Indonesian Achievement Scores for TIMSS 2003: Effects of Changing The Score Estimation Model

Diah Wihardini
Bina Nusantara Business School dwihardini@binus.edu

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

The Trends in International Mathematics and Science Studies (TIMSS), like other international assessment studies, uses a complex scaling methodology to produce population-orientated scores for participating countries. Based on item response theory (IRT), the plausible value methodology combines test information with contextual variables. This procedure enables estimates to be produced for each student providing at least some achievement or contextual information is available. Some researchers view that the combination of contextual information with achievement data to produce population measures as controversial. It is often argued that providing the assessment information dominates the scaling model, and that the plausible-value estimates are superior to other IRT measures. In this study, Indonesian mathematics data from TIMSS 2003 are used to investigate the importance of assessment data in the student plausible values.

The scored mathematics data from TIMSS, published item parameters and commercial IRT software were used to produce maximum likelihood (MLE), Warm's maximum likelihood (WML) and Expected A Posteriori (EAP) estimates for each student. The EAP estimates were produced using a number of priors, including uninformative and various normal informative priors. In some cases, the maximum likelihood and Warm's maximum likelihood procedures failed to give student estimates under a number of conditions.

We describe the model-fit plots of the theoretical item response curves against the reported scores and the WML estimates, report the MLE, WML, and various EAP averages nationally, and delineated the weighted percentage of students with missing MLE estimates for each TIMSS test booklet. In mathematics-focused books, the percentage of students not receiving valid MLE scores was trivial. However, the percentage of students without scores became non-trivial for science-focused books with relatively few mathematics items. Results showed that significant model misfits occurred on Indonesian students with atypical response patterns, especially for the low-scored students with no apparent pseudo-guessing behaviour.

TIMSS 2003, item response theory, estimation model, Indonesia, mathematics performance

The International Association for the Evaluation of Educational Achievement (IEA) has conducted a series of international assessment studies, one of which is called the Trends in International Mathematics and Science Studies (TIMSS). TIMSS employs an IRT-based plausible value methodology for scoring the assessment result. This procedure enables ability estimates to be produced for each student providing at least some achievement or contextual information is available.

In this study, Indonesian mathematics data from TIMSS 2003 were used to investigate the importance of assessment data in the student plausible values. When compared to many other participating countries, the Indonesian students performed poorly in TIMSS 1999 and

2003 (OECD, 2004; Mullis et al., 2005). After these assessments, the Indonesian government introduced a new national education curriculum called “competency based curriculum” which was a distinct reformed curriculum (Southeast Asia Ministers of Education Organisation (SAMEO), 2003). Moreover, the Indonesian National Education Department even introduced the implementation of International Curriculum parallel to the current national curriculum to several public schools few years ago in order to produce graduates with international quality comparable with those graduated in the OECD countries (Direktorat Jenderal Manajemen Pendidikan Dasar dan Menengah (Directorate General of Primary and Secondary School Management), 2008). Although a direct link between the international assessment results and the development of the reformed curriculum has not been known, there is a possibility that the TIMSS results may have provoked the Indonesian Government to rethink and possibly, redesign its national education development and curricula. However, to date no studies have been conducted to verify whether the Indonesian data fits these international assessment models.

In TIMSS 2003, the assessment framework for mathematics in the eighth-grade was organised into content and cognitive domains. The content domains included number, algebra, measurement, geometry, and data; whilst the cognitive domains measured knowledge on facts and procedure, concept usage, routine problem-solving, and reasoning skill. Taking into account time limitation for each student to do the test, while ensuring a thorough assessment of the target domains, the items were grouped into 14 blocks and distributed across 12 booklets. Each booklet was then rotated among the participating students. Similar configuration was performed on the science items. As delineated in Exhibit 2.16 of the TIMSS 2003 Technical Report, the mathematics blocks were labeled as M01 through M14, while the science blocks were labeled as S01 through S14 (Martin et al., 2004, p.53). Each booklet contained six blocks of items with three blocks on each part. One booklet took 90 minutes to complete and was administered in two 45-minute parts with a break in between. Three types of item formats were incorporated in mathematics: multiple choice, short-answer and extended response. The multiple-choice and short-answer items were scored dichotomously (1 if correct, 0 otherwise), while the extended response was scored 0,1, and 2 using a partial credit scoring system. The not-reached items located at any block-position were considered as incorrect responses.

