Cracking the Code: A Framework for Ensuring Reliable TIMSS Test Scores for South African Learners

Musa Adekunle Ayanwale, University of Johannesburg, South Africa
Daniel Olutola Oyeniran, University of Alabama, United States
Joseph Taiwo Akinboboye, Federal University of Lafia, Nigeria

The European Conference on Education 2023
Official Conference Proceedings

Abstract
In many countries, including South Africa, the Trends in International Mathematics and Science Study (TIMSS) serves as a prominent assessment tool for evaluating students' achievements in mathematics and science. It's noteworthy, however, that the extent of the reliability of TIMSS test scores in South Africa has not been extensively investigated within the existing literature. This research employs generalizability theory to assess the reliability of 2019 TIMSS test scores among South African students. The primary objective is to gauge various forms of errors linked to test scores, encompassing factors such as tester and item effects. To achieve this, a single facet crossed design was adopted alongside a systematic sampling approach to gather item responses from 150 fourth-grade learners in response to 35 mathematics items drawn from an IEA IDB Analyzer Merge module. For data analysis, the *gtheory* package within the R language and statistical computing environment was employed. The assessment encompassed the computation of the generalizability (g) coefficient, the phi (Φ) coefficient, and the decision (d) study. The results divulged a g-coefficient of 0.989 and a Φ-coefficient of 0.981, indicating a notable level of reliability. These findings emphasize that TIMSS test scores remain unaffected by diverse sources of error, including those stemming from tester and item effects. This robust level of generalizability and reliability in the scores is thus validated. In the context of South Africa, these outcomes can potentially furnish policymakers and educators with more comprehensive insights for making informed decisions concerning the utilization and interpretation of TIMSS test scores.

Keywords: Generalizability Theory, Reliability, TIMSS, Test Scores, Mathematics Achievement, South Africa

1. Introduction

The Trends in International Mathematics and Science Study (TIMSS) program, conducted by the International Association for the Evaluation of Educational Achievement (IEA), is a significant global assessment initiative that evaluates students' performance in mathematics and science across various countries. It has had a profound impact on education policies and practices worldwide. Participating countries like South Africa learn from each other's experiences to improve their education systems (Fishbein et al., 2021). South Africa selects diverse schools and students to ensure a representative sample reflecting socioeconomic, geographic, and urban-rural diversity. The TIMSS assessment targets two grade levels: Grade 4 and Grade 8. Grade 4 covers fundamental mathematics and science concepts, while Grade 8 delves into more advanced topics. Mathematics assessment topics span algebra, geometry, number theory, and data analysis, while the science assessment encompasses biology, chemistry, physics, earth science, and environmental science. To gauge students' comprehension, TIMSS focuses on three cognitive domains. It gauges students' understanding through "Knowing," "Applying," and "Reasoning" cognitive domains, assessing recall, problem-solving, and deeper comprehension (Martin et al.,2020). The program generates achievement benchmark scores (elementary level (400-474), average level (475-549, high level (550-624), and advanced level (625 or more points)) categorizing countries based on performance levels (Martin et al.,2020). However, these plausible scores are likely to be affected by measurement errors, which are variations under consistent conditions. This error arises due to differences between observed scores (X) and true scores (T). The psychometric analysis aims to estimate and reduce error variance for accurate assessments (Ayanwale, 2019; Crocker & Algina, 2008). Classical test theory (CTT) treats scores as a combination of true and random error components but can't identify specific error sources (Ayanwale et al. 2018:2019a; Brennan, 2001; Johnson & Johnson, 2009). Recognizing CTT's limitations, generalizability theory (g-theory) was developed to disentangle and estimate various sources of measurement error (Brennan, 2001; Shavelson & Webb, 1991). This theory goes beyond CTT, offering a broader framework to pinpoint, differentiate, and estimate errors for improved reliability in assessments. Unreliable scores can lead to unfair educational policies and inaccurate evaluations of students, teachers, and schools. This underscores the importance of assessing the reliability of TIMSS scores specifically for South African learners. Employing generalizability theory allows us to uncover various factors contributing to score inconsistency, like learner differences, rater variations, and task-related variability. This insight becomes instrumental in refining the testing procedures, guaranteeing precise and dependable outcomes. Ultimately, these efforts will have a positive impact on individual learners and contribute to enhancing the entire education system.

