THE EFFECT OF SAMPLE SIZE ON ITEM PARAMETER ESTIMATION FOR THE PARTIAL CREDIT MODEL

Qingping He & Christopher Wheadon

ABSTRACT

Item Response Theory (IRT) models have been widely used to analyse test data and develop IRT-based tests. An important requirement in applying IRT models is the stability and accuracy of model parameters. One of the major factors that affects the stability and accuracy of model parameters is the size of samples used to calibrate the items. Substantial research work has been undertaken in the past to study the effect of sample size on the estimation of IRT model parameters using simulations. Most of the simulation studies have focused on homogeneous item types and involved the use of model-generated response data. An important limitation of such simulation studies is that the assumptions of the IRT models are strictly met. However, data from operational tests do not normally strictly meet the model assumptions. The work reported in this paper investigates the effect of sample size on the stability and accuracy of model parameters of the Partial Credit Model (PCM) for a large data set generated from a high-stakes achievement test consisting of a mixture of dichotomous and polytomous items. Results from this study indicate that the level of stability and accuracy of model parameters is affected by the sample size, the number of categories of the items and the distribution of category scores within the items. The results obtained also suggest that the actual measurement errors associated with model parameters for polytomous items estimated from operational test data can be substantially higher than the model standard errors. It is furthermore suggested that the error introduced to true score equating using common items can be evaluated by a comparison with measurement errors inherent in the tests.

BACKGROUND

The accuracy and stability of Item Response Theory (IRT) model parameters have important implications in the analysis of test data and the development of IRT based tests. For example, Hambleton, Jones and Rogers (1993) found in their simulation study that the positive errors associated with the item discrimination index derived using the two parameter logistic model from samples of different sizes could produce tests with test information distributions which were substantially different from the true test information distribution. Chuah, Drasgow and Luecht (2006) have also found that the distribution of test information derived from parameters estimated using samples of different sizes can be substantially different from the information distribution derived using true parameters for a Computer Adaptive Sequential Test (CAST) implementing the three-parameter logistic model. However, they found little change in correlations between the ability estimate based on sample-estimated parameter values and that calculated using true parameter values among samples of different sizes. When test data are used to calibrate items using IRT software, both the model parameter estimates and associated measurement errors (model standard errors) are provided. The errors reflect the probabilistic nature of the IRT models and the degree to which the model fits the data. The probabilistic nature of the IRT models implies that sample size is an important factor that affects the accuracy and stability of the estimates of the model parameters.

Many simulation studies have been undertaken to investigate the effect of sample size on the estimation of IRT model parameters. For example, Wang and Chen (2005) have investigated how the item parameter recovery, standard error estimates and fit statistics are affected by sample size and test length for the Rasch model (Rasch, 1960) and the Rating Scale Model (RSM, see Andrich, 1979) using WINSTEPS (Linacre, 2006). De Ayala and Sava-Bolesta (1999) have studied the effect on item parameter estimation of the ratio of sample size to the number of model parameters and the latent trait distribution of the samples for the Nominal Response Model (NRM, see Bock, 1972). DeMars (2003) also studied the effect of sample size, test length, category size and sample ability distribution on parameter estimation for polytomous items with the NRM. Results from such studies generally indicate that the magnitude of the variation between sample estimates decreases with increasing sample size. The majority of the simulation studies focus on tests composed of homogeneous item types.

One of the limitations of using pure simulations to study the effect of sample size on IRT item parameter estimation is that the model assumptions are strictly met, which will seldom be true for operational test data. There have also been studies that use operational data to conduct simulation studies investigating the sample size effect on model parameter estimation. For example, Swaminathan et al (2003) have used the large test dataset generated from the Law School Admissions Council tests to study the effect of sample size on the accuracy of parameter estimation for the one-parameter, two-parameter and three-parameter IRT models. They have also developed procedures for improving parameter estimation. Stone and Yumoto (2004) have used sub-samples from a large dichotomous dataset from the Knox's Cube Test-Revised (Stone, 2002) to investigate the sample size effect for Rasch/IRT parameter estimation. Smith et al. (2008) have studied the sample size effect on the stability of model fit statistics of the PCM and the Rating Scale model, using samples drawn from a large dataset collected from medical surveys on cancer patients. Their results demonstrated that the model t-statistics were sensitive to sample size for polytomous data, while the mean square statistics remained relatively stable.

One important feature of the high-stakes tests provided by UK Awarding Bodies is the use of both dichotomous and polytomous items. Item response theory models have been increasingly used to analyse data from such high-stakes tests to ensure the comparability of standards between specifications and over time (e.g. Wheadon and Whitehouse, 2006; He and Wheadon, 2008). IRT models are particularly suitable for equating tests containing common items or involving common persons. In view of the structure of these high-stakes tests, the PCM would appear to be the most appropriate IRT model. However, the use of PCM would require the estimation for a large number of model parameters and reservations have been expressed on the use of category measures for test equating, due to uncertainties associated with their stability. Stabilised estimates of model parameters are required for test equating in order to minimise equating error. Although many studies have been conducted to investigate the effect of sample size on model parameter estimation, little attention has been given to the sample size effect on the stability of category measures for tests containing both dichotomous and polytomous items. The work reported in this paper investigates the effect of sample size on the stability and accuracy of model parameters of the PCM, using samples drawn from a large data set collected from a high-stakes achievement test administered by the Assessment and Qualifications Alliance (AQA) to students aged 16 in June 2007. The effect of sample size on test equating has also been investigated using a worked example.

