Identification of Underachievers using a Common Scale

Sik Kuan Low and Soon Chew Chia
Education Programmes Division, Ministry of Education, Singapore
Low_Sik_Kuan@moe.gov.sg and Chia_Soon_Chew@moe.gov.sg

ABSTRACT

Underachievement is wastage of human capital in the long run. Timely and accurate identification of underachievers, coupled with early intervention will go a long way to allow the students to achieve their full potential in life. On the most general level, underachievement is defined as a discrepancy between ability and expected performance. Data on student academic performance in their school Semestral Assessment (SA) was often used to identify students who underachieved in their academic performances. This was conceivable when analysis was conducted at individual school level where students sat for the same SA paper. However, when analysis was conducted at the national level involving many schools, the approach became problematic because the SA papers from different schools were of different standards and hence not comparable. In this study, the method of Rasch Analysis was employed to calibrate test scores of different SA papers from different schools onto a common scale using a common anchor for all the schools. The study assessed the reliability and validity of the common scale used to identify Year 8 underachievers across the different schools. While data on mathematics raw scores showed moderate correlation of 0.69 between Year 7 and Year 8 mathematics, the aligned common scales derived from the raw scores showed high correlation of 0.92 between the two years, indicating the common scales were more reliable and better predicted achievement for use in identification of underachievement in mathematics.

Keywords: underachievers, common scale, reliability, validity

INTRODUCTION

Cultivating the latent ability of every student is the central task of all education. Underachievement is wastage of human capital in the long run. This study examined the problem associated with the identification of underachievers when using school assessments that were of different standards and not comparable. The study applied the method of Rasch analysis to align the different assessments on a single, common scale. The study further assessed the reliability and validity of the common scale. While data on mathematics raw scores showed moderate correlation of 0.69 between the Year 7 and Year 8, the corresponding common scales scores showed high correlation of 0.92 between the two years, indicating the common scales were more reliable and better predicted achievement than the original raw scores.

BACKGROUND

At school level, outcome was typically assessed by achievement bands derived from the raw scores. The convention was to group the achievement into four bands by raw scores, band 1 (85 marks and above), band 2 (70-84 marks), band 3 (50-69 marks), and band 4 (less than 50 marks). From school's perspective, students possibly underachieved if performance dropped by at least two bands, equivalent to a drop of 16 marks or more in the raw scores. Percentiles ranks of the raw scores could also be used in place of the achievement bands. However, both the approaches had limitations because the bands or the percentiles ranks derived from the raw scores were not comparable across the schools. When underachievers were identified using percentiles ranks derived from raw scores, it was found there were over-identification of underachievers in schools with more stringent assessments. This led to the impetus to study how to align different assessments on a common scale. Attempt to align different assessments on a common scale is not new. For examples, Rasch model was separately applied to align different TCE[1] subjects on the same scale (TQA, 2000) and to calibrate 34 different

[1] Tasmanian Certificate of Education

GCSE[2] subjects on a common scale (Coe, 2008). In another study, multilevel Rasch model was applied to construct a common scale for a single variable that was repeated measured (Johnson, 2002).

METHOD

Rasch model applied to longitudinal data

In this study, the data for the 2005 Year 7 cohort was obtained from N randomly selected schools from the Ministry of Education database. The method of Rasch analysis was employed to link the mathematics scores of the Semestral Assessment (SA) from forty (N = 40) secondary schools. This approach applied the Rasch model on longitudinal data which enabled repeated measure of the construct (ability) over a specified time period. The underlying assumption was the invariance of mathematics reasoning ability over time. Students' mathematics ability was estimated by Rasch model using longitudinal data collected at three times points separated by six months apart. The mathematics grades collected at time 1 from the national examination were converted to polytomous item scores (U,E,D,C,B,A,A*→ 0,1,2,3,4,5,6). This was used as the anchor to link the data nested in N schools collected at times 2 & 3, where the raw marks were also similarly converted to polytomous item scores. Using Winsteps, the longitudinal data collected at times 1, 2 & 3 was fitted by the Rasch model using multilevel Rasch analysis on 2N + 1 items. The model fit was evaluated. Once established that the scale was productive for measurement, the mathematics common scale was constructed (Fig. 1).

