Do not quote without permission of authors

Modeling Science Achievement in South Africa Using International Achievement Data: Some Comparability Validity Issues

Dr. Kelvin Gregory
Flinders University
kelvin.gregory@flinders.edu.au

Dr. Anil Kanjee
Human Sciences Research Council
AKanjee@hsrc.ac.za

International comparative studies of achievement facilitate national reflection on educational practice and experience. Such studies use sophisticated assessment designs, detailed questionnaires, complex scaling methodologies and stringent quality control systems to produce comparable data. These data enable international ranking of countries based on a common assessment framework, the identification of indicator variables, and the modeling of achievement. At the heart of these studies is a belief that the assessment enables valid comparisons of achievement. Acknowledging that the international assessment include items that may not be covered by a specific national curriculum, the organizers argue that the removal of a small number of items unsuitable for any particular country will not appreciably change achievement scores, international ranking, or achievement model. This study investigates this claim using grade 8 science achievement TIMSS 2003 for South Africa within a comparative validity framework using construct, scale and measurement equivalence.

We performed a preliminary investigation of construct equivalence. A panel of South African science education experts reviewed the TIMSS 2003 science items, identifying items that were suitable for South Africa and those that were not. Based upon the assumption that TIMSS items are scaled appropriately, we used the published TIMSS item response theory item parameters to produce scale scores for students using all the TIMSS science items and then only those items deemed appropriate for South Africa. As expected South African science achievement scores increased as the match between TIMSS science items and the intended South African curriculum became closer. However, this increase was small, approximately one-quarter of a standard deviation. We acknowledge that the deletion of a substantial number of science items changes the assessment framework but argue that South African achievement is more meaningful in the context of what is intended to be taught to South African students.

We then investigated scale and measurement equivalence within the item response theory model used by TIMSS. Our analyses showed that there is a preponderance of item misfit, indicating a lack of scale and measurement equivalence. While undoubtedly South African students performed poorly on the TIMSS 2003 science items, we caution against using the TIMSS scales to compare South African students to students in other countries.

Introduction

International comparative assessments can have a considerable effect on national education systems (Brown, 1999), with significant opportunity to feed specific political agendas (Adler & Lerman, 2003) . For example, the Trends in International Mathematics and Science Studies (TIMSS) have forced Cypriots to admit that their "education system is not working" (Papanastasiou, 2000, p. 39) . South Africa reacted to TIMSS -95 low results with an education minister stating that particular attention would be "paid to the mismatch between the South African and international curriculum" (Howie & Hughes, 2000, p. 143).

South Africa decided to participate in TIMSS for a number of reasons with a main reason being the opportunity to benchmark progress in mathematics and science achievement in a post-apartheid era even though many of the students would complete the assessments in their second or third language and answer largely unfamiliar multiple choice questions (Bishop, Clements, Keitel, Kilpatrick, & Leung, 2003; Howie, 1998)

TIMSS assessments strive for comparative validity by addressing a number of psychometric questions (Mullis & Martin, 2007). For example, they use a test-curriculum matching analysis (TCMA) to evaluate the degree of congruence between the international mathematics and science assessments and the intended curriculum:

> "To gather data about the extent to which the TIMSS 2003 tests were relevant to the curricula of the participating countries, each NRC reported whether each item was in that country's intended curriculum at the grade tested (eighth or fourth grade in most countries). The NRC was asked to choose a person or persons who were very familiar with the curriculum at these grades to make this determination. Since an item might be in the curriculum for some but not all students in a country, an item was to be determined appropriate if it was in the intended curriculum for more than 50 percent of the students. The NRCs had considerable flexibility in selecting items and may have considered items inappropriate for other reasons" (M. O. Martin, Mullis, Gonzalez, & Chrostowki, 2004, p. 412)

TIMSS also uses measures to ensure that each country has representative samples of a defined population of students, procedures to ensure that assessment items are appropriately translated, and well-documented assessment procedures, including quality control procedures, to help ensure that assessments are implemented in a standardized manner across all participating countries (M. O. M. Martin, Ina V.S. & Chrostowski, 2004). Mullis and Martin (2007) argue that this insistence on "comparative validity" has become a hallmark of TIMSS and is one of the reasons that international comparative data on student achievement have become accepted as reliable instruments of educational policy analysis .