METHODOLOGY

The Partial Credit Model

The IRT model used in the present study is the PCM, which is widely used for analysing polytomous items. The PCM represents an extension of the Rasch model for dichotomous items (see Rasch, 1960; Wright and Stone, 1979; Masters, 1982; Wright and Masters, 1982). In the PCM, polytomous items are characterised by ordered score categories (see Masters, 1982; Wright and Masters, 1982; Masters and Evans, 1986). In the PCM, the probability $P(\theta, x)$ of a person with ability θ, who scores *x* on a polytomous item with a maximum score *m*, can be expressed as:

$$P(\theta, x) = \frac{\exp \sum_{k=1}^{x} (\theta - \delta_k)}{1 + \sum_{x=1}^{m} \exp[\sum_{k=1}^{x} (\theta - \delta_k)]} \quad (1)$$

In Equation (1), δ_k is the *k*th threshold location of the item on the latent trait continuum, which is also referred to as the item step difficulty. In PCM, the person score x on a polytomous item represents the counts of the ordered categories that have been successfully undertaken. Masters (1984) and Masters and Evans (1986) have used the model to construct calibrated item banks for questions containing multiple parts which are scored dichotomously.

The AQA GCSE Mathematics B Specification

The test studied here for illustrative purposes is an externally set and marked examination taken by pupils aged 16 in the UK. Mathematics is chosen because it offers a large sample size for which item-level data is available and contained a combination of item-types. The data is taken from the June 2007 administration of module 3 (43003H). The results from this test can be used to support applications for further study beyond age 16 and for employment purposes. Two tiers of entry are available; a lower (foundation) tier samples less demanding content from the domain, while a higher tier samples the more demanding content. This study uses the results from the higher tier test. Table 1 illustrates the structure of the test.

Table 1 Maximum marks of questions in the test.

Item ID	Maximum Mark		Item ID	Maximum Mark
1	1		17	2
2	1		18*	2
3	3		19	1
4*	3		20*	2
5*	2		21	2
6*	2		22*	3
7	3		23*	2
8	2		24	2
9	1		25	2
10	1		26	1
11	2		27	2
12	1		28	2
13	2		29	5
14	3		30	2
15	2		31	2
16	3			

* link items between the tiers.

Item Calibration Using the Population Data

Responses from 49,120 students who took the higher tier test in a single administration were used to calibrate the items in the test using the PCM implemented in WINSTEPS. Table 2 displays the category measures and associated errors exported from WINSTEPS for the test. The category measures vary from -2.82 logits to 3.48 logits, with a mean of 0 logits (the calibration was centred on item difficulty). The errors of measurement are generally either 0.01 logits or 0.02 logits. Further analysis from WINSTEPS output indicated that categories of all the polytomous items are ordered. However, as is clear from Table 2, the majority of the polytomous items have disordered thresholds, suggesting that the distribution of the scores among the categories are not uniform for the items. Ideally items should have ordered thresholds, but they do not necessarily pose a threat to measurement.

Table 2 Category measures and associated errors of measurement, estimated using responses from the entire population of 49,120 students.

Item ID	Category	Measure	Error		Item ID	Category	Measure	Error
1	1	-1.10	0.01		17	1	2.49	0.01
2	1	-0.76	0.01			2	-1.81	0.01
3	1	0.24	0.02		18*	1	-1.07	0.01
	2	-2.41	0.02			2	-1.25	0.01
	3	-1.34	0.01		19	1	-0.95	0.01
4*	1	0.11	0.02		20*	1	1.72	0.01
	2	-0.56	0.01			2	-2.54	0.01
	3	-2.82	0.01		21	1	0.06	0.01
5*	1	0.98	0.01			2	-0.7	0.01
	2	-2.48	0.01		22*	1	0.57	0.01
6*	1	1.18	0.01			2	-0.68	0.01
	2	-2.68	0.01			3	-1.71	0.01
7	1	-0.13	0.01		23*	1	3.28	0.02
	2	3.48	0.01			2	-0.21	0.02
	3	-2.56	0.01		24	1	-0.59	0.02
8	1	-0.93	0.01			2	-2.51	0.01
	2	-0.03	0.01		25	1	-0.06	0.01
9	1	-0.33	0.01			2	1.47	0.01
10	1	0.69	0.01		26	1	0.50	0.01
11	1	1.80	0.01		27	1	2.35	0.01
	2	2.47	0.03			2	-1.06	0.01
12	1	-0.07	0.01		28	1	2.04	0.01
13	1	1.98	0.01			2	-1.26	0.01
	2	-1.96	0.01		29	1	2.45	0.01
14	1	1.47	0.01			2	-0.98	0.01
	2	-0.09	0.01			3	0.28	0.01
	3	-1.21	0.01			4	0.45	0.01
15	1	1.38	0.01			5	-0.42	0.02
	2	-0.47	0.01		30	1	2.52	0.01
16	1	2.94	0.01			2	-0.27	0.02
	2	1.43	0.02		31	1	0.67	0.01
	3	-1.48	0.02			2	1.98	0.02

* link items between the tiers.

Investigating the Sample Size Effect on Parameter Estimation

To study the effect of sample size on partial credit model parameter estimation, random samples with different sizes were drawn from the population using replacement sampling. Six sample sizes were investigated: 150, 300, 500, 1000, 2000 and 4000. For each sample size, 10 replicates (repeated samples) were produced. The values of the PCM parameters and the associated statistics estimated using the whole population were treated as the standard or true model parameter values. The differences between the parameter values estimated using individual samples and the model true values reflect the influence of sample size on sample estimates. The root mean square errors (RMSEs) have frequently been used to compare sample estimates with true values for model parameters. The RMSE σ_i for a parameter is the

square root of the average of the square of the difference between the sample estimate and the model true value over the replications in a sample size class:

$$\sigma_i = \sqrt{\frac{\sum_{k=1}^{K}(\delta_{i,k} - \delta_{i,true})^2}{K}} \quad (2)$$

where

i represents a specific category;
k is the kth replicate in a size class;
K is number of replicates in a sample class (10 in this study);
$\delta_{i,k}$ is sample estimate of category i from the kth replicate;
$\delta_{i,true}$ is the true model value for category i.