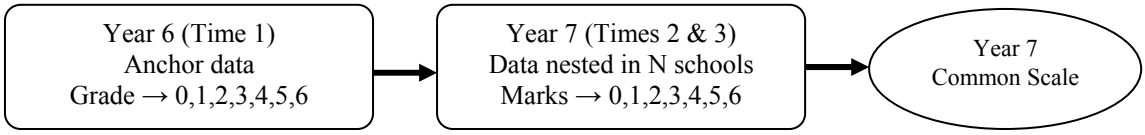

Figure 1: Rasch model for mathematics common scale for Year 7

The common scale scores obtained in time 3 were converted to polytomous item scores (0,1,2,3,4,5,6) for use as the new anchor to link the data nested in N schools collected at times 4 & 5. Using Winsteps, the longitudinal data was fitted by the Rasch model to generate the mathematics common scale for Year 8 (Fig. 2).

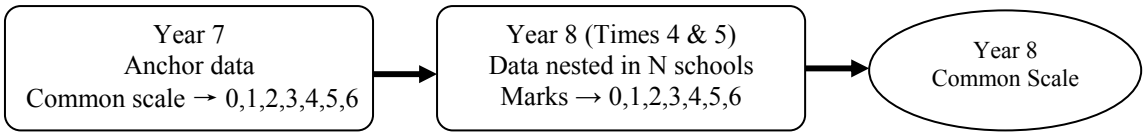

Figure 2: Rasch model for mathematics common scale for Year 8

The Rasch model required that the items worked together to define a single unidimensional construct (unidimensionality), and that the items were not related to each other (local independence). The requirement of unidimensionality and local independence were met when the data fit the model and item reliability was established. Model fit was interpreted using fit statistics mean-square values shown in Table 1 (Linacre, 2002). Item reliability of 0.90 (or greater) indicated good range of item measures and was considered an excellent scale (Waugh, 1998).

Table 1: Interpretation of parameter-level mean-square fit statistics

Fit Statistics	Implication for Measurement
> 2.0	Distorts or degrades the measurement system
1.6 to 2.0	Unproductive for construction of measurement, but not degrading
0.5 to 1.5	Productive for measurement
< 0.5	Less productive for measurement. May produce misleadingly good reliabilities & separations

[2] General Certificate of Secondary Education

Reliability & validity evidence

The reliability of the mathematics common scale was established through evidence of internal consistency and the test-retest reliability. Internal consistency showed that items in the common scale were measuring the same construct. Test-retest reliability showed the stability of common scales scores over time. The internal consistency measure for this study was the Person reliability from Rasch analysis. A lower bound for the test-retest reliability of the common scale was also estimated by correlating common scale scores for the same group of students with the common scale scores obtained over a one-year period. The correlation between these two set of common scales scores were expected to be smaller than the test-retest reliability and therefore formed the lower bound.

The validity of the mathematics common scale was the extent that inferences made from the common scale were appropriate, meaningful, and useful. The study first established the predictive validity the mathematics common scale by showing that the mathematics common scale was effective in predicting the mathematics performance at international examination (TIMSS). Second, the study established the construct validity by showing that the mathematics common scale correlated more strongly with mathematics than with other the less related subjects.

Relative difficulty of SA papers

Rasch analysis checked the difficulty of SA papers and calibrated them on a single, common scale. It assumed that each SA paper measured an underlying common trait which was the mathematics reasoning ability. By analyzing how students performed on the range of SA papers it was possible to arrange the SAs on a "difficulty scale". At the same time as the SAs were calibrated on the common scale, Rasch analysis also mapped the achievement of the students on the same scale. Correlation analysis of relative difficulty with achievement on the common scale enabled the comparison of achievement bands from the schools with range of relative difficulty of SA papers.

Identification of underachievers

Underachievement is defined as a discrepancy between ability and expected performance. Another level of validation of the common scale was to examine the efficacy of the common scale for identification of underachievers (UA). In this validation study, ability was measured by prior attainment in the subject. Attainments were respectively described by achievement bands, raw scores percentiles ranks and common scales percentiles ranks. Three methods of identification were evaluated whereby the underachievers were respectively identified by discrepancies using (1) achievement bands, (2) raw scores percentiles ranks, and (3) common scales percentiles ranks. With achievement bands, UA were identified by a drop of at least two achievement bands (Table 2). With percentiles ranks, UA were identified by a drop of at least D percentiles ranks (Fig. 3). The threshold ($D \geq m$) was chosen to coincide with the median (m) of discrepancies of UA identified by (1) to ensure 50% overlap of UA identified by the Methods (1), (2) & (3).