Validity is a property of the meaning attached to, and use of, scores (Kane, 2006; Messick, 1989) and TIMSS with its extensive documentation enables an exploration of the comparative validity of its assessment scores. Comparative validity may be defined as “the appropriateness, meaningfulness, and usefulness of comparative inferences made from” assessment scores (Bechger, van den Wittenboer, Hox, & de Glopper, 1999, p. 19). This definition builds upon a widely accepted conceptualization of validity developed by Messick (Kane, 2006; Messick, 1989). The appropriateness of comparative statements hinges on three basic ideas of construct equivalence, scale equivalence, and measurement equivalence (Bechger et al., 1999).

Construct Equivalence

Construct equivalence indicates that the international assessment measures the same construct for each country. If the analyses are conducted for subgroups within countries, then construct equivalence extends to these groups as well. For example, TIMSS routinely reports differences in achievement between girls and boys (M. O. Martin et al., 2004). Supporting arguments for construct equivalence can come from evidence showing that the international assessment frameworks apply to all groups. That is, the content and cognitive components of each assessment framework are universally applicable. While multi-group confirmatory factor analysis is one source of support for construct equivalence, the TCMA used by TIMSS serves as a simple indicator of construct equivalence.

Scale and Measurement Equivalence

Scale equivalence requires that the assessment responses can be placed on a common scale. Earlier IEA studies used classical test theory (Keeves & Schleicher, 1992) while IEA studies since TIMSS 1995 have used item response theory (Yamamoto & Kulick, 2000). Measurement equivalence is established when all countries (and subgroups) satisfy the conditions for measurement according to the formal measurement theory being used. TIMSS argues for measurement equivalence, in part, through its presentation of item response functions graphs.

This paper explores comparative validity from a South African perspective. It examines construct validity through the TCMA and the impact on scale scores of the inclusion of science items not covered by the South African curriculum. The paper also explores scale and measurement equivalencies from a South African perspective using available TIMSS 2003 information.

TIMSS 2003 Construct Equivalence and South Africa

TIMSS studies are built upon a fundamental curriculum framework used since the Second International Mathematics Study (Robitaille, Beaton, & Plomp, 2000). This study highlighted “the distinction between that which a society would like to have taught, the intended curriculum; that which is taught, the implemented curriculum; and that which students learn, the attained curriculum" (Robitaille et al., 2000, p. 12) . From TIMSS 2003 onwards, TIMSS has shifted from a curriculum focus towards an assessment-orientated approach as evidenced by the development of assessment frameworks. For example, the "*TIMSS Assessment Frameworks and Specifications 2003* is the publication that describes in some detail the mathematics and science content to be assessed" (Mullis et al., 2003).

The quantitative comparison of countries demands a common frame of reference (Mislevy, 1995). Beaton et al (1996) noted that "when comparing student achievement across countries, it is important that the comparisons be as 'fair" as possible" (p. B-1). Noting that TIMSS achievement results may be interpreted as "bad news" Robitaille, Beaton, and Plomp (2000) foresaw the problem that the TIMSS item pool might not adequately or appropriately reflect the national curriculum and the items, especially the multiple choice items, may be unfamiliar to students and therefore negatively affect performance. Since TIMSS 1995, study coordinators have required National Research Coordinators (NRCs) from each country to formally approve the TIMSS tests, "thus accepting it as being sufficiently fair to compare their students' science achievement with that of students from other countries" (Beaton et al., 1996, pp. B-1).

The Test-Curriculum Match Analysis reported by TIMSS 2003 shows that approximately 52% of the science items were deemed to be appropriate for South Africans (M. O. Martin et al., 2004). In terms of score points available, only 106 out of the available 206 points mapped onto the South African science curriculum. Given the large discrepancy between the TIMSS assessed science curriculum and the intended South African science curriculum, it is highly doubtful that there is construct equivalence.