Because of the sample size effect on parameter estimation, the RMSEs will vary between sample sizes, and the magnitudes will reflect the degree of the departure of the sample estimates from the model true values. Evaluating what is an acceptable level of error is to some extent subjective and should be weighed against practical considerations; however, if the error due to sampling is comparable to the model standard error then the sampling effect is not substantial.

Investigating the Sample Size Effect on Test Equating Using Common Items

To study the sample size effect on test equating using common items, which is one of the most widely used test equating methods, another index, the root mean square error of equating (RMSEofEq) σ_{Eq}, is used for the linking/common items:

$$\sigma_{Eq} = \frac{1}{Ne}\sqrt{\frac{\sum_{k=1}^{K}(\sum_{j=1}^{Ne}(\delta_{j,k} - \delta_{j,true})^2}{K}} \quad (3)$$

In Equation (3), Ne is the total number of categories of the link items between two tests to be equated, and j is the jth link category of the link items. When the equating method between two tests suggested by Wright and Stone (1979) is used, σ_{Eq} will provide a measure for the errors associated with the equating constant introduced by sampling in equating using common items.

RESULTS

Model Parameter Estimation

As an example of the variation of sample parameter estimation, Figure 1 depicts the relationships between sample category estimates and true parameter values for samples with different sizes (the estimated values were derived using the first sample from each size class). The straight line in Figure 1 is the identity line (y=x). Figure 1 indicates that as sample size increases, the linear relationship between sample estimates and model true values becomes

stronger. When the sample size is 150, the differences between sample estimates and model true values vary from -1.08 logits to 0.81 logits. When the sample size is 300, the differences vary between -1.03 logits to 1.22 logits. When the sample size increases to 1000, the differences vary from -0.39 logits to 0.37 logits.

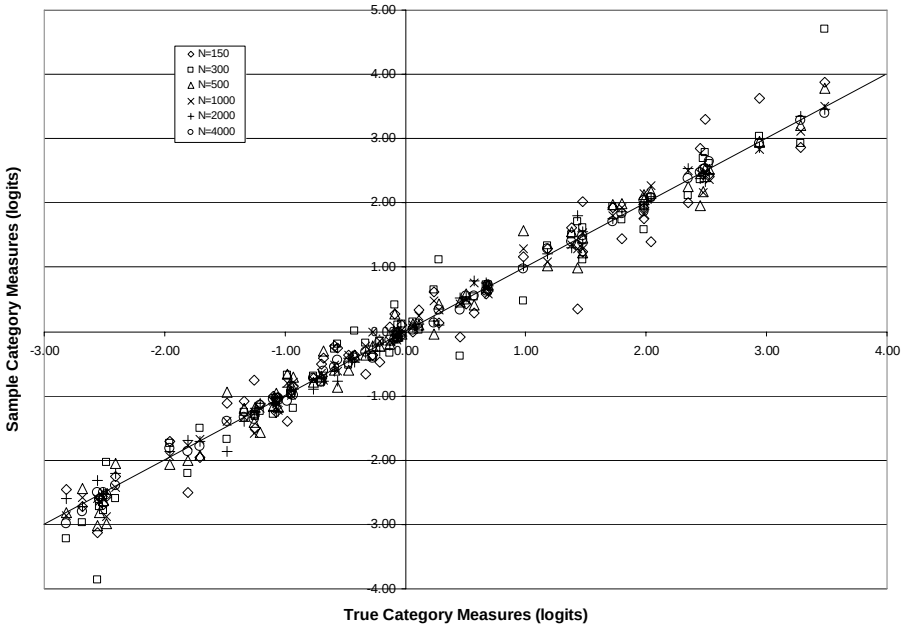

Figure 1 The relationships between sample parameter estimates and true model parameter values.

Although to some degree the differences between sample estimated values and the true values reflect the sample size effect on model parameter estimation, it is the variation of the differences between the replicates for a specific sample size class that is more useful, as one will not be able to obtain replicates in practical operations. The root mean square errors (RMSEs) reflect the variation of sample parameter estimates within replications for a specific size class. Figure 2 illustrates the distributions of the RMSEs for category measures for the different sample size classes (see Equation (2)). As is clear from Figure 2, the RMSEs generally decrease with increasing sample size. When the sample size is 150, the RMSEs for some categories are as high as 0.90 logits. When sample size increases to 1000, the RMSEs for the majority of the categories are within 0.20 logits.

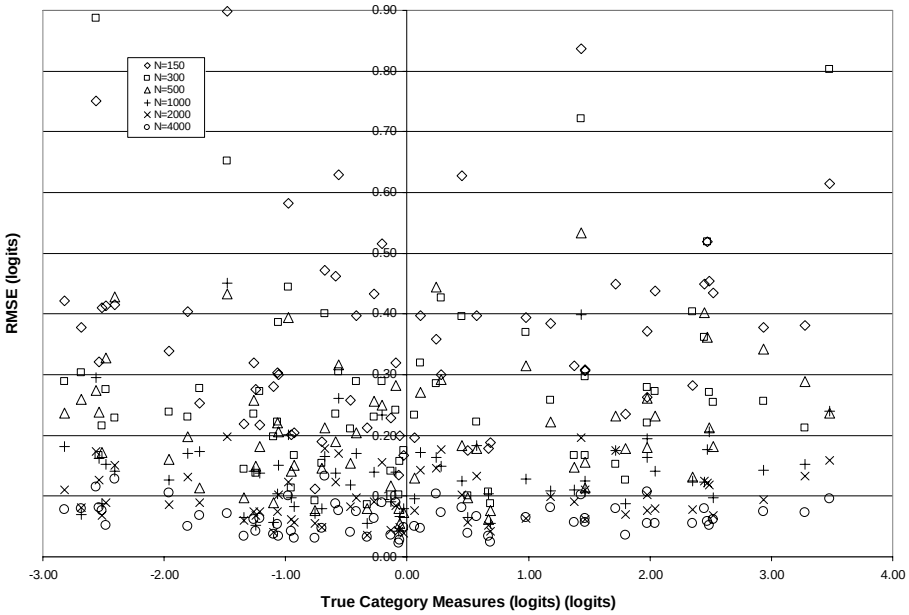

Figure 2 The distributions of RMSE for category measures.