Table 2: UA identified by Achievement Bands

		Year 8			
		Performance			
Year 7		Band 1	Band 2	Band 3	Band 4
Ability	Band 1			UA	UA
	Band 2				UA
	Band 3				
	Band 4				

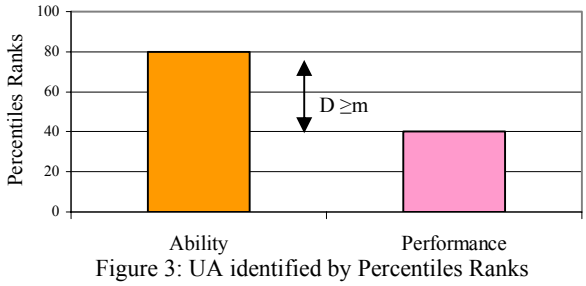

Figure 3: UA identified by Percentiles Ranks

RESULTS

Model fit

The overall assessment of the model fit for the common scales was good with mean item fit statistics range from 0.86 to 0.92 (Appendix). This showed that the common scales were productive for measuring the underlying common trait. Analysis of the individual item fit statistics showed that majority (over 98%) of the items was productive for measurement (Table 3).

Table 3: Mean-square Fit Statistics of individual items

		Year 7	Year 8
	Fit Statistics	No. of items	No. of items
Distorting	> 2.0	0	0
Unproductive	1.6 to 2.0	1	2
Productive	0.5 to 1.5	77	79
Less productive	< 0.5	0	0

Reliability

Person reliability was 0.90 and 0.93 for Year 7 and Year 8 respectively, indicating high internal consistency and high reliability of the common scales derived from the Rasch model. Item reliability was 1.0 for both Year 7 and Year 8. This showed the common scale spanned across a wide difficulty range and had the hallmark of an excellent measurement scale. While data on mathematics raw scores (Year 7 & Year 8) showed moderate correlation of 0.69 between the two years, the common scales scores from Rasch model showed high correlation of 0.92 between Year 7 and Year 8. This showed the common scales scores had a test-retest reliability of at least 0.92.

Predictive validity

External validation of the common scales scores through the analysis by Research and Evaluation Section, Ministry of Education using TIMSS 2007 International Database showed that the correlation between mathematics common scales scores (Year 7) and the TIMSS scores was 0.77. This showed the predictive validity of the mathematics common scales scores which accounted for about 60% of the performance in TIMSS.

Construct validity

Validation of both the convergent and divergent validity of the common scales scores was through analysis using the GCE 'O' level data from the School Cockpit database. The mathematics common scales scores (Year 8) correlated with the GCE 'O' level examination (Year 10) with mathematics (0.72), with science (0.53), with English (0.42), with mother tongue (0.20). This showed that mathematics common scales was more correlated with mathematics than with the less related subjects as Science, English and mother tongue, the evidence of construct validity.

Relative difficulty of SA papers

The relative difficulty of the SA papers on the Rasch scale spanned from easy (–1.9 logit) to very difficult of 4.7 logits. The distribution of bands across schools with differing difficulty logits showed that the achievement bands were not comparable across schools. For example, Band 3 students from the school with difficulty of 4.7 logit could be seen to be comparable to Band 1 from schools with difficulty logits of 1.0 and below (Fig. 4).

