APPLICATIONS OF ITEM RESPONSE THEORY TO IDENTIFY AND ACCOUNT FOR SUSPECT RATER DATA

Nathan Zoanetti
Patrick Griffin
Ray Adams
University of Melbourne

ABSTRACT

This paper describes a plausible values imputation approach for deriving population scores on several language proficiency domains. The approach harnessed a multi-dimensional item response analysis combining student responses, rater judgements and student background variables. The target population was grade one and grade two primary school students enrolled in the Hong Kong schooling system. The raters were local teachers of English employed within the sampled schools. The primary objective of this research was to impute plausible values for data where no data was provided or where rater data was deemed suspect. By necessity, a secondary objective of this study was to establish rules for justly excluding particular data on the basis of questionable validity. Surveys such as TIMSS, PISA and NAEP have used such “plausible value” methodologies to account for incomplete test designs and person non-response (Beaton & Johnson 1990, Adams, Wu & Macaskill 1997). The point of difference between this study and other similar studies was the use of item response theory (in particular plausible values) to replace and quantify the impact of potentially invalid rater judgements in a large-scale educational survey.

INTRODUCTION

The purpose of this study was to impute plausible values for students on the *Profiles in English as a Second Language* (*ESL Profiles*) (Griffin, Smith & Martin 2003) assessment including cases where responses were either missing or deemed suspect. Suspect data were defined where response patterns represented random, missing or deterministic or fixed responses. The data in these cases were regarded as being either protest responses, missing responses or perhaps teacher response not related to any relevant student language behaviour. Several approaches were taken to their identification. Clearly missing responses were simple to identify. Fixed responses consisted of the same rating assigned to every item and hence no between item variance. Deterministic patterns were those in which the pattern was predictable (e.g. 12321232123) and random patterns of response were identified using fit indices and item response modelling. This approach could also be used to identify deterministic patterns. Once these response patterns were identified, all such cases were set to ‘missing’ and their values imputed using patterns of responses from students with similar characteristics. Such imputation has become an accepted method for survey data analysis (Adams, Wu & Macaskill 1997, Gonzalez, Galia & Li 2004, Beaton & Johnson 1992).

Identification of dubious rater data was to be the first purpose of this study. This was carried out primarily using goodness-of-fit statistics associated with the chosen scaling model. If this were to be done as a routine check of data collection and feedback given to raters, they could conceivably be approached, cautioned or re-trained depending on the policies of the project. As such, the initial phase of this investigation served as a quality control mechanism at the rater level. There was however no opportunity for feedback.

It was hoped that differences in population parameter estimates upon selective exclusion of rater data would provide insight into the impact of the suspect rater judgements. Several variants of the original observed data set were used in separate imputation procedures to meet this end.

A post-hoc analysis of rater judgements was also undertaken. This second rater-targeted analysis served to identify raters who may have been systematically harsh or lenient. The outcomes and assumptions associated with this investigation are presented in the section *Diagnostic Charts Using Mean EAP Shift by Rater*.

Measurement Variables

The outcome variables of interest were individual student measures of latent ability on the three *ESL Profiles* dimensions; *Speaking*, *Reading* and *Writing*. Measures of latent ability on the ancillary *Interview Test of English* (*ITEL-ed*) (Griffin, Tomlinson, Martin, Adams & Storey 2004) were also estimated. These student measures were used to estimate population parameters.

A marginal item response model (Adams, Wu & Macaskill 1997) was implemented to estimate a distribution of latent ability for each student on the basis of their *ITEL* responses, their *ESL Profiles* scores and their background variables. A consequence of this plausible value methodology was that at no stage were reportable point estimates of individual student abilities obtained. This is because plausible values contain random error variance at the individual level. In other words, two students achieving the same test score could be assigned different plausible value estimates (Wu 2004). While this limits the validity of comments on student performance at the individual level, unbiased estimates of population (and population sub-group) ability on each of the dimensions are attainable (Forsyth *et al*. 1996, Wu 2004). *ITEL* achievement was also measured as part of the multi-dimensional scaling. This served to enhance the reliability of *ESL Profiles* posterior distributions by virtue of being a correlated ancillary test (de la Torre & Patz 2005).

Target Population and Sampling

Two-stage cluster sampling was conducted as part of the *Hong Kong Primary Native English-speaking Teacher* (HKPNET) project from which the data have been obtained. The sampling frame comprised the 144 targeted schools as directed by the Education and Manpower Bureau (EMB) (Griffin, Woods, Nadebaum & Tay 2005). From these schools, clusters of 15 students were selected at random from each year level in which a Native English-speaking Teacher (NET) operated. Students were sampled from two grades; Primary One and Primary Two.