To look at the impact of sample size on model parameter estimates for items with different numbers of categories, the items were further classified into four types, based on the number of their categories: items with two categories (1 mark dichotomous items), items with 3 categories (2 mark items), items with 4 categories (3 mark items) and items with 6 categories (the 5 mark item). There were no items with 5 categories or 4 marks in the test. For each sample size class, the RMSEs for category measures in each item type were averaged to produce the category mean RMSEs for the item type. Figure 3 depicts the mean RMSE for each item type. This represents a crude way to look at the effect of the number of categories in an item on parameter estimation. As is clear from Figure 3, the mean category RMSE generally decreases with increasing sample size. The mean RMSE for dichotomous items (two category items) is 0.19 logits when the sample size is 150 and 0.03 logits when the sample size is 4000. The mean RMSE for the item with six categories is 0.47 logits when the sample size is 150 and 0.08 logits when the sample size is 4000.

Within each sample size class, the mean RMSE for category measures generally increases with increasing number of categories in the items, except for the case of sample size of 1000 in which the mean RMSE for items with four categories is higher than that for the item with six categories. The mean RMSE for category measures for items with three categories is 0.37 logits, while that for items with five categories is 0.38 logits when the sample size is 300. The increase of mean RMSE for category measures with increasing number of categories in items for a specific sample size class reflects the fact that higher category items have more model parameters than lower category items and that a test taker will only respond to one of the categories of an item regardless the number of categories of the item. As a results, higher category items receive fewer responses for a specific category than lower category items. The score distribution between categories for an item will also affect the parameter estimation. The differences in RMSEs between different item types also decrease with increasing sample size. When the sample size is 4000, the category RMSEs for all item types are within 0.10 logits.

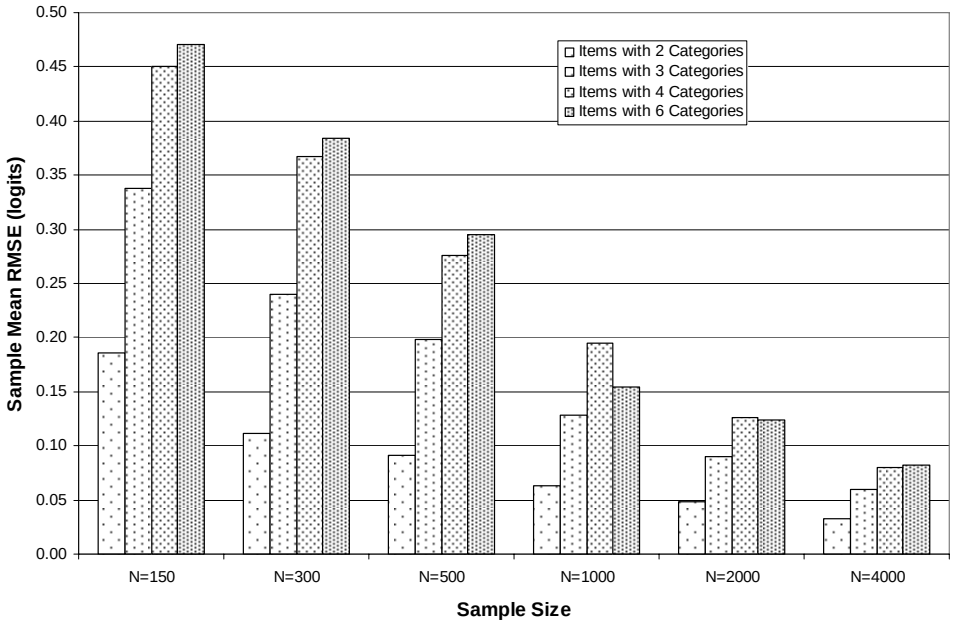

Figure 3 The distributions of the mean RMSE for category measures of individual item type.

Figure 4 shows the distribution of the measurement errors of the category step difficulties exported from WINSTEPS. The errors of measurement exported from an IRT analysis tool normally represent model standard error (see Wright, 1995), which decrease with increasing sample size. When the sample size is 150, the mean errors of measurement for category measures vary from 0.17 logits to 0.50 logits. When sample size increases to 500, the mean measurement errors vary from 0.09 logits to 0.25 logits. The largest measurement errors are associated with the third category measure (with a true model value of 2.47 logits) of a three-category item.