The current research intends to provide insights into model-fit in TIMSS by focusing on the calculation of ability (θ). PARSCALE, the IRT program used in TIMSS, uses three parameter estimators, namely MLE, WML, and EAP. Firstly, the maximum likelihood estimation (MLE) procedure aims to identify an unobserved variable by maximising the likelihood of obtaining a particular response pattern from data that were actually observed. In terms of a student assessment, this technique intends to find the unknown θ from a set of student responses to test-items so that the probability of obtaining a particular response pattern is maximised. Introduced by Warm (1989), the weighted maximum likelihood (WML) method aims to reduce the bias from the MLE procedure by incorporating a test-information function as the weighting function. This WML is not a Bayesian estimator since the weighting function used is actually the reciprocal of the standard error of the maximum likelihood function. Meanwhile, the EAP method is a Bayesian statistics in which combines the information from the students’ responses contained in the likelihood function and that of the students’ background represented in the prior distribution function in order to make an inference about the θ-estimate. The prior ability distribution is usually expressed as a normal distribution. When the number of students is large, the variance tends to be large, and thus, the shape of the normal curve flattens. In this case, the prior distribution will not give a substantial contribution to the final product as the likelihood function will dominate the result. On the other hand, when the variance is small, the prior ability distribution will take shape as a thin and peaked-curve, said to be informative, and will tend to dominate the posterior distribution which may result in a biased estimate. Hence, the prior distribution function plays a crucial role in EAP method.

A comparison among these θ-estimation techniques is used to answer two research questions.

1. How well does the Indonesian data fit the international assessment models evaluated using (a) PV and (b) WML methodology?

For a large number of test-items in mathematics, their theoretical item response curves generated using the TIMSS 2003 item parameters was compared against the empirical data based on (a) the θ-metrics transformed from the reported plausible values and (b) the WML θ-estimates. Here, the WML estimates are chosen to generate the empirical plots because this particular parameter estimator would tend to produce less biased estimates than the MLE (Warm, 1989; Kim & Nicewander, 1993). Results indicated that a considerable model misfit had occurred on most items, regardless of the position of such items in the design of item-blocks as shown in Figure 1 through Figure 4.
In Figure 1, the plot of the probability of correctly responding the plot of the probability of correctly responding to a multiple-choice item coded as M022010 and its corresponding “free” item, i.e. MF22010. This particular item appears in Block M04 in the second position of Booklet 3 and in the first position of Booklet 4, whereas the corresponding “free” item is contained in the third position of Booklet 9, as indicated in Table 1. Here, the horizontal x-axis represents the θ-scale, whereas the vertical left and right y-axes denote the probability of a correct response. The smooth curve with diamond markers illustrates the theoretical curve based on the given item parameters, while the empirical data are symbolised by circles of which size indicates the proportions of the correct response estimated at the corresponding θ-value. It is demonstrated in Figure 1(a) and 1(b) that the empirical data generated from the published plausible value did not show a good fit with the 3-PL model being used. Besides, when the WML θs were used in generating Figure 1(c) and 1(d), similar misfit was detected in both plots. A worse misfit occurred for the “free” item which was allocated in the last block of the first half of Booklet 9. This evidence suggests that many students might not have responded, nor completed such a item resulting in many low scorers. This could also be due to the positioning effect of the particular items and the biased item parameters. In the case of the “free” item (MF22010), the empirical data suggest that the low scorers did not seem to show a pseudo-guessing behaviour. Since the respective item was represented by a 3-PL model that allowed pseudo-guessing to occur and of which item parameters were calibrated from participating countries wherein pseudo-guessing behaviour in a test was encouraged, the ability of Indonesian students who did not display such behaviour could be estimated incorrectly.
Figure 2 presents model fit plots of a short-answer type of item (M022148) and its “free” item version on which the 2-PL model was applied. No pseudo-guessing behaviour would be expected for this type of item because the student had to provide a direct short answer. This item was included in M03 and located in the second position in Booklet 2, the first position in Booklet 3, and the third position in Booklet 10. Meanwhile, the model fit plots given in Figure 3 also indicate that the partial credit model applied to item M22234B and MF22234B could not correspond to the Indonesian data well.