A sample of 2873 students was established. The data chosen for use in this study were intentionally not screened or cleaned prior to use.

Theory

Several theoretical topics in educational measurement will now be explained. The first of these topics is Item Response Theory (IRT). IRT models the relationship between a person's ability on some latent trait (measured by a test) and the person's observed responses to items on that test (Stocking 1999, Lord 1980).

IRT enables construction of metrics where both items and persons are assigned values representing their respective difficulty and ability. The most elementary IRT model is the *Rasch*, or simple logistic, model (Rasch 1960, 1980). In the case of the *Rasch* model, the distances between item difficulties (δ) and person abilities (θ) on the common metric can be interpreted as the logarithm of the odds of success (X=1) of a person on a given item *i*. This can be derived from equation (1) and is summarised in equation (2):

$$\text{p}=\text{P}_i(X=1)=\frac{e^{(\theta-\delta_i)}}{1+e^{(\theta-\delta_i)}} \tag{1}$$

Re-arranging equation (1) demonstrates the meaning of distances on the latent scale:

$$\ln(\frac{p}{1-p})=\theta-\delta \tag{2}$$

In some cases, item responses may reflect a degree of correctness in the answer to a question, rather than simply correct/incorrect. Both the *ESL Profiles* and the *ITEL* instruments consist of items comprising more than two ordered score categories. Analysis of such polytomously scored data is possible using the Partial Credit Model (PCM) as derived by Masters (1982). The PCM is an extension of the dichotomous *Rasch* model.

Masters (1982) derived the PCM by applying the dichotomous *Rasch* model to adjacent pairs of score categories. That is, given that a student's score is k-1 or k, the probability of being in score category k has the form of the simple *Rasch* model. The PCM can be mathematically summarised: if item *i* is a polytomous item with score categories 0, 1, 2, …, m_i , the probability of person *n* scoring *x* on item *i* is given by:

$$\Pr(X_{ni}=x)=\frac{\exp\sum_{k=0}^{x}(\theta_n-\delta_{ik})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{x}(\theta_n-\delta_{ik})} \tag{3}$$

$$\text{Where }\exp\sum_{k=0}^{0}(\theta_n-\delta_{ik})\equiv 1 \tag{4}$$

Multiple estimation algorithms have been applied in the context of IRT, each yielding different parameter estimators with particular benefits and limitations (Monseur & Adams 2002).

Two software packages were utilized for estimation throughout this study; *Quest* (Adams and Khoo 1995), and *ConQuest* (Wu, Adams and Wilson 1998). *Quest* can be used to estimate person abilities of the form Maximum Likelihood Estimates (MLEs) and Weighted Maximum Likelihood Estimates (WLEs) (Warm 1983). *ConQuest* has the additional capacity to estimate conditioned posterior ability distributions for each person on multiple dimensions (see Adams, Wu and Macaskill 1997). From the conditioned student posterior ability distributions, plausible values (PVs) can be randomly drawn. The mean value of the conditioned posterior distribution is known as the Expected A-Priori (EAP) estimate.

When deriving PVs, instead of obtaining a point estimate (θ_n) for ability, a probability distribution $h(\theta|x_i)$ is produced for each person as determined by the response vectors, x_i (Wu, 2004). Plausible values are randomly drawn from this distribution, which is generally known as a posterior distribution. Due to the continuous nature of the posterior distribution (and therefore non-zero spread), $h(\theta|x_i)$ incorporates measurement error unlike the MLE and WLE point estimate methods. It is also possible to condition the posterior distributions with a vector of individual background variables, y_i, to increase the accuracy of associated population parameter estimates (Reckase 2002). The corresponding posterior distribution for a person with responses x_i and background variables y_i would then be denoted $h(\theta|x_i,y_i)$.

Another important element of this study regarded the fit of student data to the underlying measurement model. In large scale surveys there will undoubtedly be a number of responses which, when the characteristics of the respondent are considered, appear largely unexpected.

One means of identifying aberrant response patterns is through the use of residual-based IRT fit statistics. It is possible to use fit statistics to identify subsets of items and persons whose behaviour is not consistent with the measurement model (Meijer & Stitsma 2001, Wu 1997). In this paper it is argued that model fit, while desirable from a statistical viewpoint, is not a necessity for identification of misfitting raters.