We explored scale and measurement equivalence in two ways. Scored cognitive assessment data was obtained from the TIMSS website and scored using provided scoring programs. Item response theory parameters were obtained from the TIMSS 2003 technical report (M. O. M. Martin, Ina V.S. & Chrostowski, 2004). The TIMSS technical report refers to a number of derived items. Since these are not documented, but probably reflect the recoding of partial-credit items, these items were removed from the analyses. The appropriateness of the grade 8 science items for South African students was obtained from the test –curriculum matching analysis (M. O. M. Martin, Ina V.S. & Chrostowski, 2004).

Since 1999 TIMSS has performed its scaling with item response theory using the three-parameter Birnbaum model for dichotomous items and the generalized partial credit model for polytomous items. We followed TIMSS by using the commercial version of PARSCALE 4.1. TIMSS uses plausible value methods to obtain group-orientated proficiency estimates. These Bayesian-derived estimates combine the assessment information with an informative prior. The strength and location of the prior is established using specialized software that essentially regressing proficiency onto principal components obtained from all available contextual information. We used the simpler Warm's weighted maximum likelihood estimation available in PARSCALE. All statistical analyses were performed using SAS 9.1. Since TIMSS uses a stratified, cluster sample design we used jackknife methods to compute standard errors.

The mean of the first TIMSS 2003 science plausible value for South Africa, on the theta metric, is -2.24 with a standard error of .046. we obtained a mean for the Warm's weighted likelihood estimate for South Africa , using all science items and TIMSS item parameters, of -1.92 (.042). When only items deemed suitable for South Africa are used the weighted likelihood estimate mean is significant higher at -1.67 (.037). In practical terms, this difference between when all TIMSS science items and only South African-match items are used would be approximately 25 points on the TIMSS reporting metric. South African students do better when the assessment more closely matches the South African curriculum; that is, when the TIMSS attained curriculum model matches the intended South African curriculum model. In contrast, the average difference between Warm's weighted likelihood estimates obtained using the two groups of items is -.03 on the theta metric across the other 44 countries included in our analyses of TIMSS 2003 science data. Only seven of the these forty-four countries had absolute differences in the means greater than .1, and no other country had a difference greater than .2. However, the ranking of South Africa stayed the same as that reported in the TIMSS 2003 regardless of the item pool. Statistically, South African science achievement was lower than all other countries, except Ghana, regardless of the method used to estimate the science proficiency.

Distribution of South African Science Achievement

We explored the shape of South African science achievement distributions, plotting histograms of the first science plausible value (in theta metric; see Figure 1), the weighted likelihood estimate obtained using all TIMSS 2003 science items (see figure 2) and the weighted likelihood estimate obtained using TIMSS 2003 science items considered to be part of the intended curriculum (see Figure 3).

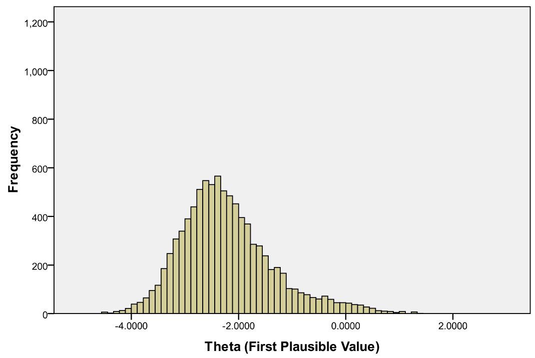

Figure 1: Distribution of first science plausible value for South Africa (TIMSS 2003)

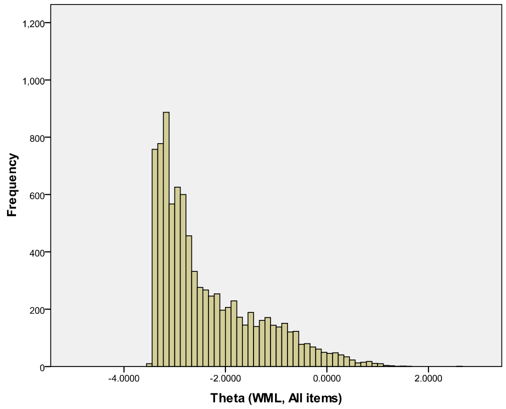

Figure 2: Distribution of weighted likelihood estimates of achievement using all science items for South Africa (Data: TIMSS 2003)