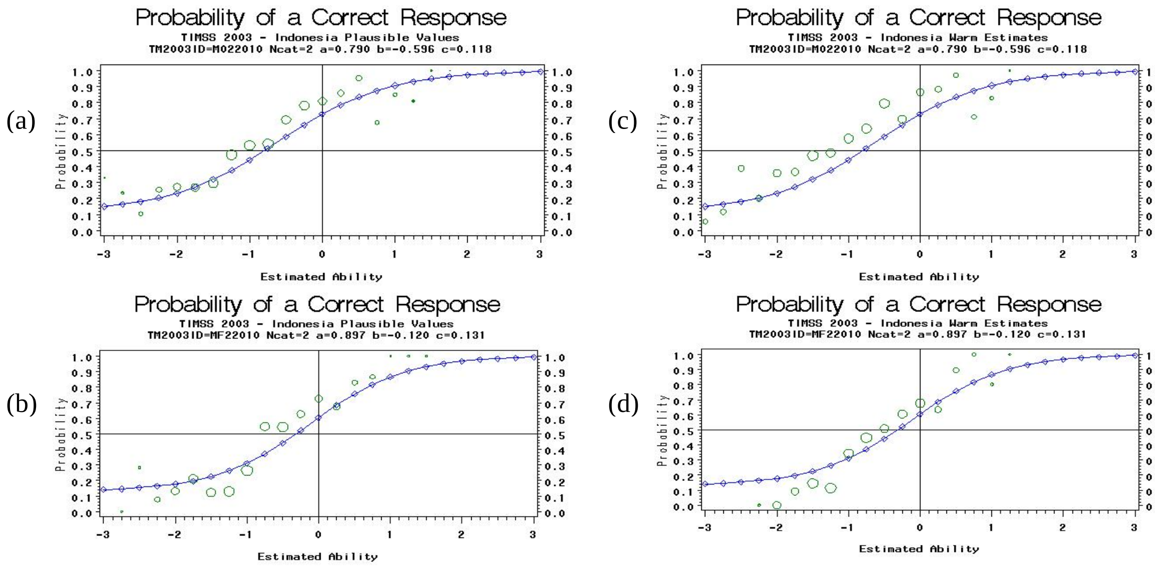

Figure 1. Examples of item response curves for item M022010 and MF22010 generated with the published TIMSS 2003 item parameters using the reported plausible values ((a) and (b)) and the WML θ-estimates ((c) and (d)), respectively.

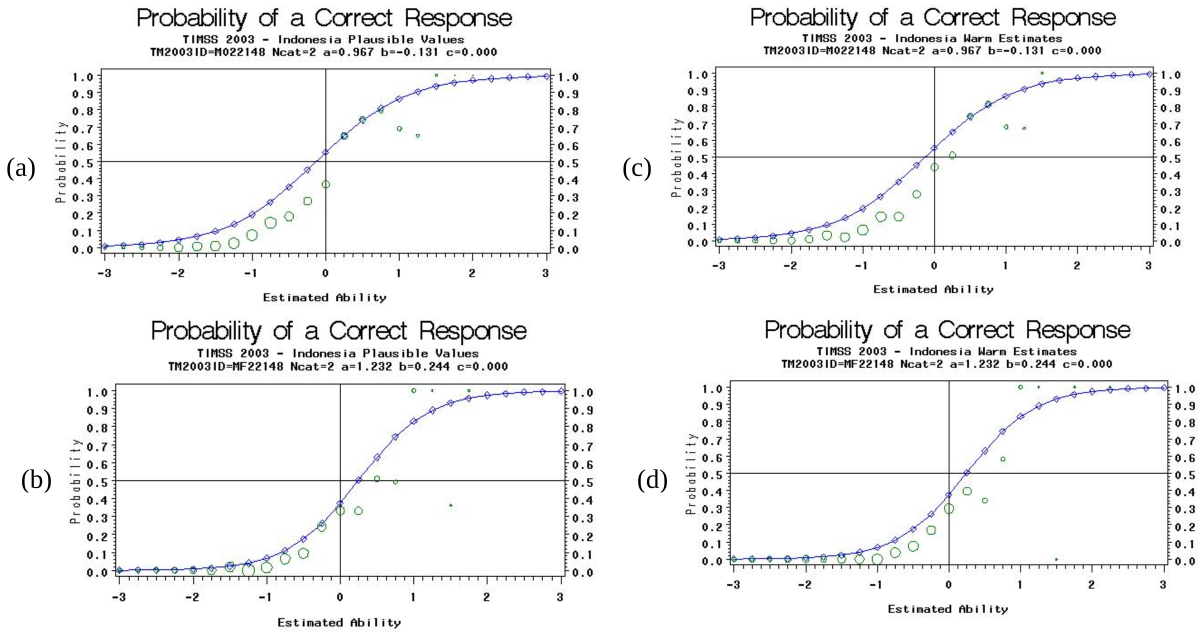

Figure 2. Examples of item response curves for item M022148 and MF22148 generated with the published TIMSS 2003 item parameters using the reported plausible values ((a) and (b)) and the WML θ-estimates ((c) and (d)), respectively.

The last example of model misfit is also shown for an M01 item (M012037) as given in Figure 4. This particular item is located in the first position of Booklet 1 and second position of Booklet 6, indicating that a positioning effect was less likely to occur. The empirical plot generated from the PVs and the WML behaved in a similar manner, deviating from the assumed theoretical model.

The apparent misfits shown in Figures 1 through 4 were largely illustrated by the plots of the “free” items, which were possibly due to a “no pseudo-guessing” behaviour among the Indonesian students who could not respond to such items. The ability of these students, who were more likely to be the poor scorers, was inappropriately estimated since the 3-PL model used in TIMSS included the pseudo-guessing parameter.