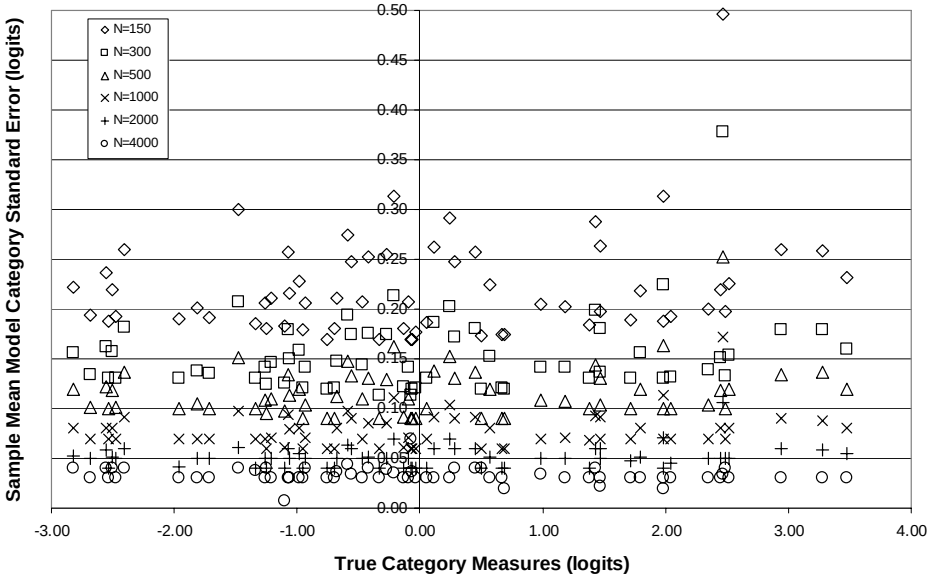

Figure 4 The distributions of sample mean model category standard errors.

Figure 5 further shows the mean model category measurement errors for specific item types in each sample size class. Within each class, the mean errors of measurement increase with increasing categories in the items. When the sample size is 150, the mean measurement errors are 0.17 logits for items with just two categories, 0.22 logits for items with three categories, 0.22 logits for items with 4 categories and 0.24 logits for items with 6 categories. As sample sizes increase, the differences between the mean category measurement errors within the same sample size class decrease. When sample size reaches 2000, the mean measurement errors for all items are lower than 0.06 logits.

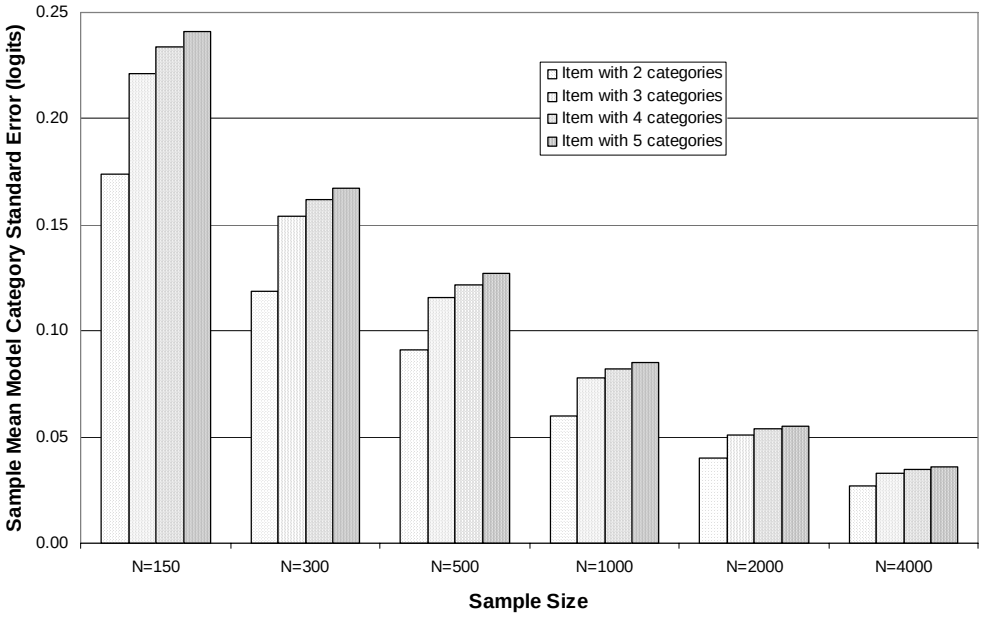

Figure 5 The distributions of category mean model measurement errors against item types.

The RMSEs discussed above reflect both the sampling effect and the probabilistic nature of the Partial Credit Model and may be viewed as the real standard errors of the model parameter estimates as suggested by Wright (1995). It is noted that the model standard error for an item exported from an IRT analysis software system does not take into account whether the item fits the IRT model. Wright (1995) has suggested that misfit items can increase the measurement error above the model standard error. To study how both the RMSEs and the model standard errors are affected by sample size, Figure 6 further depicts the ratio of RMSEs to the model standard errors for the different sample classes. If the data fits the model perfectly, it would be expected that the ratios be close to 1. It is clearly from Figure 6 that for the majority of the items, the ratios are between 1.00 and 3.00, indicating that the real measurement errors are substantially higher than the model standard errors. Figure 7 further illustrates the mean ratios of the RMSEs to the model standard errors for specific item types in each sample size class. Within each sample size class, the ratios increase with increasing number of categories in the items. For dichotomous items, the mean ratios are close to 1.