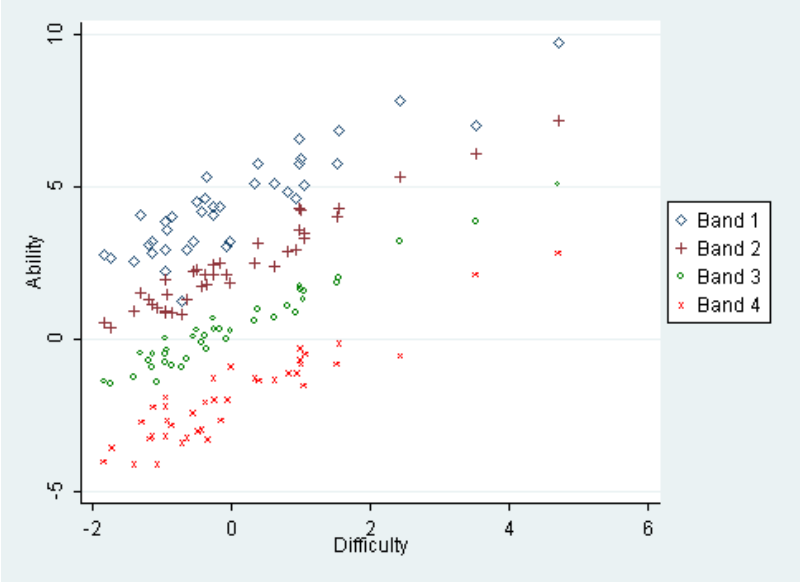

Figure 4: Relative difficulty and ability

The correlation of relative difficulty with achievement was 0.83, 0.83 and 0.71 for Year 7, Year 8 and Year 10 respectively, where the Year 7 and Year 8 achievement was measured by the common scales and the Year 10 by the national examination (Fig. 5). This showed that relative difficulties of the SA papers were positively correlated with academic achievement.

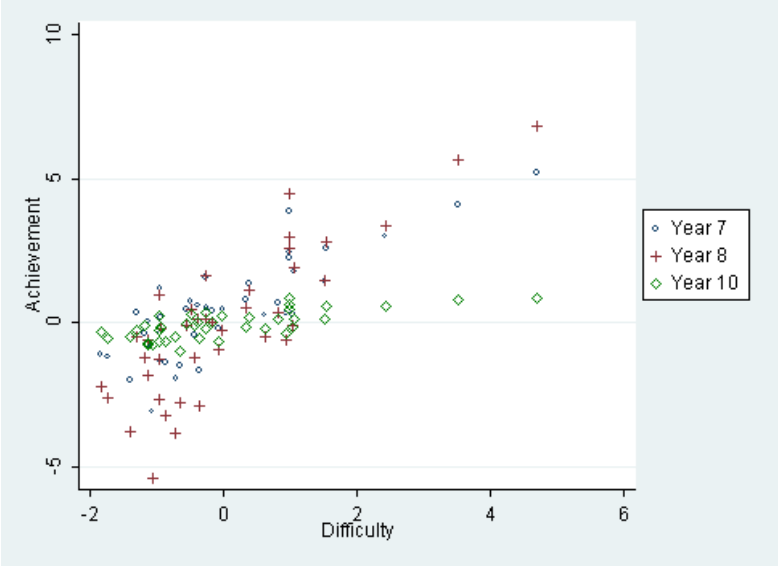

Figure 5: Relative difficulty and achievement

Identification of underachievers

The proportion of underachievers (UA) identified by Methods 1, 2, and 3 were 3%, 20%, and 9% respectively. Identification using achievement bands (Method 1) were conservative in that the base was restricted to students with prior achievement of Band 2 or better. Identification of UA using raw scores percentiles (Method 2) was also presented with difficulties because both the raw scores and the students' attainment were not measured on the same scale. The impact was the observation of large gaps in the proportions of UA identified by Methods 2 and 3. This was especially pronounced for schools with high difficulty logits. For example, the two schools with difficulty of 3.5 and 4.7 logits showed discrepancies of 25% and 23% respectively between Method 2 & 3 of identification of UA With the common scale (Method 3), the proportion of UA was also noted to correlate negatively with relative difficulty with a correlation value of – 0.64 (Fig. 6).