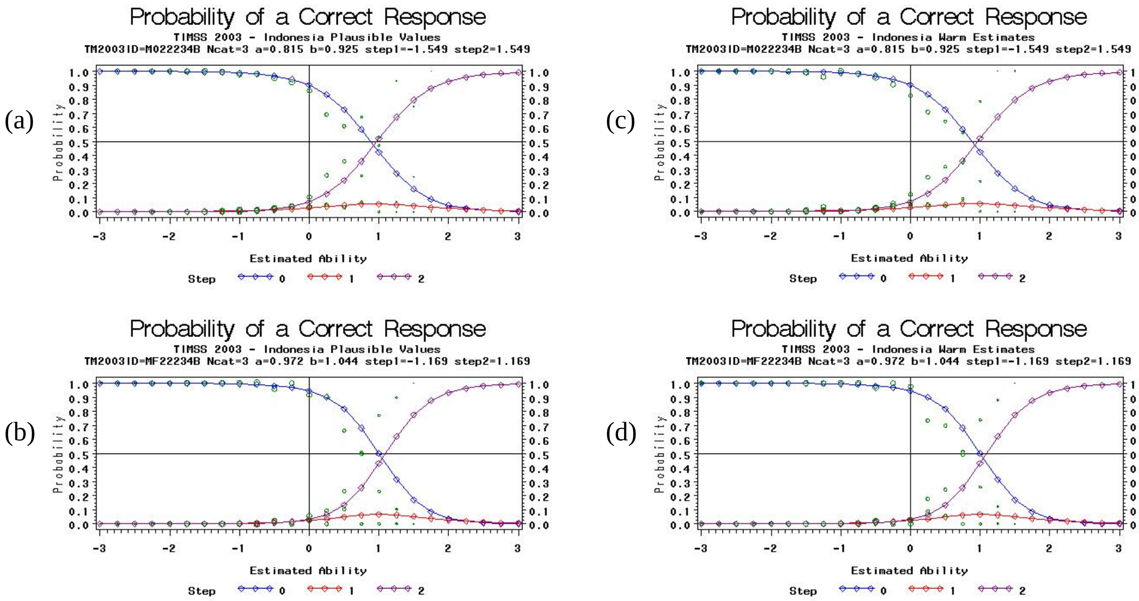

Figure 3. Examples of item response curves for item M022234B and MF22234B generated with the published TIMSS 2003 item parameters using the reported plausible values ((a) and (b)) and the WML θ-estimates ((c) and (d)), respectively.

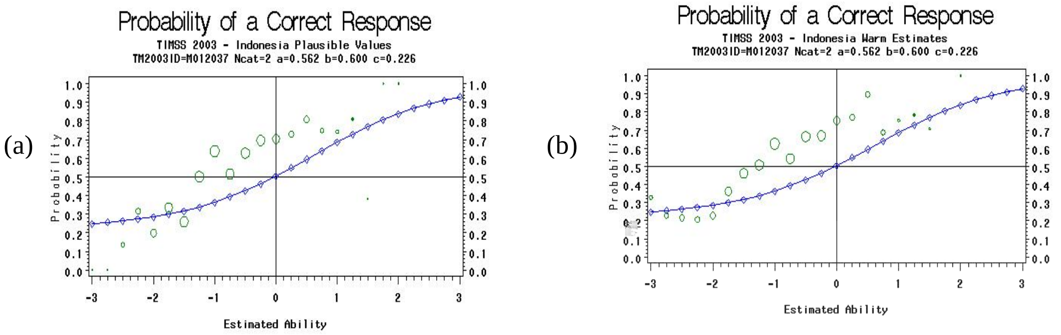

Figure 4. Examples of item response curves for item M012037 based on the TIMSS 2003 item parameters with (a) the θ-metrics of the reported plausible values and (b) the WML θ-estimates.

2. How well do the ability estimates produced by three different parameter estimation methods, i.e. MLE, WML, and EAP methods, compare with the reported ability estimates of TIMSS 2003?

(a) What effect do the Bayesian priors in the EAP method have on the ability estimates?

(b) When does the MLE estimate fail? What can be done if it fails?

(c) What are the effects on the mean estimates when EAP measures are substituted into the missing MLE?

This study found that some differences did occur on the θ-estimates produced by the different parameter estimators some of which produced incomparable estimates with the reported national average ($\bar{\theta} \approx -0.9$) as demonstrated in Table 2. The MLE and WML estimators produced lower estimates than those transformed from the reported plausible values (PV θ-metrics). The MLE and WML estimators produced mostly lower estimates than those transformed from the reported plausible values (PV θ-metrics), especially for students who did Booklet 7 through 12. In these booklets, since the number of math items were approximately half of those in the earlier booklets and their positions within the item-block design prompted not-reached responses, the estimates for the corresponding item parameter values would be less accurate resulting even more biased MLE and WML θs.