Generalizability Theory (G-theory) is a statistical framework employed to assess the reliability of test scores (de Vet et al. (2011); Thompson (2003)) like those from the TIMSS assessment. It accounts for various factors that can influence scores, such as different raters, learners, and tasks. G-theory helps identify and quantify these factors to understand score reliability better. In this context, G-theory examines multiple facets like learners, raters, and tasks (Brennan, 2001). When the same test is given to different learner groups, score variation can occur due to their abilities. High learner variability can hinder accurate ability measurement, while high rater variability makes consistent evaluation challenging. Similarly, different raters grading the same test might have varying scoring criteria. Additionally, using different tasks to measure the same skill can lead to score variation due to task difficulty, not suited for all learners. To ensure reliable TIMSS scores, each facet's impact on reliability

must be considered. G-theory helps identify which facets affect reliability the most, aiding in adjustments. This enhances score accuracy, ensuring valid measures of learner abilities. The conditions of G-theory encompass variables impacting TIMSS score reliability, including learner, rater, and task numbers. Limited learners might yield non-representative results for a larger learner population. A small rater group could introduce significant bias into score reliability. Likewise, a few tasks might not comprehensively reflect learner skills. All these conditions are vital in evaluating TIMSS score reliability using G-theory.

In G theory, there are two types of coefficients that can be computed: G and Phi. These coefficients serve to distinguish between relative and absolute decision-making within G theory. Specifically, G and Phi coefficients enable the independent assessment of norm-referenced testing and criterion-referenced testing. Coefficients derived from relative error variance are determined by the interactions between different aspects of measurement and the items being measured, and they are referred to as generalizability coefficients (G coefficients). On the other hand, coefficients obtained from absolute error variance are based on the main effects of all factors involved in measurement, including different facets, and the interactions between these facets and the items being measured are represented by the Phi coefficient (symbolized as Φ). It's crucial to note that in assessing behavioral measurements' reliability, a G-study is crafted. It aims to separate and estimate variations from the measured object and possible measurement error facets. This approach emphasizes practicality and efficiency in examining these facets. Afterward, a D-study (decision study) utilizes insights from the G-study to customize measurement applications for specific purposes. During D-study planning, decision-makers outline the scope of generalization, specifying facets and levels for extending conclusions and the intended interpretation of the measurement (Rentz, 1987; Shavelson & Webb, 1991: 2003).

1.1. The Present Study

In the context of South Africa, where the TIMSS test is commonly utilized, it's crucial to assess the reliability of these test scores to ensure their accuracy and usefulness in educational decision-making. The 2019 Trends in Mathematics and Science Study (TIMSS) results revealed a decline in Grade 4 mathematics achievement in South Africa compared to the 2015 average score of 376 (Mullis et al., 2020). This has sparked a debate among South African educators about the reliability of TIMSS scores in gauging the performance of South African students relative to those in other countries. While TIMSS employed a relevant instrument to assess mathematics skills among South African learners based on the framework and content, there's a lingering question about the reliability of these test scores. However, the literature lacks a comprehensive assessment of the reliability of TIMSS scores in South Africa using a robust statistical framework. Existing research mainly relies on classical test theory, which has limitations in addressing multiple sources of error and various measurement facets. Various studies have explored the reliability of test scores using generalizability theory in different contexts. For instance, Akindahunsi and Afolabi (2021) evaluated the reliability of English Language examination scores in Nigeria and found high-reliability coefficients. Uzun et al. (2018) assessed the score reliability of dentistry students' communication skills and identified issues related to the task component's variance. Nalbantoglu-Yilmaz (2017) examined score reliability from self-, peer-, and teacher-assessments and found acceptable limits of reliability. Gugiu et al. (2012) investigated the reliability of grades assigned to research papers and discovered high interrater reliability. Atilgan (2008) used generalizability theory to assess the score reliability of the special ability selection examinations for music education programs in higher education and concluded that

both the relative severity of raters and the relative difficulty of tasks are reported as the variance component of facets.