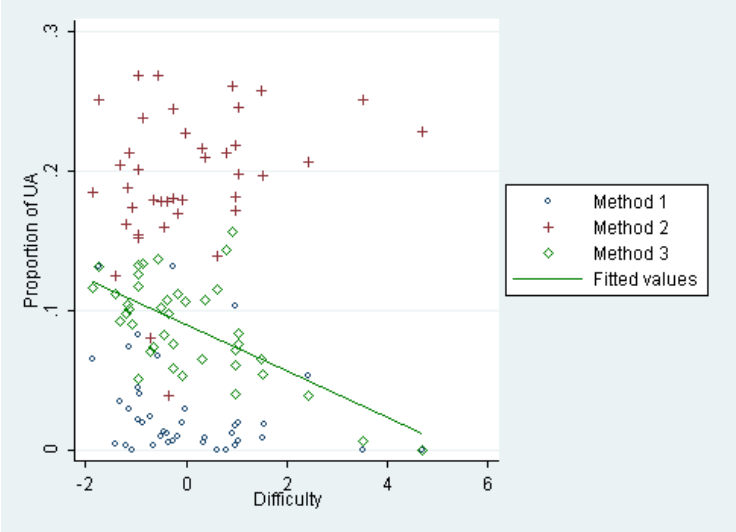

Figure 6: Relative difficulty and proportions of UA

DISCUSSION

Tasmanian Qualifications Authority (TQA, 2000) applied Rasch analysis to align different TCE subjects on the same scale. TQA used the subject score scaling to compare the relative difficulties of achieving each award in each subject. Coe (2008) also applied Rasch analysis to align 34 GCSE subjects on the same scale using the partial credit model within Winsteps, which estimated the difficulty of individual grades, and the overall subject difficulty. In both TQA and Coe, the data was nested in different subjects taken by the same group of test-takers over the same sitting, the linking was through the test-takers who took the same subjects in the same sitting; the overall assessment of the subject grades were used as the test items and the underlying common trait measured was the general academic ability.

In the study reported in this paper, Rasch analysis was applied to align the mathematics SA papers from different schools on the same scale. This study differed from TQA and Coe in that, the data was nested in N different schools collected over three time points, the linking was through a common national examination, and the underlying common trait measured was the mathematics reasoning ability. While TQA and Coe were able to compare the relative difficulty of different subjects taken by the same group of test-takers in the same sitting, this study compared the relative difficulty of the mathematics SA papers from different schools taken at two time points over a 6-month period. One contention was whether it was valid to apply Rasch analysis on longitudinal data collected over time. This amounted to whether the assumption of the existence of the unidimentional mathematical reasoning ability invariant over time was acceptable. On this note, it was noted that assumption of measurement invariance over time was not new. Johnson (2002), a repeated measures, multilevel Rasch model was applied to longitudinal data on criminal behavior under the assumptions of conditional independence, additivity, and measurement invariance over time. In the final analysis, the issue was whether the data fitted the Rasch model under the purported assumption.

CONCLUSION

Assessment for the future generations could well be served not so much by more collection of data, but rather by more effective and creative use of existing data. The goal of this study was to create a common scale linking the SA papers. The challenge was to provide evidence that supported the unidimensionality assumption of a common trait in the SA papers and to show evidence to establish for the reliability and the validity of the common scale created using the Rasch model. The good model fit statistics suggested that the longitudinal data fitted the Rasch model. Person reliability was 0.90 and 0.93 in two separate Rasch models applied to Year 7 and Year 8 data was strong support for the existence of a common mathematics trait across the SAs. The Item reliability was 1.0 for both Year 7 and Year 8 showed the items measured a wide range of item difficulties. Evidence of reliability of the common scale also came from the establishment of a lower bound test-retest

reliability of 0.92. Predictive validity was established when the common scales correlated with an international mathematics examination TIMSS. Construct validity was established when the common scale (Year 8) was shown to be more correlated with mathematics than with the less related subjects as Science, English and mother tongue in the national examination (Year 10).

With the establishment of a reliable and valid mathematics common scale, the relative difficulty of mathematics SA papers was also established, together with the knowledge that relative difficulty of mathematics SA papers positively correlated with mathematics achievement.

Finally, identification of students who did not achieve at their potential was more meaningful with a common scale. Without the common scale, students were more likely to be assessed as underachievers in schools with more difficult SA papers. With the common scale, the proportion of underachievers correlated negatively with the relative difficulty of SA paper, meaning that schools which had relatively harder SA papers also had relatively lower proportions of underachievers identified using the common scale.

ACKNOWLEDGEMENTS

The research reported here was supported by the Senior Specialist Research Fund (SSTRF) of Ministry of Education. The external validation of the mathematics common scales scores, analysis by Research and Evaluation Section, Ministry of Education using TIMSS 2007 International Database.

REFERENCES

Coe, R. (2008). "Comparability of GCSE examinations in different subjects: an application of the Rasch model." Oxford Review of Education 34(5): 609 – 636.

Johnson, C. and S. W. Raudenbush (2002). A Repeated Measures, Multilevel Rasch Model with Application to Self-reported Criminal Behavior, University of Michigan.