Each of the tested Bayesian normal priors, i.e. a flat prior, $N(-1,0.5)$, $N(-1,1)$ and $N(-1,2)$, gave a significant impact on the difference of the EAP outcomes since its informative or non-informative characteristics determined its contribution to the final EAP θ-estimates. When a Bayesian prior is very informative indicated by having a small variance in the prior, the resulting θ-estimates were more comparable with the θ-metrics. In this case, the contribution of the item responses, depicted in the likelihood part, would be overridden giving a biased EAP θ-estimate. On the other hand, when a Bayesian prior was non-informative, its impact could be overridden by the likelihood part instead, so that the final θ-estimate was directly obtained from the item responses.

It was noted that for some students their MLE θ-estimates were undefined, summing up to approximately 6% to 23% on each booklet (see Table 3). There were some occasions that the WML θ-estimates were also missing. A reasonable justification for this problem to occur was that the commonly used Newton-Raphson (NR) procedure had failed to obtain a unique maximum solution of the likelihood functions in MLE/WML for cases either with multiple maxima or with extremely low or high scores (Kreyszig, 1988; Yen et al., 1991).

Table 2. Means of θ-estimates on mathematics items calculated by different parameter estimators.

Booklet	EAP N(-1,0.5)	EAP N(-1,1)	EAP N(-1,2)	EAP Flat	MLE	WML	theta_ m1	theta_ m2	theta_ m3	theta_ m4	theta_ m5
1	-0.9495	-0.9495	-0.9495	-1.2759	-1.0454	-1.0424	-0.8626	-0.8721	-0.8641	-0.8607	-0.8651
	(0.046)	(0.046)	(0.046)	(0.066)	(0.061)	(0.060)	(0.053)	(0.051)	(0.055)	(0.054)	(0.052)
2	-0.9102	-0.9621	-1.0135	-1.0990	-0.8686	-0.8656	-0.9203	-0.9109	-0.9253	-0.9041	-0.9119
	(0.045)	(0.051)	(0.057)	(0.064)	(0.049)	(0.052)	(0.051)	(0.053)	(0.055)	(0.053)	(0.052)
3	-1.0263	-1.1113	-1.1887	-1.3047	-1.0341	-1.0940	-0.9097	-0.9344	-0.9190	-0.9220	-0.9271
	(0.047)	(0.054)	(0.060)	(0.069)	(0.061)	(0.060)	(0.056)	(0.058)	(0.057)	(0.055)	(0.055)
4	-1.1390	-1.2326	-1.3106	-1.4183	-1.4153	-1.2852	-0.8216	-0.8292	-0.8380	-0.8041	-0.8218
	(0.041)	(0.048)	(0.053)	(0.059)	(0.057)	(0.059)	(0.049)	(0.048)	(0.047)	(0.044)	(0.046)
5	-0.9867	-1.0503	-1.1094	-1.2008	-0.9753	-0.9953	-0.8292	-0.8422	-0.8672	-0.8282	-0.8330
	(0.045)	(0.052)	(0.057)	(0.065)	(0.059)	(0.057)	(0.059)	(0.056)	(0.054)	(0.054)	(0.052)
6	-1.0506	-1.1309	-1.2049	-1.3176	-1.0704	-1.1388	-0.8906	-0.8931	-0.8944	-0.8679	-0.8864
	(0.050)	(0.059)	(0.067)	(0.078)	(0.062)	(0.079)	(0.063)	(0.063)	(0.056)	(0.061)	(0.064)
7	-0.9098	-0.9589	-1.0137	-1.1093	-0.8154	-0.8364	-0.8804	-0.8881	-0.8607	-0.8600	-0.8976
	(0.041)	(0.049)	(0.055)	(0.064)	(0.053)	(0.052)	(0.055)	(0.053)	(0.049)	(0.056)	(0.054)
8	-0.9401	-1.0150	-1.0945	-1.2252	-0.9708	-0.9124	-0.8644	-0.8298	-0.8508	-0.8146	-0.8305
	(0.043)	(0.052)	(0.060)	(0.072)	(0.066)	(0.061)	(0.061)	(0.061)	(0.061)	(0.055)	(0.060)
9	-0.9524	-1.0478	-1.1602	-1.3666	-0.9986	-0.8003	-0.8860	-0.8717	-0.8973	-0.8947	-0.8750
	(0.039)	(0.048)	(0.056)	(0.070)	(0.080)	(0.046)	(0.051)	(0.047)	(0.051)	(0.050)	(0.048)
10	-0.9598	-1.0600	-1.1755	-1.3812	-0.6748	-0.8691	-0.8956	-0.8625	-0.9057	-0.8852	-0.8998
	(0.041)	(0.050)	(0.059)	(0.072)	(0.052)	(0.058)	(0.064)	(0.056)	(0.061)	(0.061)	(0.057)
11	-0.9480	-1.0367	-1.1367	-1.3072	-0.6492	-0.9250	-0.8641	-0.8834	-0.8616	-0.9012	-0.8673
	(0.039)	(0.048)	(0.056)	(0.069)	(0.045)	(0.061)	(0.052)	(0.055)	(0.055)	(0.061)	(0.055)
12	-0.8049	-0.8612	-0.9302	-1.0549	-0.6467	-0.7009	-0.7999	-0.8076	-0.8055	-0.7666	-0.7625
	(0.046)	(0.054)	(0.062)	(0.074)	(0.068)	(0.058)	(0.062)	(0.057)	(0.064)	(0.058)	(0.060)