However, most of these studies focus on test scores outside of the TIMSS context, particularly those for grade 4 mathematics. Notably, there's no comparable study that examines the reliability of TIMSS test scores using generalizability theory, specifically for South African learners in the 2019 TIMSS mathematics assessment for Grade 4. This study addresses this gap by assessing the reliability of TIMSS test scores for South African learners using the Generalizability Theory. The study's uniqueness lies in its utilization of TIMSS mathematics achievements and scores from the IEA IDB Analyzer Merge module. By applying Generalizability Theory, the study aims to offer a more accurate and detailed understanding of the reliability of TIMSS test scores for South African learners. Consequently, the research seeks to fill the existing gap in the literature by answering the research question: What is the reliability of TIMSS test scores for learners in South Africa, analyzed through the lens of Generalizability Theory?

The next section outlines the methodology, encompassing participant details, used instruments, and conducted statistical analyses. Subsequently, the third section presents the obtained results, and the paper concludes with a discussion section, which includes final remarks and practical implications.

2. Methodology

The study employed a crossed one-facet design where all conditions of one facet are observed alongside all conditions of every other facet. For instance, in this design, denoted as a pxi design, each individual's measurement is taken for each item, symbolized as X_{pi}. The research utilized mathematics achievement data from the TIMSS 2019 4th-grade assessment in South Africa, accessible from the IEA repository (https://www.iea.nl/data-tools/repository/timss) (IEA, 2021). All 4th-grade students who participated in TIMSS 2019 from South Africa were included in the study. The selection of participants for TIMSS 2019 was meticulously carried out using a systematic random approach to ensure a representative sample of all 4th-grade students in South Africa, encompassing a diverse range of schools. The assessment included 11,891 students, hailing from 298 distinct schools nationwide. The gender distribution in the sample was almost balanced, with 49.4% male, 50.4% female, and a minimal exclusion of 12 data entries (0.2%). Notably, the assessment received a high response rate for all questions, highlighting the comprehensive completion of the evaluation.

The research incorporated data gathered from students who participated in the 2019 TIMSS assessment cycle, responding to a 35-item test that included both multiple-choice and constructed response questions. TIMSS is a global evaluation of math and science skills, conducted every four years since 1995, targeting students in 4th and 8th grades. Notably, it's important to highlight that South Africa joined the TIMSS initiative in 2015. The primary goal is to analyze trends in student achievement alongside contextual data. In TIMSS 2019, 58 countries participated, constructing the assessment based on frameworks established by each country for various curriculum areas and grades. The majority of items are designed to evaluate students' application and reasoning skills (Mullis et al., 2020). In TIMSS, student achievements are represented using five plausible values. For this study, these five plausible values were employed as a measure of mathematics achievement for South African grade 4 students in the 2019 assessment. The values were obtained using IEA IDB Analyzer 4.0.12 (2018) and SPSS software version 26.0. The analysis involved the utilization of various tools,

including the "*gtheory*" package (Moore, 2016), along with functions like *aov(), gstudy(), dstudy(),* and others within the R programming language and statistical computing environment (R Core Team, 2021). These tools were employed to compute parameters such as the g-coefficient, phi coefficient, and the D study. The D study, which identifies the most suitable number of conditions for each aspect to optimize reliability, was deduced from the G study variance components. The specific R code implementations, adapted from Huebner and Lucht (2019), can be found in the appendix.

3. Results

The G-study procedure calculates the variances associated with the measured entities (e.g., students/persons) and the different aspects (e.g., tasks/items), along with variance that cannot be explained. This analysis quantifies the level of error when extending a student's 2019 TIMSS 4th grade mathematics test score to the overall score of the population. An effective way to interpret the estimated variances in a G-study is by determining the proportion of the total variance that each variance component signifies. Table 1 displays the assessed variance corresponding to each of these components.

