H Brookes Publishing.
- Harzing, A. -W., 2006. Response styles in cross-national survey research. *International Journal of Cross Cultural Management*, 6(2), pp. 243-266.
- Heckman, J. J., 2006. Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, 312(5782), pp. 1900-1902.
- He, J., Barrera-Pedemonte, F. & Buchholz, J., 2018. Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, 26(4), pp. 369-385.
- Hofer, S. I. et al., 2024. Self-perceptions as mechanisms of achievement inequality: evidence across 70 countries. *Science of Learning*, 9(2).

- Hoff, E., 2003. The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, 74(5), pp. 1368-1378.
- Hofmeyr, H., 2022. Why do girls do better? Unpacking South Africa's gender gap in PIRLS and TIMSS. *International Journal of Educational Development*, Volume 94.
- Horn, J. L. & Mcardle, J. J., 1992. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3), pp. 117-144.
- Isdale, K., Reddy, V., Juan, A. & Arends, F., 2017. *TIMSS 2015 Grade 5 National Report: Understanding mathematics achievement amongst Grade 5 learners in South Africa*, Pretoria: HSRC.
- Jussim, L. & Harber, K. D., 2005. Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. *Personality and Social Psychology Review*, 9(2), pp. 131-155.
- Kline, R. B., 2016. *Principles and practice of structural equation modeling*. 4th Edition ed. New York: Guildford Press.
- Kuang, X., Mok, M. M. C., Chiu, M. M. & Zhu, J., 2019. Sense of school belonging: Psychometric properties and differences across gender, grades, and East Asian societies. *PsyCh Journal*, 8(4), pp. 449-464.
- Kuzucu, Y. et al., 2013. Developmental Change and Time-Specific Variation in Global and Specific Aspects of Self-Concept in Adolescence and Association with Depressive Symptoms. *Journal of Early Adolescence*, 34(5), pp. 638-666.
- Lafontaine, D., Dupont, V., Jaegers, D. & Schillings, P., 2019. Self-concept in reading: Factor structure, cross-cultural invariance and relationships with reading achievement in an international context (PIRLS 2011). *Studies in Educational Evaluation*, Volume 60, pp. 78-89.
- Law, D. M. et al., 2022. Measurement Invariance and Relationships Among School Connectedness, Cyberbullying, and Cybervictimization: A Comparison Among Canadian, Chinese, and Tanzanian Adolescents. *Journal of Psychoeducational Assessment*, 40(7), pp. 865-879.
- Leitgeb, H. et al., 2023. Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*, 110(102805).
- Liou, P. & Lin, J. J., 2021. Comparisons of Science Motivational Beliefs of Adolescents in Taiwan, Australia, and the United States: Assessing the Measurement Invariance Across Countries and Genders. *Frontiers in Psychology*, Volume 12.
- Lombardi, E. et al., 2019. The Impact of School Climate on Well-Being Experience and School Engagement: A Study With High-School Students. *Frontiers in Psychology*, Volume 10:2482.
- Mag-aso, S. J., Albiso, K. M., Condeza, K. B. & Edaño, R. V., 2025. The Influence of School Climate and Subjective Well-Being on Students' Engagement. *International Journal of Research and Innovation in Social Science*, 9(3), pp. 1844-1855.
- Major, B. & Schmader, T., 1998. Coping with stigma through psychological disengagement. In: J. K. Swim & C. Stangor, eds. *Prejudice: The target's perspective*. Santa Barbara: Academic Press, pp. 219-241.
- Marsh, H. W., 1986. Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, 23(1), pp. 129-149.

- Marsh, H. W. et al., 2013. Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: a comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, Volume 105, p. 108.
- Marsh, H. W. et al., 2014. The big-fish-little-pond effect in mathematics: a cross-cultural comparison of US and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology*, 45(5), pp. 777-804.
- Marsh, H. W. et al., 2015. The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology*, Volume 107, pp. 258-271.
- Marsh, H. W., Bryne, B. M. & Shavelson, R. J., 1988. A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, 80(3), pp. 366-380.
- Marsh, H. W., Cheng, J. & Martin, A. J., 2008. How we judge ourselves from different perspectives: Contextual influences on self-concept formation. In: M. Maehr, T. Urdan & S. Karabenick, eds. *Advances in Motivation and Achievement*. New York: Elsevier, pp. 315-356.