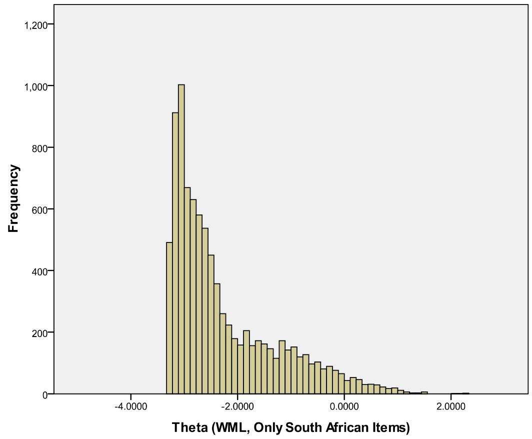

Figure 3: Distribution of weighted likelihood estimates of achievement using TCMA-matched science items for South Africa (Data: TIMSS 2003)

The shape of the two weighted likelihood estimates distributions are remarkabley similar, with both showing a pronouced floor effect and both being positively skewed. In contrast, the TIMSS 2003 first plausible value histogram lacks the floor effect, has a more obvious normal shape, and has a large number of theta values below that found in either of the weighted likelihood estimates distributions. Plausible values are imputed numbers calculated using a normal prior which is in turn based upon a linear model linking a Bayesian achievement estimate to contextual factors. While outside the scope of this paper, we suspect that the veracity of this model to a developing country context like South African merits further research. Since parametric regression approaches to modeling achievement seek to account for the variance of the proficiency estimate, the shape of the score distribution will impact the model.

Thus, the need to verify the veracity of the plausible value methodology in a developing country context should precede the use of TIMSS data-based modeling.

TIMSS 2003 Scale and Measurement Equivalence for South Africa

As noted above, TIMSS 2003 showed the applicability of the item response theory models using item response function graphs. These graphs show the predicted probability of obtaining a correct response as a solid line. The empirical probability of attaining a correct response, at a range of theta points, is calculated as the percentage of students obtaining a correct score. These probabilities are presented on the graphs as circles, with the size of the circles indicating the relative number of students with that theta value. Good item fit, that is close agreement between the predicted and observed probabilities, is one indicator that the item response theory model is appropriate. Figures 4 and 5 show indicative empirical and predicted item response curves obtained using TIMSS item parameters and scored South African data for students who completed assessment book number 11. All three items show considerable disagreement between the observed percent correct and the predicted percent correct at each theta point. The prevalence of misfit in these and other items for South Africa supports the notion that the TIMSS 2003 scaling model does not adequately model South African achievement. That being the case, scale and measurement equivalence is not established for South Africa and this country's science achievement should not be compared to other countries using the TIMSS metric.