Source	Effect	df (α)	SS (α)	MS (α)	$\sigma^2(\alpha)$	Percent of variability
person	σ^2_p	149	9372938	62906	1777.14	59.9
Item/task	σ^2_i	34	2487749	73169	483.09	16.3
residual	$\sigma^2_{pi,e}$	5066	3573100	705	705.31	23.8

Note: SS (α) – sum square, MS (α)- mean square, df (α)- degree of freedom, $\sigma^2(\alpha)$ - estimated variance

Table 1. Summary of G-study statistics for the p x i design

Table 1 demonstrates that the variance stemming from students, representing the overall score variability, constitutes 59.9% of the entire variance, which appears notably substantial. Additionally, the variance attributed to items making up 16.3% of the total variance is somewhat smaller compared to the universe score variance, but it is also less than the residual variance amounting to 23.8% of the total variance. This implies that the variation in students' scores comprises a significant portion of the overall variability, possibly suggesting a substantial diversity in the abilities and backgrounds of the students. The variance arising from the individual items contributes a moderate degree of variability, signifying potential differences in item difficulty or clarity. The residual variance, on the other hand, represents unexplained factors that impact scores beyond students and items, highlighting the presence of other sources of variability that the analysis may not have captured. This underlines the complexity of factors affecting test scores that go beyond individual characteristics and item quality. Furthermore, Table 2 illustrates the calculation of the generalizability coefficient employing the following mathematical expression:

$$\mathbf{E}\rho 2 = \frac{\sigma 2(\tau)}{\sigma 2(\tau) + \sigma 2(\delta)} = \frac{1777.142}{1777.142 + 20.151} \approx 0.989 \qquad \text{Eqn.1}$$

Source	Effect	Estimate
Universe score variance	$\sigma^2(\tau)$	1777.142
Relative error variance	$\sigma^2(\delta)$	20.151
Generalizability coefficient	$\mathbf{E}\rho^2$	0.989

Note: $\sigma^2(\tau) = \sigma^2(p)$ due to the consideration that all facets are treated as "random." This signifies that the sample size is significantly smaller than the population.

Table 2. Generalizability coefficient *(N = 35)*

Table 2 provides the essential elements utilized to calculate the generalizability coefficient for the 35-item 2019 TIMSS 4th-grade mathematics test. According to the table, the computed generalizability coefficient for the test stood at 0.989. This high generalizability coefficient value signifies a considerable level of reliability associated with the test. In essence, the test demonstrates a high degree of consistency in measuring the mathematical abilities of 4th-grade students in South Africa. The implication of this high generalizability coefficient is that the test results are dependable and consistent.

Moreover, in order to evaluate the dependability coefficient, a D-study was conducted, utilizing the foundation established by the preceding G-study. This procedure enabled the determination of the reliability of the TIMSS test, as showcased in Table 3. Much like the approach adopted for the generalizability coefficient, the analysis utilized the dstudy () function. The computation of the dependability coefficient follows this mathematical expression:

$$\boldsymbol{\phi} = \frac{\sigma 2(\tau)}{\sigma 2(\tau) + \sigma 2(\Delta abs)} = \frac{1777.14}{1777.14 + 33.95} \approx 0.981 \qquad \text{Eqn.2}$$

Source	Effect	Estimate
Universe score variance	$\sigma^2(\tau)$	1777.14
Absolute error variance	$\sigma^2(\Delta abs)$	33.95
Dependability coefficient	ϕ	0.981

Table 3. Dependability coefficient

Table 3 presents the elements utilized to calculate the dependability coefficient for the 35-item 2019 TIMSS 4th-grade mathematics test in South Africa. The findings reveal that the calculated dependability coefficient for the test amounted to 0.981. This outcome underscores that the test scores were notably reliable and consistent. The implication of this high dependability coefficient is that the test results are dependable and stable, thus reflecting the students' mathematical abilities consistently. This reliability suggests that the test scores can be confidently used to assess students' math skills, guide educational decisions, and monitor progress over time. Consequently, educators and policymakers can place strong trust in the test outcomes as a reliable tool for evaluating students' mathematics proficiency and making informed educational choices.