- Marsh, H. W. & Parker, J. W., 1984. Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well?. *Journal of Personality and Social Psychology*, 47(1), pp. 213-231.
- Marsh, H. W. et al., 2019. The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, 111(2), p. 311.
- Marsh, H. W., Scalas, L. F. & Nagengast, B., 2010. Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, 22(2), pp. 366-381.
- Marsh, H. W. & Shavelson, R., 1985. Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20(3), pp. 107-123.
- Masitsa, G., 2008. Discipline and disciplinary measures in the Free State township schools: unresolved problems. *Acta Academica: Critical Views on Society, Culture and Politics*, 40(3), pp. 234-270.
- Michael, D. & Kyriakides, L., 2023. Mediating effects of motivation and socioeconomic status on reading achievement: a secondary analysis of PISA 2018. *Large-scale Assessments in Education*, 11(31).
- Mndawe, D. M., Oduaran, C. A. & Onyencho, V. C., 2024. Students' Experiences Regarding Diversity and Their Academic Self-concept in a South African Institution. *International Journal of Educational Sciences*, 44(2), pp. 94-102.
- Möller, J. & Marsh, H. W., 2013. Dimensional comparison theory. *Psychological Review*, 120(3), pp. 544-560.
- Mullis, I. V. & Martin, M. O., 2013. *TIMSS 2015 Assessment Frameworks*, Boston: TIMSS & PIRLS International Study Center.
- Mullis, I. V., Martin, M. O. & Loveless, T., 2016. *20 years of TIMSS. International trends in mathematics and science achievement, curriculum, and instruction*, Boston: TIMSS and PIRLS International Study Center.
- Nagengast, B. & Marsh, H. W., 2014. Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In: M. von Davier & D. Rutkowski, eds. *Handbook*

- of international large-scale assessment: Background, technical issues, and methods of data analysis.* Boca Raton: CRC Press, pp. 317-344.
- OECD, 2013. *PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs*, Paris: PISA, OECD Publishing.
- OECD, 2017. *PISA 2015 Results (Volume III): Students' Well-Being*, Paris: PISA, OECD Publishing.
- Oginni, O. I. & Olabode, O. T., 2020. Effect of Mother Tongue and Mathematical Language on Primary School Pupils Performance in Mathematics. *British Journal of Educational Studies*, 4(3), pp. 542-546.
- Osterman, K., 2000. Students' Need for Belonging in the School Community. *Review of Educational Research*, 70(3), pp. 323-367.
- Part, R., Perera, H. N., Marchand, G. C. & Bernacki, M. L., 2020. Revisiting the dimensionality of subjective task value: towards clarification of competing perspectives. *Contemporary Educational Psychology*, Volume 62: 101875.
- Pettersson, M., 2023. *Within and Beyond Borders: Intrinsic Reading Motivation in PIRLS*. Oslo, Norway, Abstract from Frontier Research in Educational Measurement 2023.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. -Y. & Podsakoff, N. P., 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), pp. 879-903.
- Porter, S. C., Jackson, C. K., Kiguel, S. & Easton, J. Q., 2023. *Investing in adolescents: High school climate and organizational context shape student development and educational attainment*, Chicago, IL: University of Chicago Consortium on School Research.
- Putnick, D. L. & Bornstein, M. H., 2016. Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review*, Volume 41, pp. 71-90.
- R Core Team, 2020. A language and environment for statistical computing. *R*.
- Raju, N. S., Laffitte, L. J. & Byrne, B. M., 2002. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), pp. 517-529.
- Raufelder, D. & Kulakow, S., 2021. The role of the learning environment in adolescents' motivational development. *Motivation and Emotion*, Volume 45, pp. 299-311.
- Reardon, S. F., 2011. The widening of the socioeconomic status achievement gap: New evidence and possible explanation. In: G. J. Duncan & R. J. Murnane, eds. *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*. New York: Russell Sage Foundation, pp. 91-116.
- Reardon, S. F., 2018. The Widening Academic Achievement Gap Between the Rich and the Poor. In: D. Grusky & J. Hill, eds. *Inequality in the 21st Century*. New York: Routledge, pp. 177-189.
- Reddy, V. et al., 2016. *TIMSS 2015: highlights of mathematics and science achievement of grade 9 South African learners*, Pretoria: HSRC (Commissioned by the Department of Basic Education, December).