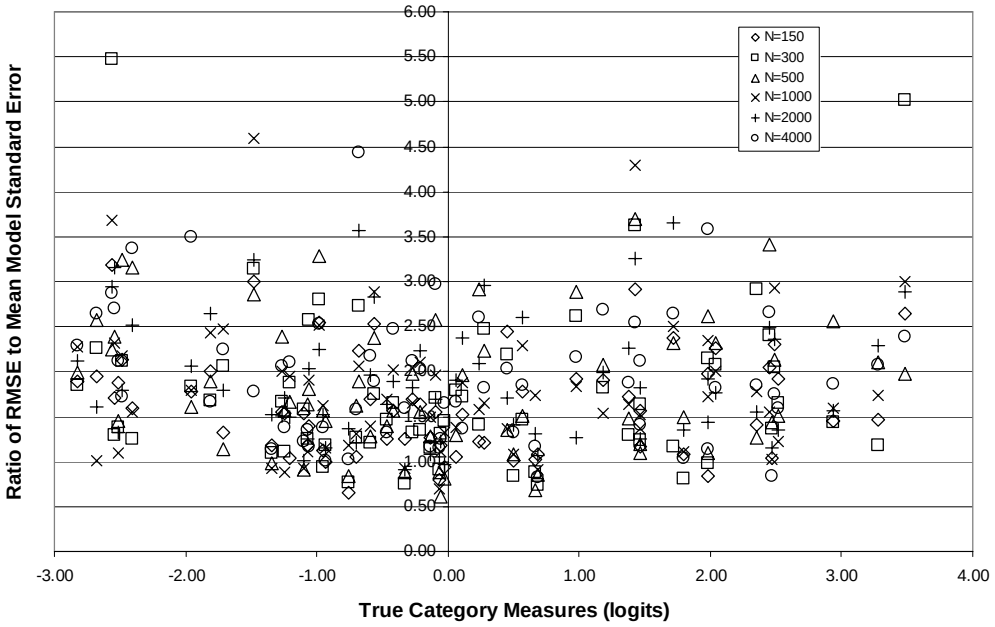

Figure 6 The distributions of the ratio of sample mean model category standard error to RMSE.

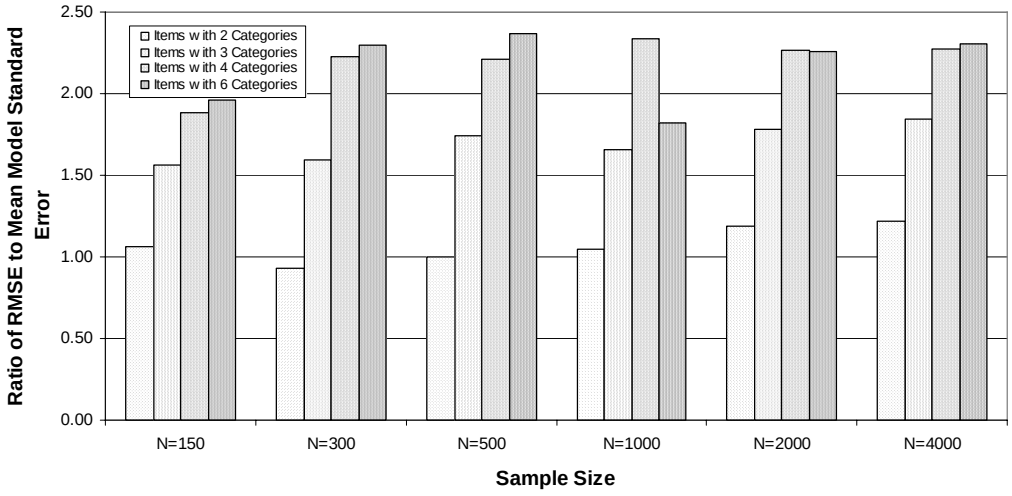

Figure 7 The distributions of the ratio of sample mean model category standard error to RMSE against item types.

Test Equating

Seven items in the test (accounting for 25% of the maximum mark of the test) analysed here are common to both the Foundation Tier test and the Higher Tier test. The sample size effect on equating can be measured using the root mean square error of equating (RMSEofEq) defined by Equation (3). Figure 8 shows the distribution of RMSEofEq against sample size. With seven questions (16 marks) common to both tiers, the magnitude of the equating error induced by sampling is quite small. This is because, although the RMSEs associated with individual

category measures could be large, the variation of the mean of the measures of the link items will be reduced by averaging over all link items. When the sample size is 150, the RMSEofEq is just over 0.06 logits. When the sample sizes are 300 and 500, the equating error is below 0.04 logits. When the sample size approaches 2000, the error is just above 0.01 logits.

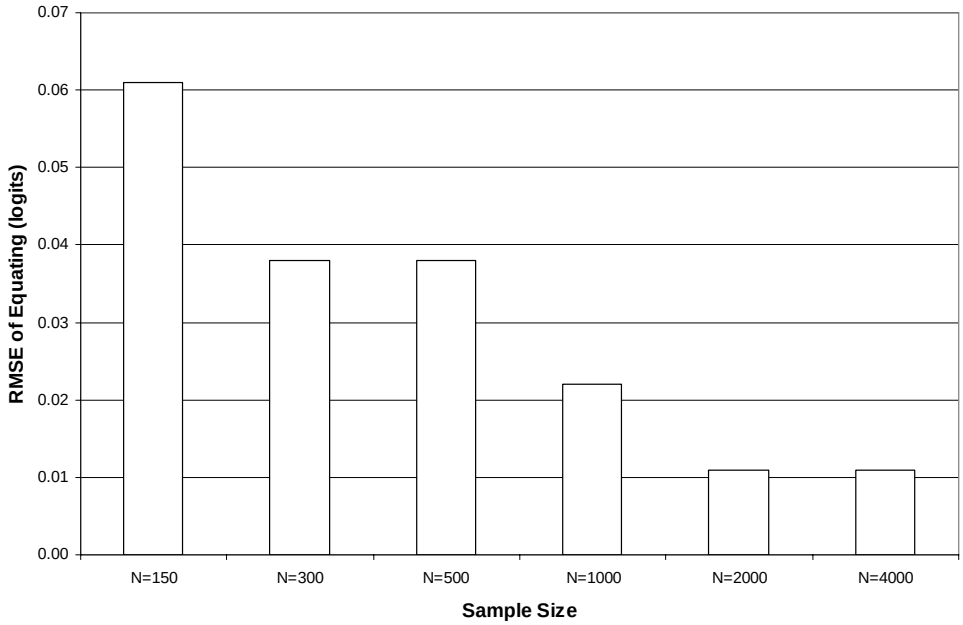

Figure 8 The distributions of root mean square error of equating.