Additionally, to establish the D-study concerning the test items, the corresponding values for various alternative n values (found in columns three through six of Table 4) can be established using the equations provided in (1) and (2). As an illustration, the variance components *σ2(I)* and *σ2(pI)* for the D-study, corresponding to the case where n=2, are computed by dividing the variance components from the G-study, *σ2(i)* and *σ2(pi)*, by 2. Table 4 provides an overview of the D-study statistics for the design based on *p x i*.

$\sigma^2(\tau)$	$\acute{n}_i$	25	20	15	10
$\sigma^2(p) = 1777.14$	$\sigma^2(p)$	1777.14	1777.14	1777.14	1777.14
$\sigma^2(i) = 483.09$	$\sigma^2(I)$	19.32	24.15	32.21	48.31
$\sigma^2(pi) = 705.31$	$\sigma^2(pI)$	28.21	35.27	47.02	70.53
	$\sigma^2(\delta)$	28.21	35.27	47.02	70.53
	$\sigma^2(\Delta abs)$	1.36	1.69	2.26	3.39
	$\boldsymbol{E}\rho^2$	0.984	0.981	0.974	0.962
	ϕ	0.999	0.999	0.999	0.999

Table 4. Summary of D-study statistics for the p x i design

Table 4 depicts that as the number of items decreases, the proportion of variance that can be explained by the study, denoted as Eρ2, displays a declining pattern. Starting at 0.984 for 25 items, it decreases to 0.981 for 20 items, further drops to 0.974 for 15 items, and finally reaches 0.962 for 10 items. This pattern suggests that with fewer items, the extent to which the study can account for the variability in the measurements diminishes, leading to a gradual reduction in the explained variance. However, on the other hand, the reliability coefficient, represented as ϕ, remains consistently high throughout the reductions in item numbers. It remains at an elevated level of 0.999 regardless of whether the number of items is 25, 20, 15, or 10. This consistently high-reliability coefficient indicates that even with fewer items in the measurement, the results remain dependable and stable.

4. Discussions and Conclusion

The performance of Grade 4 students in South Africa concerning the 2019 TIMSS Mathematics assessment has been scrutinized using generalizability theory through single-facet designs, and the ensuing findings are outlined below. In this framework, items or tasks are considered as the focal measurement object in completely crossed designs. In this particular setup, the estimated variance component associated with items exhibits a notably limited impact in elucidating the overall variance. While conventional wisdom suggests that the measurement object should significantly contribute to explaining total variance, existing literature demonstrates instances where variance percentages of the measurement object are low, particularly when the attributes being measured do not exhibit substantial differentiation. In this study, it has been deduced that the items or tasks within the 2019 TIMSS mathematics test exhibit insignificant disparities in terms of difficulty levels. This observation resonates with the findings of Akindahunsi and Afolabi (2021), Atilgan (2008), and de Vries (2012), which imply that a considerable portion of the error variance in the examination might be attributed to the interplay between individuals and items. Lowering this variance could result in heightened dependability. Furthermore, upon examining individual students, it becomes apparent that the variance associated with the student component is notably significant. This observation suggests that the 2019 TIMSS participants differ in terms of their performance within the context of the assessment. This aligns with previous findings reported by Yılmaz and Gelbal (2011). In conclusion, the G-Study analysis offers a comprehensive panorama of the contributing elements to Grade 4 TIMSS mathematics scores in South Africa. It underscores the significance of students' unique attributes, the quality of test items, and recognizes the existence of unexplained variance. These insights present the opportunity for refining mathematical education strategies, test development, and assessment methodologies, ultimately culminating in more meticulous and insightful evaluations of students' mathematical capabilities.