Note: The parentheses denote the standard error of the θ-estimates, while theta_m1 through theta_m5 are the θ-metrics transformed from the associated five plausible values reported in TIMSS 2003

Table 3. Weighted percentage of Indonesian students whose MLE θ-estimates were undefined/missing in TIMSS 2003 with the standard error of θ-estimates in brackets.

Booklet	N	Weighted N	Weighted percentage of Students with missing MLE θ-estimate	
1	486	485.502	7.28	(0.014)
2	488	490.213	6.27	(0.015)
3	480	479.702	10.01	(0.015)
4	482	482.597	4.02	(0.010)
5	480	480.587	7.41	(0.015)
6	485	483.788	6.71	(0.014)
7	478	479.876	7.97	(0.018)
8	481	478.849	9.36	(0.022)
9	476	474.906	18.37	(0.022)
10	473	473.346	23.23	(0.024)
11	479	476.863	22.14	(0.029)
12	474	473.318	12.04	(0.020)
Total	5762	5759.547		

In addition, the positioning effect of the mathematics block-design made the problem worse, particularly with the “free” items since they prompted not-reached responses. As a fewer number of items was used in calibrating these item parameters, the resulting θ-estimates would contain some bias of order $O(n^{-1})$. When using the NR iteration procedure, the numerical error of these “free” item parameters would then propagate giving an even more biased MLE/WML θ (Warm, 1989; Kim & Nicewander, 1993).

Meanwhile, the EAP flat yielded the highest standard error of the mean estimates in all booklets as illustrated in Table 2, followed by the MLE or the WML estimators. The WML θ produced a slightly smaller standard error of the mean θ-estimates than the MLE θs, corroborating a past study that the WML estimator would yield a less biased results when compared to MLE (Warm, 1989; Kim & Nicewander, 1993).

Drawing an advantage of using EAP as a parameter estimator from these findings showed that EAP would always be able to produce a θ-estimate regardless of the characteristics of its prior or the item responses. Since the EAP estimates compared closer to the θ-metrics as the variance of the priors got smaller than those computed by the MLE and WML methods, these results then showed that the reported scores seemed to be derived largely from the prior distributions, not from the students’ actual responses.

When the MLE method failed to produce a θ-estimate, a solution method was proposed by substituting the corresponding EAP estimate to the missing value. A number of substitutions were made, each of which used different priors and produced lower means in comparison with the θ-metrics. The mean MLE θ-estimates on the mathematics items for each booklet were then recalculated after substituting the missing MLE estimates with estimates obtained from using the following EAP estimators: the EAP flat (labelled as *mle1*), the EAP with a prior of $N(-1,1)$ (labelled as *mle2a*), the EAP with a prior of $N(-1,2)$ (labelled as *mle2b*), and the EAP with a prior of $N(-1,0.5)$ (labelled as *mle2c*) as illustrated in Table 3. It can be observed from the comparison plots for Booklet 1 through 12 in Table 3 that the substitution method was more likely to produce lower θ-estimates than the estimates initially produced by the MLE method itself and those from the θ-metrics of the reported PVs. This could be due to the fact that before substitution, students who were mostly the poor scorers with aberrant response patterns, were excluded from the calculation of the mean so that the non-substituted datasets could yield higher means. The substitution of EAP θ-estimates with different priors into the MLE estimates has seemed to introduce an even larger variability to the datasets, and thus yielded standard errors still in the range of 0.04 – 0.08. Comparable with the discussion in the preceding section, the substitution using EAP flat (*mle1*) gave the lowest θ-estimates and the largest standard error in all booklets, followed by those substituted from the EAP with a prior of $N(-1,2)$ (*mle2b*). These results had been anticipated since both EAP methods utilised non-informative priors with a high degree of variability. The corresponding standard errors were mostly larger than those of the MLE's, particularly with those substituted by the EAP flat θ-estimates. This solution, however, might provide an alternative approach in overcoming the failure of the MLE estimator and hence, the final imputed score produced would better represent the ability of students with aberrant item responses and conditions. These results had been anticipated since both EAP methods utilised non-informative priors with a high degree of variability.