One way of evaluating the acceptability of the error introduced to true score equating using common items is to compare it with measurement errors inherent in the tests. The following example, taken from a different equating scenario, demonstrates how this can be done. Using a post-test equating design, a group of 250 candidates was recruited to take an anchor test which sampled the content of two other tests. These two tests were shorter than that studied above, with a maximum mark of 36 and only two polytomous items with a maximum category score of 4. The purpose of the equating was to link the difficulty of these two tests. In this situation, the sample size is given, but the number of linking items can be varied. As the first test had already been administered to a large sample (N=20250), it was possible to take ten samples from this first administration to calculate the RMSEofEq according to Equation (3). Two scenarios were trialled: one in which the linking items were chosen in order from the start of the test and a second in which the linking items were chosen to best target the population ability (the mean item difficulty was closely matched to mean person ability). Only dichotomous items were chosen as link items. Figure 9 shows how the RMSEofEq decreases as the number of linking items is increased according to each selection method. Increasing the number of link items in each case yields diminishing returns; targeting the items to mean person ability results in a lower RMSEofEq than selecting them in sequence order. This is to be expected, as the parameters of well-targeted items can be calibrated with greater precision. Figure 10 shows the error introduced to true score equating by the shift represented by the RMSEofEq. This shift is slight compared to the standard error of raw scores. It would seem in this situation, therefore, that even with a low sample size and relatively few link items, the error introduced by equating is outweighed by the measurement error inherent in the tests themselves.

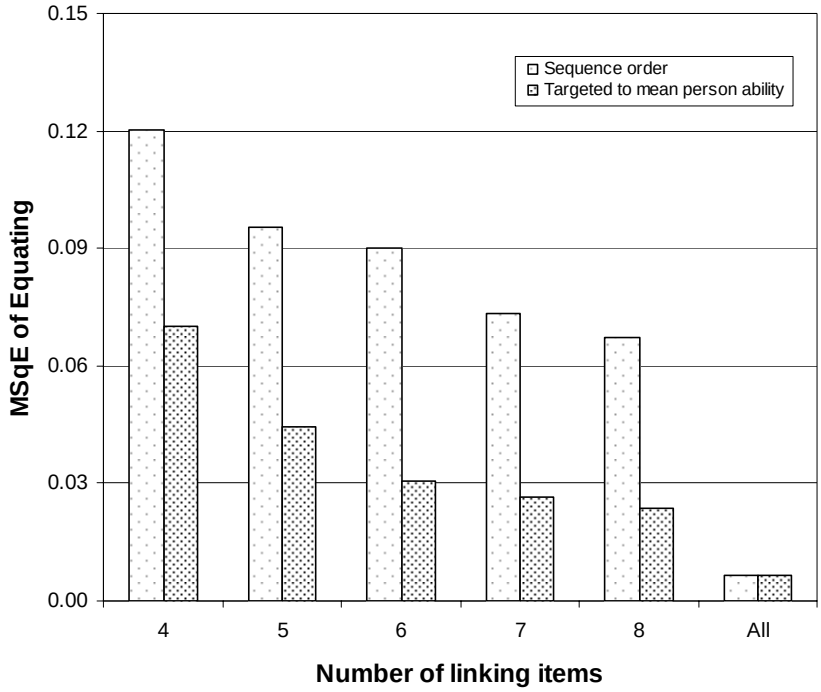

Figure 9 MSqE of equating according to the number of link items.

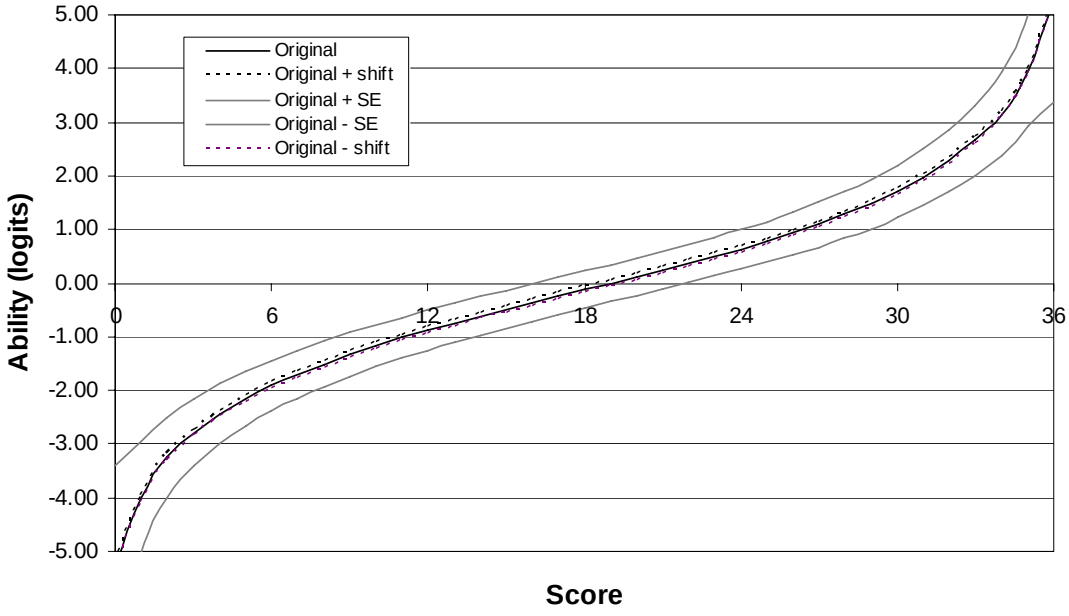

Figure 10: The impact of the MSqE of equating on true score equating.