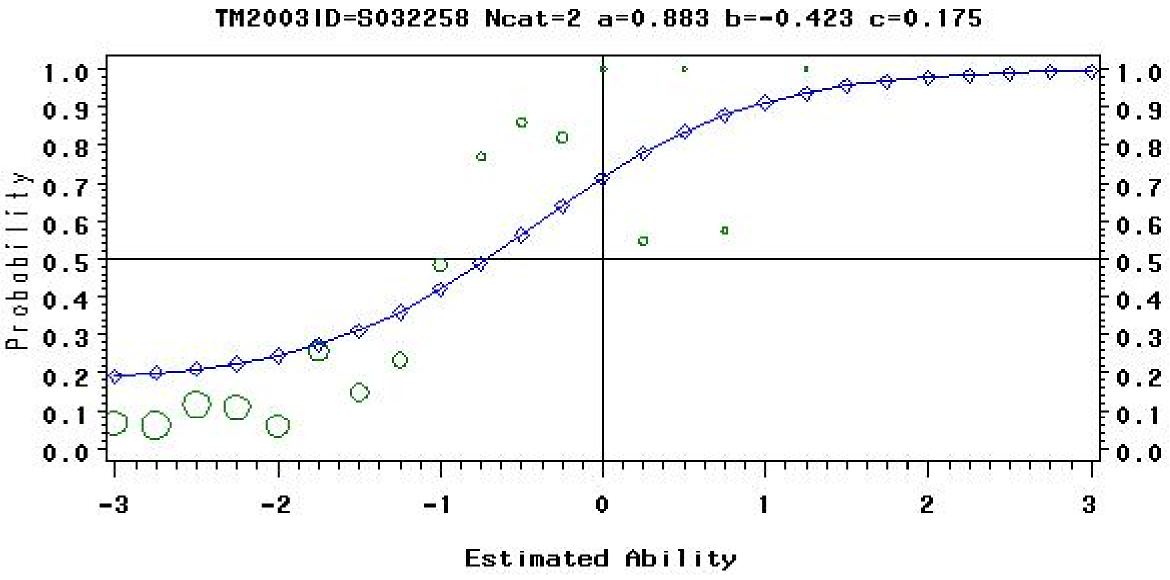

Figure 4: Empirical and predicted item curves for S032258 (Data: TIMSS 2003, test book 11)

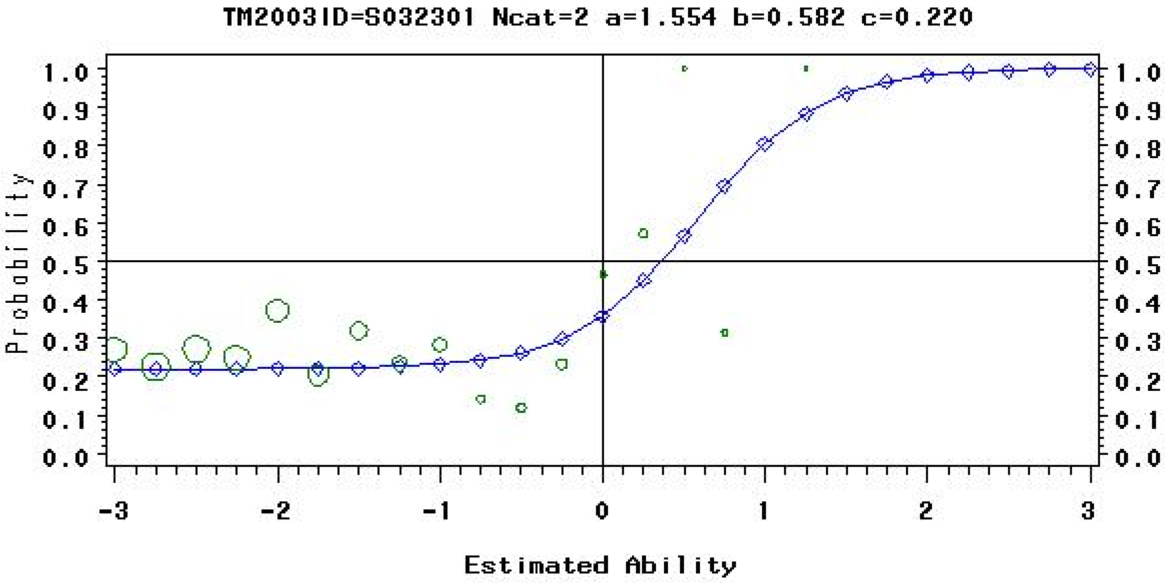

Figure 5: Empirical and predicted item curves for S032301 (Data: TIMSS 2003, test book 11)

Discussion

There is little doubt that in TIMSS demonstrates considerable psychometric expertise. The technical documentation alone is a testament to this expertise. However, as demonstrated in this paper, there are significant shortcomings in this documentation and specifically the psychometric demonstrations of scale and measurement equivalence in TIMSS is wanting. Without demonstrable scale and measurement equivalence, TIMSS as a comparative assessment survey is likely to provide misleading information.

As argued by others (Keitel & Kilpatrick, 1999) – something missing here has not addressed several crucial questions that extend well beyond scale and measurement considerations and impinge on the nature of construct equivalence. Pedagogy is “a window on the culture of which it is a part, and on that culture’s underlying tensions and contradictions as well as its publically declared educational policies and purposes” (Alexander, 2001, p. 4). Curriculum, whether it be intended, implemented or attained, is intimately linked with culture and pedagogy and this is effectively ignored by the TIMSS assessment frameworks. From the South African perspective, with its rainbow of cultures and specific challenges, there is a danger that the international assessments will force curricula changes on the country which “ may hamper learning and as a consequence will not contribute to a binding of cultures but to isolation and feelings of inferiority” (Vedder, 1994, p. 5). South African ministerial responses to TIMSS 1995, 1999, and 2003 indicate that this threat is very real.