Table 3. Mathematics mean θ-estimates with the corresponding standard errors for each booklet produced by the MLE, WML and the plausible value methodology as well as from the substitution of the undefined MLE estimates.

Booklet	MLE	WML	mle1	mle2a	mle2b	mle2c	theta_ m1	theta_ m2	theta_ m3	theta_ m4	theta_ m5
1	-1.0454	-1.0424	-1.1996	-1.1240	-1.1240	-1.1240	-0.8626	-0.8721	-0.8641	-0.8607	-0.8651
	(.061)	(.06)	(.07)	(.062)	(.062)	(.062)	(.053)	(.051)	(.055)	(.054)	(.052)
2	-0.8686	-0.8656	-0.9861	-0.9505	-0.9658	-0.9332	-0.9203	-0.9109	-0.9253	-0.9041	-0.9119
	(.049)	(.052)	(.062)	(.056)	(.058)	(.054)	(.051)	(.053)	(.055)	(.053)	(.052)
3	-1.0341	-1.0940	-1.2072	-1.1439	-1.1721	-1.1113	-0.9097	-0.9344	-0.9190	-0.9220	-0.9271
	(.061)	(.06)	(.068)	(.062)	(.064)	(.06)	(.056)	(.058)	(.057)	(.055)	(.055)
4	-1.4153	-1.2852	-1.4856	-1.4614	-1.4732	-1.4454	-0.8216	-0.8292	-0.8380	-0.8041	-0.8218
	(.057)	(.059)	(.063)	(.06)	(.061)	(.058)	(.049)	(.048)	(.047)	(.044)	(.046)
5	-0.9753	-0.9953	-1.1128	-1.0712	-1.0898	-1.0491	-0.8292	-0.8422	-0.8672	-0.8282	-0.8330
	(.059)	(.057)	(.066)	(.061)	(.063)	(.059)	(.059)	(.056)	(.054)	(.054)	(.052)
6	-1.0704	-1.1388	-1.2135	-1.1662	-1.1877	-1.1407	-0.8906	-0.8931	-0.8944	-0.8679	-0.8864
	(.062)	(.079)	(.077)	(.069)	(.073)	(.066)	(.063)	(.063)	(.056)	(.061)	(.064)
7	-0.8154	-0.8364	-0.9904	-0.9310	-0.9559	-0.9042	-0.8804	-0.8881	-0.8607	-0.8600	-0.8976
	(.053)	(.052)	(.064)	(.056)	(.059)	(.054)	(.055)	(.053)	(.049)	(.056)	(.054)
8	-0.9708	-0.9124	-1.1669	-1.0947	-1.1258	-1.0606	-0.8644	-0.8298	-0.8508	-0.8146	-0.8305
	(.066)	(.061)	(.081)	(.071)	(.075)	(.066)	(.061)	(.061)	(.061)	(.055)	(.06)
9	-0.9986	-0.8003	-1.2847	-1.1460	-1.1992	-1.0949	-0.8860	-0.8717	-0.8973	-0.8947	-0.8750
	(.08)	(.046)	(.08)	(.072)	(.075)	(.07)	(.051)	(.047)	(.051)	(.05)	(.048)
10	-0.6748	-0.8691	-1.1753	-0.9816	-1.0583	-0.9065	-0.8956	-0.8625	-0.9057	-0.8852	-0.8998
	(.052)	(.058)	(.073)	(.057)	(.063)	(.052)	(.064)	(.056)	(.061)	(.061)	(.057)
11	-0.6492	-0.9250	-1.1481	-0.9678	-1.0410	-0.8938	-0.8641	-0.8834	-0.8616	-0.9012	-0.8673
	(.045)	(.061)	(.071)	(.052)	(.059)	(.045)	(.052)	(.055)	(.055)	(.061)	(.055)
12	-0.6467	-0.7009	-0.9154	-0.8221	-0.8606	-0.7824	-0.7999	-0.8076	-0.8055	-0.7666	-0.7625
	(.068)	(.058)	(.077)	(.069)	(.072)	(.066)	(.062)	(.057)	(.064)	(.058)	(.06)

Note: The parentheses denote the standard error of the θ-estimates, "mle1" uses a substitution of EAP Flat, "mle2a" uses EAP N(-1,1), "mle2b" uses EAP N(-1,2), "mle2c" uses EAP N(-1,0.5), while theta_m1 through theta_m5 are the θ-metrics transformed from the associated five plausible values reported in TIMSS 2003.