- Reddy, V. et al., 2020. *TIMSS 2019: Highlights of South African Grade 9 results in mathematics and science. Achievement and achievement gaps*, Pretoria: Department of Basic Education.

- Regner, I., Huguet, P. & Monteil, J. -M., 2002. Effects of Socioeconomic Status (SES) Information on Cognitive Ability Inferences: When Low-SES Students Make Use of a Self-Threatening Stereotype. *Social Psychology of Education*, 5(3), pp. 253-269.
- Reise, S. P., Widaman, K. F. & Pugh, R. H., 1993. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), pp. 552-566.
- Rogelberg, S. L., Starrett, A., Irvin, M. J. & DiStefano, C., 2021. Examining motivation profiles within and across socioeconomic levels on educational outcomes. *International Journal of Educational Research*, Volume 109.
- Rosseel, Y., 2012. An R package for structural equation modeling. *Journal of Statistical Software*, 48(1), pp. 1-36.
- Roux, K., van Staden, S. & Pretorius, E. J., 2022. Investigating the differential item functioning of a PIRLS literacy 2016 text across three languages. *Journal of Education*, Volume 87, pp. 135-155.
- Rutkowski, L. & Svetina, D., 2014. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74(1), pp. 31-57.
- Ryan, R. M. & Deci, E. L., 2009. Promoting self-determined school engagement: Motivation, learning, and well-being. In: K. R. Wentzel & A. Wigfield, eds. *Educational psychology handbook series. Handbook of motivation at school*. New York: NY: Routledge, pp. 171-195.
- Ryan, R. M. & Deci, E. L., 2016. Facilitating and hindering motivation, learning and well-being in schools: Research and observations from self-determination theory. In: K. R. Wentzel & D. B. Miele, eds. *Handbook of motivation at school*. New York: Routledge, pp. 96-119.
- Sarstedt, M. & Wilczynski, P., 2009. More for less? A comparison of single-item and multi-item measures. *Die Betriebswirtschaft*, 69(2), p. 211.
- Schmitt, N. & Kuljanin, G., 2008. Measurement invariance: Review of practice and implication. *Human Resources Management Review*, 18(4), pp. 210-222.
- Shen, C., 2002. Revisiting the relationship between students' achievement and their self-perceptions: a cross-national analysis based on TIMSS 1999 data. *Assessment in Education: Principles, Policy & Practice*, 9(2), pp. 161-184.
- Shen, C. & Talavera, O., 2002. The Effects of self-perception on students' Mathematics and Science Achievement in 38 countries Based on TIMSS 1999 Data. *North American Stata Users' Group Meetings 2003*, Volume 8.
- Shen, C. & Tam, H. P., 2008. The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation*, 14(1), pp. 87-100.
- Shepherd, D. L., 2015. Learn to teach, teach to learn: A within-pupil acrosssubject approach to estimating the impact of teacher subject knowledge on South African grade 6 performance. *Stellenbosch Economic Working Paper*, Volume WP01/2015, Stellenbosch University.
- Shepherd, D. L., 2017. Gender, Self-Concept and Mathematics and Science Performance of South African Grade 9 Students. *Stellenbosch Economic Working Paper*, Volume WP11/2017, Stellenbosch University.

- Shepherd, D. L. & van der Berg, S., 2020. Analysing Matric Data to Identify "promising" Schools in Mathematics Performance. *Stellenbosch Economic Working Paper*, Volume WP16/2020, Stellenbosch University.
- Shuukwanyama, T. T., Long, C., Nkosi, A. D. & Maseko, J., 2022. The language of instruction in mathematics teacher education for the early grades. *South African Journal of Childhood Education*, 12(1).
- Simpkins, S., Fredricks, J. A. & Eccles, J. S., 2012. Charting the Eccles' Expectancy-Value Model from Mothers' Beliefs in Childhood to Youths' Activities in Adolescence. *Developmental Psychology*, 48(4), pp. 1019-1032.
- Sirin, S. R., 2005. Socioeconomic status and academic achievement: a meta-analytic review of research. *Review in Education Research*, Volume 75, pp. 417-453.
- Sirin, S. & Rogers-Sirin, L., 2004. Exploring school engagement of middle-class African American adolescents. *Youth & Society*, 35(3), pp. 323-340.
- Spaull, N., 2013. Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, 33(5), pp. 436-447.