DISCUSSION

As indicated by the work carried out by Hambleton, Jones and Rogers (1993) and the work by Swaminathan et al. (2003), estimation errors of IRT model parameter estimates due to sampling can have important implications in analysis of test data and the development of IRT-based tests. Efforts should be made to reduce estimation errors to a minimum.

Existing studies on sample size effect on IRT model parameter estimation have primarily focused on model parameter recovery using simulations to generate responses data. One of the limitations with simulations using model produced responses data is that the data used meet the model assumptions. As operational test data normally do not strictly meet the model assumptions, results from simulation investigations may not correctly reflect real situations. Although there have been studies using operational test data in simulations to investigate the sample size effect on IRT model parameter estimation, many used homogeneous item types. While this study used heterogeneous item types and operational test data, the results are specific to the test used and caution must be exercised in generalising from the findings reported here.

Nevertheless, the results obtained from this study demonstrate both the sample size and the item type have an important influence on model parameter estimation. The RMSEs for sample category measures decrease with increasing sample size. When sample size is fixed, the RMSEs generally increase with increasing number of categories in an item. Model parameter estimation is also affected by the score distribution between categories of the items. Although some researchers have suggested using the ratio of sample size to the number of model parameters as a guide to determine sample size (see De Ayala and Sava-Bolesta, 1999; DeMars, 2003), the structure of the test also needs to be taken into account. For example, for tests comprising homogeneous item types, the ratio of sample size to number of model parameters may be used as a good indicator for choosing sample size. In such a case, responses will be relatively evenly distributed between all categories. However, when a test is composed of heterogeneous item types (i.e. items with varying number of categories), a reasonable number of responses to categories of the items with the highest number of categories must be ensured in order to maintain the sampling error for all items with acceptable level. This study clearly shows that when sample size is fixed, the sampling errors associated with category measures for items with high numbers of categories are generally larger than those for items with low numbers of categories. It is also worth noting that the sample estimated RMSEs for polytomous items can be substantially higher than the model standard errors. Further work is needed to study the effect of misfit items on parameter estimation.

The size of samples required to estimate model parameters depends on the distribution of scores among categories within items and the degree of sampling errors that can be accepted. For the test data used here, when the sample size reaches 1000, the RMSEs for the category measures vary from 0.04 to 0.45 with an average of 0.14. When the sample size is 4000, the RMSEs vary from 0.02 to 0.13 with an average of 0.06. In this case, if an average of 0.06 for the RMSEs is required, then the sample size would need to be 4000. Results from this study also show that the equating error depends on both the number of link items and size of the sample used to derive parameter estimates. The test studied here had a maximum score of 64, and the link items account for 25% of the maximum mark. A sample of the size of 300 in this case would produce equating errors below 0.04 logits. Whether this level of error is acceptable is largely a

subjective decision; however, the final example shows that such equating errors can be negligible compared to the model standard errors of estimates of true scores.

Qingping He and Chris Wheadon
Research and Policy Analysis Department
AQA
December 2008

REFERENCES

Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika* **43**, 561-573.

Bock, R. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* **37**, 29-51.

Chuah, S., Drasgow, F. and Luecht, R. (2006) How big is big enough: Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education* **19**, 241-255.

De Ayala, R. and Sava-Bolesta, M. (1999) Item parameter recovery for the Nominal Response Model. *Applied Psychological Measurement* **23**, 3-19.

DeMars, C. (2003) Sample size and the recovery of Nominal Response Model item parameters. *Applied Psychological Measurement* **27**, 275-288.

He, Q. and Wheadon, C. (2008) Using the Rasch model to analyse dichotomous and polytomous items. *AQA Internal Report, RPA_08_QH_RP_017*.

Hambleton, R., Jones, R and Rogers, H. (1993) Influence of item parameter estimation errors in test development. *Journal of Educational Measurement* **30**, 143-155.

Linacre, J. M. (2006) WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com

Masters, G. (1982) A Rasch model for partial credit scoring. *Psychometrika* **47**, 149-174.

Masters, G. (1984). Constructing an item bank using partial scoring. *Journal of Educational Measurement*, 21, 19-31.

Masters, G. (1999). Partial credit model. *In Advances in measurement in educational research and assessment* (Ed. by G. Masters and J. Keeves), 98-109. The Netherlands: Elsevier Science.

Masters, G. and Evans, J. (1986) Banking non-dichotomously scored items. *Applied Psychological Measurement* **10**, 355-367.

Rasch, G. (1960) Rasch G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Denmark Paedagogiske Institute, Copenhagen, Denmark.

Smith, A., Rush, R., Fallowfield, L., Velikova, G. and Sharpe, M. (2008) Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology* **8:33**.

Stone, M. (2002) Knox's cube test – revised (KCT-R). Wood Dale, IL: Stoelting.

Stone, M. and Yumoto, F. (2004) The effect of sample size for estimating Rasch/IRT paramters with dichotomous items. *Journal of Applied Measurement* **5**, 48-61.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D. and Rizavi, S. (2003) Small sample estimation in dichotomous item response models: effect of prios based on judgemental information on the accuracy of item parameter estimates. *Applied Psychological Measurement* **27**, 27-51.

Wang, W. and Chen, C. (2005) Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement* **65**, 376-404.

Wheadon, C. and Whitehouse, C. (2006) GCSE Science multiple-choice tests new specification (4460) First award: November 2006 Technical report

Wright, B.D. (1995) Which standard error? *Rasch Measurement Transactions* **9**, p 436.

Wright, B.D. and Master, G (1982) Rating scale analysis*. Rasch Measurement.* Chicago, IL: MESA Press.

Wright, B.D. and Stone, M.H. (1979) *Best Test Design. Rasch Measurement.* Chicago, IL: MESA Press.