Furthermore, the outcomes derived from the generalizability coefficient underscore the likelihood that the test results accurately mirror students' mathematical skills and knowledge, unaffected by random or extraneous influences. This conclusion resonates with earlier investigations by Gugiu et al. (2012) and Yilmaz (2017). Additionally, the dependability coefficient (Φ), a metric reflecting the measurement procedure's contribution to the test score's reliability, emerges as highly dependable. This aligns with the assertions of Akindahunsi and Afolabi (2021); Brennan (2003), who posit that values approaching unity (1) indicate the capability to discern scores of interest with notable accuracy, even amidst random measurement fluctuations. Notably, Φ offers the advantage of pinpointing error sources that undermine classification precision and devising strategies for enhancing these classifications. While most authors typically explore variability across facets to identify the most beneficial factor for generalizability, this outcome aligns with the findings by Fosnacht and Gonyea (2018). Additionally, the consistent high-reliability coefficient confirms that even when the measurement employs a reduced number of items, the results maintain their reliability and stability. Yin and Wiley (2015) corroborate this notion by affirming that expanding the number of items reduces error variance while simultaneously elevating both G and phi coefficients. Succinctly, the findings underscore the robustness of the assessment. The generalizability coefficient implies trustworthy reflections of student abilities, supported by previous studies. The dependability coefficient reinforces measurement precision, is consistent with expert opinions. Additionally, the enduring high-reliability coefficient endorses the reliability of results even with fewer items. These observations not only confirm existing research but also contribute to a better understanding of measurement quality and the factors influencing it.

5. Implications

The findings carry significant implications for the realm of educational assessment. To begin with, the reduction in the proportion of explained variance emphasizes a delicate equilibrium between assessment comprehensiveness and practical constraints like time limitations or participant fatigue. It is imperative to carefully navigate this equilibrium, considering the interplay between item count and the extent of measurement precision. Additionally, the enduring high-reliability coefficient signifies that, even in situations where there's a necessity to curtail the number of items, educators and policymakers can still place confidence in the retained items to generate consistent and reliable outcomes. Furthermore, the identification of sources of variability in the scores provides policymakers with valuable insights to make well-informed decisions concerning the enhancement of educational quality within the country. This offers a pathway to strategically address areas that contribute to variance, enabling targeted interventions to uplift educational practices.

Acknowledgments

We acknowledge the International Association for the Evaluation of Educational Achievement (IEA) for providing the Trends in International Mathematics and Science Study (TIMSS) data for this study. Their commitment to making this data available for research purposes is greatly appreciated, as it contributes to the advancement of educational research and our understanding of global educational trends. Thank you, IEA, for your valuable contribution to the academic community.

Appendix- R codes

```
# get working directory

getwd()
# set working directory

setwd("C:/Users/DELL/OneDrive/Download 6/Generalisability theory")

# call package for the generalizability theory analysis

library(gtheory)

# read dataset into R environment

Person <- as.factor(rep(1:150,each = 35))

Item <- as.factor(rep(1:35,times = 150))

Score<-
c(582.69,616.59,570.78,648.9,658.69,682.77,667.15,624.15,611.69,621.51,682.32,671.44,58
5.7,616.21,609.22,673.87,671.28,616.33,607.96,701.09,596.1,660.19,626.15,616.2,699.75,58
9.11,628.68,582.25,628.33,682.67,579.98,623.44,619.26,600.61,657.1,526.63,497.17,561.69,
469.71,548.32,525.4,537.62,582.21,533.37,551.16,500.12,562.58,548.75,519.32,549.18,540.
5....)

Timss_dat <- data.frame(Person,Item,Score)

# to perform analysis of variance

ANOVA<- summary(aov(Score~Person+Item, data  = Timss_dat))

# extracting the ANOVA output

sink()

sink("ANOVA_ANALYSIS.TXT")

ANOVA

sink()

# to perform G-study variance components from ANOVA results

formula1 <- Score ~ (1|Person)+(1|Item)

g_study <- gstudy(data = Timss_dat, formula1)
g_study$components

# extracting the G-study output
```

```
sink()

sink("GSTUDY_ANALYSIS.TXT")

g_study$components

sink()

# to perform D-study component

d_study <- dstudy(g_study,colname.objects="Person",colname.scores="Score",data=
Timss_dat)
d_study$components

# extracting the D-study component output

sink()

sink("DSTUDY_ANALYSIS.TXT")

d_study$components

sink()

# to perform universe score variance

d_study$var.universe

# to perform relative error variance

d_study$var.error.rel

# to perform generalizability coefficient

d_study$generalizability

# to perform dependability coefficient(phi)

d_study$dependability
```

References

Akindahunsi, O., & Afolabi, E. R. I. (2021). Using Generalizability Theory to Investigate the Reliability of Scores Assigned to Students in English Language Examination in Nigeria. Journal of Measurement and Evaluation in Education and Psychology, 12(2), 147-162.