- Spaull, N., 2013. *South Africa's Education Crisis: The quality of education in South Africa 1994-2011*, Johannesburg: Centre For Development & Enterprise.
- Spaull, N., 2015. Schooling in South Africa: How Low Quality Education Becomes a Poverty Trap. In: A. de Lannoy, S. Swartz, L. Lake & C. Smith, eds. *The South African Child Gauge*. Cape Town: Children's Institute, pp. 34-41.
- Spaull, N. & Taylor, S., 2012. Effective enrolment - creating a composite measure of educational access and educational quality to accurately describe education system performance in sub-Saharan Africa. *Stellenbosch Economic Working Papers*, Volume 21/12, pp. 1-25.
- Steele, C. M., 1997. A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52(6), pp. 613-629.
- Steenkamp, J.-B. E. & Baumgartner, H., 1998. Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1), pp. 78-90.
- Steinberg, L., Graham, S., O'Brien, L. & Woolard, J., 2009. Age Differences in Future Orientation and Delay Discounting. *Child Development*, 80(1), pp. 28-44.
- Tabachnick, B. G. & Fidell, L. S., 2007. *Using multivariate statistics*. 5th ed. New York: Allyn and Bacon.
- Takalani, M. G. & Shepherd, D. L., 2023. *Who and How Matters: Using situated expectancy value theory to explore the mathematics performance of Grade 9 learners in South Africa*, Stellenbosch: Research in Socio-Economic Policy.
- Taylor, N., Muller, P. & Vinjevold, P., 2003. *Getting Schools Working: Research and Systemic School Reform in South Africa*. Cape Town: Pearson Education South Africa.
- Taylor, S. & Yu, D., 2009. The importance of socio-economic status in determining educational achievement in South Africa. *Stellenbosch Economic Working Paper*, Volume WP01/09, Stellenbosch University.
- van Broekhuizen, H., 2016. Graduate unemployment and Higher Education Institutions in South Africa. *Stellenbosch University Working Paper, Department of Economics*, Volume 08/2016.

- Vandenberg, R. J. & Lance, C. E., 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), pp. 4-69.
- Venkat , H. & Spaul, N., 2015. What do we know about primary teachers' mathematical content knowledge in South Africa? An analysis of SACMEQ 2007. *International Journal of Educational Development*, Volume 41, pp. 121-130.
- von Maurice, J., Dörfler, T. & Artelt, C., 2014. The Relation between Interests and Grades : Path Analyses in Primary School Age. *International Journal of Educational Research* , 64(1), pp. 1-11.
- Wang, M. -T. & Eccles, J. S., 2013. School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction*, Volume 28, pp. 12-23.
- Wang, M. -T. & Holcombe, R., 2010. Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal*, 47(3), pp. 633-662.
- Wang, Z., 2015. Examining big-fish-little-pond-effects across 49 countries: a multilevel latent variable modelling approach. *Educational Psychology*, 35(2), pp. 228-251.
- Wang, Z. & Bergin, D. A., 2017. Perceived relative standing and the big-fish-little-pond effect in 59 countries and regions: Analysis of TIMSS 2011 data. *Learning and Individual Differences*, Volume 57, pp. 141-156.
- Widaman, K. F. & Reise, S. P., 1997. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In: K. J. Bryant, M. Windle & S. G. West, eds. *The science of prevention: Methodological advances from alcohol and substance abuse research*. Washington, D.C.: American Psychological Association, pp. 281-324.
- Wigfield, A. & Eccles, J. S., 2000. Expectancy-Value Theory of Achievement Motivation. *Contemporary Educational Psychology* , 25(1), pp 68-81.
- Wigfield, A. & Cambria, J., 2010. Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, 30(1), pp. 1-35.
- Wurster, S., 2022. Measurement invariance of non-cognitive measures in TIMSS across countries and across time. An application and comparison of Multigroup Confirmatory Factor Analysis, Bayesian approximate measurement invariance and alignment optimization approach. *Studies in Educational Evaluation*, 73(4), p. 101143.
- Yang, Y., Chen, Y. H., Lo, W. J. & Turner, J. E., 2012. Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment*, Volume 30, pp. 509-519.
- Zoch, A., 2017. The effect of neighbourhoods and school quality on education and labour market outcomes in South Africa. *Stellenbosch Economic Working Papers* , Volume WPO8/2017.
- Zumbo, B. D., 2007. Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), pp. 223-233.