While there are very real dangers in participating in international assessments, there is also much a developing country can learn from such assessments. Assessments like TIMSS, PIRLS, and PISA represent some of the best practices in large-scale assessments. However, we concur with Adler and Lerman (2003) who argue that there is an ethical dimension to such studies that has been largely overlooked. Simply put, since studies like TIMSS have the capacity to impact students' learning, the test developers have an ethical requirement to demonstrate that their assessments, at the very least, meet the standards for comparative validity.

Adler, J., & Lerman, S. (2003). Getting the description right and making it count: Ethical practice in mathematics education research. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (Vol. 10). Dordrecht: Springer International Handbooks of Education

Alexander, R. (2001). Pedagogy and culture. In J. Soler, A. Craft & H. Burgess (Eds.), *Teacher development: Exporing our own practice*. London: Paul Chapman Open University.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.

Bechger, T. M., van den Wittenboer, G., Hox, J. J., & de Glopper, C. (1999). The validity of comparative educational studies. *Educational Measurement: Issues and Practice, 18*(3), 18-26.

Bishop, A. J., Clements, M. A., Keitel, C., Kilpatrick, J., & Leung, F. K. S. (Eds.). (2003). *Second International Handbook of Mathematics Education* (Vol. 10). Dordrecht: Springer International Handbooks of Education

Brown, M. (1999). Problems of Interpreting International Comparative Data. In B. Jaworski & D. Phillips (Eds.), *Comparing Standards Internationally: Research and Practice in Mathematics and Beyond*. Oxford: Symposium Books.

Howie, S. (1998). *South Africa in TIMSS:The value of international comparative studies for a developing country.* Paper presented at the Perspectives on the Third International Mathematics and Science Study, Johannesburg: University of Witwatersrand.

Howie, S., & Hughes, C. (2000). South Africa. In D. F. Robitaille, A. E. Beaton & T. Plomp (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science*. Vancouver: Pacific Educational Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Portsmouth, NH: American Council for Education/Praeger.

Keeves, J. P., & Schleicher, A. (1992). Changes in Science Achievement. In J. P. Keeves (Ed.), *The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984* (pp. 263-290). Oxford: Pergamon Press.

Keitel, C., & Kilpatrick, J. (1999). The rationality and irrationality of international comparative studies In G. Kaiser, L. Eduardo & I. Huntley (Eds.), *International Comparisons in Mathematics Education* (pp. 241-256). London: Falmer.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowki, S. J. (2004). *TIMSS 2003 International Science Report*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M. O. M., Ina V.S. , & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York, NY: Macmillan Publishing Company.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*(4), 419-437.

Mullis, I. V. S., & Martin, M. O. (2007). TIMSS in Perspective: Lessons Learned from IEA's Four Decades of International Mathematics Assessments. In T. Loveless (Ed.), *Lessons Learned: What International Assessments Tell Us about Math Achievement* (pp. 9-36). Washington, D.C.: Brookings Institution Press.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., et al. (2003). *TIMSS Assessment Frameworks and Specifications 2003* (2nd ed.). Chestnut Hill,: International Study Center, Lynch School of Education, Boston College.

Papanastasiou, C. (2000). Cyprus. In D. F. Robitaille, A. E. Beaton & T. Plomp (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 35-40). Vancouver: Pacific Educational Press.

Robitaille, D. F., Beaton, A. E., & Plomp, T. (Eds.). (2000). *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science*. Vancouver: Pacific Educational Press.

Vedder, P. (1994). Global measurement of the quality of education: A help to developing countries? . *International Review of Education/Internationale Zeitschrift für Erziehungswissenschaft/Revue internationale l'éducation, 40*(1), 5-17.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science studies. In M. O. Martin, K. D. Gregory & S. E. Stemler (Eds.), *TIMSS 1999 technical report: IEA"s repeat of the third international mathematics and science study at the eighth grade.* . Chestnut Hill, MA: Boston College.