Linacre, J. M. (2002). "What do Infit and Outfit, Mean-square and Standardized mean?" Rasch Measurement Transactions 16:2: p.878.

TQA (Tasmanian Qualifications Authority) (2000). "Using Rasch Analysis to Scale TCE subjects." from http://www.tqa.tas.gov.au/4DCGI/_WWW_doc/003675/RND01/Rasch_Intro.pdf.

Waugh, R. F. and P. A. Addison (1998). "A Rasch measurement model analysis of the revised approaches to studying inventory." Bristish Journal of Educational Psychology 68(1): 95-112.

APPENDIX

Figure 1: WINSTEPS Table 3.1 (Year 7)

TABLE 3.1 Year 7
INPUT: 12782 Persons 81 Items MEASURED: 12782 Persons 78 Items 7 CATS 3.68.2

SUMMARY OF 12453 MEASURED (NON-EXTREME) Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	10.3	2.9	.62	.80	.97	-.2	.99	-.1
S.D.	3.8	.3	2.54	.17	1.03	1.2	1.07	1.2
MAX.	17.0	3.0	10.05	1.74	9.90	5.3	9.90	8.3
MIN.	1.0	1.0	-6.30	.63	.00	-3.2	.00	-3.2

REAL RMSE	.94	ADJ.SD	2.36	SEPARATION	2.51	Person RELIABILITY .86
MODEL RMSE	.82	ADJ.SD	2.40	SEPARATION	2.93	Person RELIABILITY .90
S.E. OF Person MEAN = .02						

MAXIMUM EXTREME SCORE: 157 Persons
MINIMUM EXTREME SCORE: 172 Persons
VALID RESPONSES: 3.7%

SUMMARY OF 78 MEASURED (NON-EXTREME) Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	1670.2	478.1	.00	.07	.88	-1.8	.86	-2.0
S.D.	5157.2	1402.3	1.39	.01	.25	3.5	.24	3.4
MAX.	46866.0	12777.0	5.65	.10	1.61	9.9	1.69	9.9
MIN.	451.0	194.0	-2.60	.01	.46	-8.7	.45	-8.8

REAL RMSE	.08	ADJ.SD	1.39	SEPARATION	18.46	Item RELIABILITY 1.00
MODEL RMSE	.07	ADJ.SD	1.39	SEPARATION	18.97	Item RELIABILITY 1.00
S.E. OF Item MEAN = .16						

LACKING RESPONSES: 3 Items

Figure 2: WINSTEPS Table 3.1 (Year 8)

TABLE 3.1 Year 8
INPUT: 12782 Persons 81 Items MEASURED: 12782 Persons 81 Items 7 CATS 3.68.2

SUMMARY OF 11706 MEASURED (NON-EXTREME) Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	9.5	3.0	.42	.82	.96	-.2	.98	-.1
S.D.	4.5	.2	3.24	.17	1.01	1.2	1.06	1.2
MAX.	17.0	3.0	10.30	1.70	9.90	5.6	9.90	6.0
MIN.	1.0	1.0	-7.25	.67	.00	-3.2	.00	-3.2

REAL RMSE	.96	ADJ.SD	3.09	SEPARATION	3.24	Person RELIABILITY .91
MODEL RMSE	.84	ADJ.SD	3.13	SEPARATION	3.74	Person RELIABILITY .93
S.E. OF Person MEAN = .03						

MAXIMUM EXTREME SCORE: 301 Persons
MINIMUM EXTREME SCORE: 775 Persons
VALID RESPONSES: 3.7%

SUMMARY OF 81 MEASURED (NON-EXTREME) Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	1427.8	466.6	.00	.08	.92	-1.1	.91	-1.3
S.D.	4129.7	1377.6	1.84	.01	.25	3.1	.23	3.0
MAX.	38278.0	12782.0	5.85	.12	1.86	9.9	1.77	9.3
MIN.	265.0	194.0	-3.42	.01	.56	-6.3	.57	-6.3

REAL RMSE	.08	ADJ.SD	1.84	SEPARATION	22.30	Item RELIABILITY 1.00
MODEL RMSE	.08	ADJ.SD	1.84	SEPARATION	23.07	Item RELIABILITY 1.00
S.E. OF Item MEAN = .21						