Conclusions and Recommendations

The choice of IRT models depend upon the characteristics of both the test-items and the test-takers (Baker, 1992; van der Linden & Hambleton, 1997). The more item parameters included in the model, the more sample size it requires to better represent the observed data. A quick and sufficient examination of model fit between the theoretical and the empirical IRF functions can be performed by comparing the theoretical IRF curve generated using the item parameter estimates and the empirical curve based on the ability estimates. It has been suggested that a decision made on how sufficiently close these two curves are depends on the objectives of the assessment, the degree of robustness, and the sample size (Harris, 1999). When a 3-PL model shows less information, then a model with fewer parameters should be used to avoid complexity in the parameter estimation process and model misfit (Yen et al., 1991). To date, no method, however, has been found to definitely determine whether or not an IRT model fit is entirely satisfactorily (Harris, 1999; Embretson & Reise, 2000).

Due to significant model misfits for the majority of the test items and biased resulting scores as described in the summary above, the TIMSS 2003 results on Indonesian students therefore need to be interpreted with caution. This study has shown that a considerable number of students could not have their MLE θ-estimates defined, whose aberrant item responses failed the MLE/WML methods to work as expected. Hence, some of the issues emerging from these findings relate specifically to the large variability in the student's item responses, the choice of the item response models, and the failure of the commonly used NR iteration technique in finding the maximum likelihood. These findings have also important implications for the development of an international or a large-scale assessment procedure that should take into account the large variability in the characteristics of the participating students since the results will often influence a country's specific national decision-making. As it was found that the commonly used assessment models for scoring students' achievement did not serve their purposes, an appropriate measurement model and solution techniques for any large-scale assessment imposed on students in developing countries such as Indonesia needs to be sought to account for the varied student characteristics.

With regard to the research findings summarised above, the following recommendations can be suggested. First of all, using a 2-PL or even 1-PL model may give more appropriate ability estimates considering that the 3-PL model employed for the multiple-choice items in TIMSS 2003 did not seem to represent the Indonesian students' responses. Since no pseudo-guessing behaviour was apparent, especially the low scorers, item response models with no pseudo-guessing parameters may better represent the characteristics of the population being investigated.

Secondly, a better estimation procedure, including a more robust numerical technique to solve for the maximum likelihood function, is required since there was some evidence that the student background information and the ability estimate from MLE used for generating the imputation scores in TIMSS 2003 were both weak. Given a non-informative background information, if the contribution of the likelihood of the item responses is weak, the final ability estimate produced will also be biased, far from the real latent trait. In order to avoid a weak likelihood function, an inclusion of several easy items on which all students will pass and several hard items on which all of them will fail may set the lower and upper limits of the ability scale better solvable by the NR technique. An alternative numerical method should still be sought for solving the MLE/WML functions as there is no guarantee that the commonly used Newton-Raphson procedure in finding the maximum likelihood will work all the time.

Finally, it is impractical, however, to keep adjusting the item response models or the parameter estimation methods in an international assessment in order to fully accommodate any particular characteristics of each participating country, unless a complex adaptive testing mechanism is employed. Consequently, further research should be directed to seek a more appropriate method for scoring or generating imputation scores that takes into account the large variability in student responses.

Reference:

Baker, F. B. (1992). *Item response theory. Parameter estimation techniques*. Reading, NY: Marcel Dekker.

Direktorat Jenderal Manajemen Pendidikan Dasar dan Menengah (Directorate General of Primary and Secondary School Management). (2008). 200 SMA dirintis jadi sekolah bertaraf internasional (Development of 200 Senior high schools to become internationally-standard school). Retrieved 1 July 2009, 1 July 2009, from http://mandikdasmen.aptisi3.org/index.php?option=com_content&task=view&id=60&Itemid=11

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. Mahwah, NJ: Lawrence Erlbaum.

Harris, D. (1999). Comparison of 1-, 2-, 3-parameter IRT models. In F. Brown (Ed.), *Instructional Topics in Educational Measurement* (Vol. Spring, pp. 35-41). Iowa, IA: National Council on Measurement in Education.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*(4), 587-599.

Kreyszig, E. (1988). *Advanced mathematics for engineering* (6th ed.). New York: Wiley.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 Technical Report*. Boston, MA: Boston College.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 International report on achievement in the mathematics cognitive domains. Findings from a developmental project.* Boston: TIMSS & PIRLS International Study Center.

OECD. (2004). *Learning for tomorrow's world. First result from PISA 2003*. Paris: Organisation for Economic Co-operation and Development (OECD).

Southeast Asia Ministers of Education Organisation (SAMEO). (2003). Indonesia. The educational process. Retrieved 20 June 2008, from http://www.seameo-innotech.org/resources/seameo_country/educ_data/indonesia/indonesia8.htm

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Yen, W. M., Burket, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika, 56*(1), 39-54.