Atilgan, H. (2008). Using generalizability theory to assess the score reliability of the special ability selection examinations for music education programmes in higher education. International Journal of Research & Method in Education, 31(1), 63-76.

Ayanwale, M. A. (2019). Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous response mathematics achievement tests in Osun State, Nigeria. Nigeria.

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2018). An assessment of item statistics estimates of basic education certificate examination through classical test theory and item response theory approach. International Journal of Educational Research Review, 3(4), 55-67.

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2019a). Invariance Person Estimate of Basic Education Certificate Examination: Classical Test Theory and Item Response Theory Scoring Perspective. Journal of the International Society for Teacher Education, 23(1), 18-26.

Brennan, R. L. (2001). Generalizability theory. New York, NY: Springer.

Brennan, R. L. (2003). *Coefficients and indices in generalizability theory* (CASMA Research Report Number 1). Iowa: Centre for Advanced Studies in Measurement and Assessment.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. U.S.A: Cengage Learning.

de Vet, H.C., Terwee, C.B., Mokknink, L. B., & Knol, D.L. (2011). Measurement in medicine: A practical guide. New York, NY: Cambridge University Press.

Fishbein, B., Foy, P., & Yin, L. (2021). TIMSS 2019 User Guide for the International Database (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-database/

Fosnacht, K., & Gonyea, R. M. (2018). The dependability of the updated NSSE: A generalizability study. *Research and Practice in Assessment 13*, 62-73. Retrieved from https://eric.ed.gov/?id=EJ1203503

Gugiu, M. R., Gugiu, P. C., & Baldus, R. (2012). Utilizing Generalizability Theory to Investigate the Reliability of Grades Assigned to Undergraduate Research Papers. Journal of Multidisciplinary Evaluation, 8(19), 26–40. https://doi.org/10.56645/jmde.v8i19.362

Huebner, A. & Lucht, M. (2019). Generalizability Theory in R. Practical Assessment, Research & Evaluation, 24(5). Available online: http://pareonline.net/getvn.asp?v=24&n=5

IDB Analyzer V4 (version 4.0.12.0) [computer software]. (2018). Hamburg: IEA Data Processing and Research Center. Retrieved from https://www.iea.nl/data

IEA (2021). TIMSS 2019 international database. Retrieved from the IEA website 20230716. https://www.iea.nl/data-tools/repository/timss

Johnson, S., & Johnson, R. (2009). Conceptualising and interpreting reliability. Coventy: Ofqual.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). Methods and Procedures: TIMSS 2019 Technical Report. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods

Moore, C. T. (2016). gtheory: Apply Generalizability Theory with R. R package version 0.1.2. Retrieved from https://CRAN.R-project.org/package=gtheory

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international results

Nalbantoglu-Yilmaz, F. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. Educational Sciences: Theory & Practice, 17(2), 395-409. doi:10.12738/estp.2017.2.0098

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rentz, J. O. (1987). Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures. Journal of Marketing Research, 24(1), 19–28. doi:10.1177/002224378702400102

Shavelson, R. J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Shavelson, R.J. & Webb, N.M. (2003). “Generalizability Theory”. In Encyclopedia of Social Measurement, Edited by: Kempf□Leonard, K. San Diego: Academic Press.

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Ed.), Score reliability: contemporary thinking on reliability issues (pp. 43-58). Thousand Oaks, CA: Sage Publications.

Uzun, N. B., Aktas, M., Asiret, S., & Yormaz, S. (2018). Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students. Asian Journal of Education and Training, 4(2), 85-90.

Contact email: ayanwalea@uj.ac.za