Wright and Masters (1982) describe two fit statistics based on standardised residuals (z_{ni}), one weighted and one un-weighted. This study does not utilise the un-weighted form. These are more sensitive to singular aberrant responses (Wright & Mok 2000). Instead, the *information weighted mean square residual goodness of fit,* or *infit,* statistic is used. The *infit* statistic is calculated by taking the weighted average of squared residuals so that responses that are extremely easy or difficult to a person are given less weight than proximal responses. Thus, the *infit* examines unexpected responses to items targeted at the ability of the person in question. Mathematically, the weighted fit statistic, (v_i), is

$$v_i = \sum_{n=1}^{N} z_{ni}^2 W_{ni} \bigg/ \sum_{n=1}^{N} W_{ni} \qquad (5)$$

where

$$W_{ni} = P_{ni}(1 - P_{ni}) \qquad (6)$$

For person fit, (u_n), the equation is the same but the residuals, z_{ni}, are summed over the total number of items rather than persons.

In this study, reliability and validity considerations were always at the fore. The implications of recoding suspect data to missing needed to be considered in light of reliability reductions and validity violations.

Reliability is classically expressed as the ratio between true variance and the observed variance (Thorndike 1988). A recent conception of reliability when using plausible values is to examine the *predictive reliability (R_P)* of person posterior distributions. This is a measure of the proportion of variance in the data (σ_θ^2) that is accounted for by the measurement model for a given student's estimate (Mislevy, 1991). It is calculated as:

$$R_P = \frac{\sigma_\theta^2 - \sigma_e^2}{\sigma_\theta^2} = 1 - \frac{\sigma_e^2}{\sigma_\theta^2} \qquad (7)$$

In (1), σ_e^2 is the variance of the student posterior distribution. Averaging R_p over all students produces a reliability index referred to as the PV/EAP reliability (produced as standard output by *ConQuest* (Wu, Adams & Wilson 1998)).

Validity was of concern for the HKPNET project when the data was considered suspect. Clearly for the sub-set of pupils for whom the data was considered as non valid, there was little or no chance of generalising to the population of school students in the Hong Kong PNET scheme. However, it remained important to ensure that as much valid data as possible was retained and not unnecessarily discarded.

METHODOLOGY

Suspect Data Identification

Two methods for identifying suspect data were applied. Firstly, *Quest* was used to scale each of the three *ESL Profiles* strands separately in order to evaluate person fit statistics for each student. This would quantify the agreement of student response data to the item response model (Wright & Mok 2000). The fit statistic of interest was the *infit* statistic.

Secondly, where the number of responses per *ESL Profiles* dimension was limited to less than two, the student data were flagged for exclusion. These response patterns were defined as ‘missing’ for the purposes of the analysis. It was reasoned that estimates derived from response strings featuring less than two scores would be inherently unstable and unreliable due to high associated measurement errors. The rule for excluding students for which insufficient data had been recorded was at first formulated by intuition. It was not until the posterior distributions were analysed in section *Suspect Data on the Basis of Minimal Response Information* that the rule was confirmed as being appropriate.

The first method outlined above requires further explanation. *Quest* produces fit statistics for both items and persons (*ConQuest* presently produces item fit statistics only). The infit is one such statistic. Infit values of one generally indicate perfect overall fit of the person response data to that predicted by the item response model (Smith, Schumacker & Bush 1995). For values less than one, the data are said to over-fit the model. For the *ESL Profiles* assessment, over-fit had an anticipated prevalence due to the determinacy built into the progression of items in the instrument (Zoanetti 2006). For fit values greater than one, the data are said to under-fit the model. This usually implies a higher level of randomness (and reduced item discrimination) in the response data than that predicted by the model. In terms of mathematically defining the acceptable range of infit values, Smith, Schumacker & Bush (1995) state that as a guide values greater than 1.2 and less than 0.8 for samples of more than 500 are unacceptable. Karabatsos (1997) suggests that such cut-off values tend to vary depending on the purpose for which the ratings are used. Wu (1997) demonstrates the suitability of known reference distributions for determining when infit values are statistically unacceptable according to the scaling model. These results apply to data demonstrating good model fit. Systematic under-fit at the rater level was viewed as the primary indicator of suspect data in this report.

Conditioned Posterior Distribution Imputation

ConQuest (Wu, Adams and Wilson 1998) was used for the majority of the necessary item response modelling. A four step process was applied, described as follows:

1. An initial four dimensional scale analysis was carried out using only item response data for the three *ESL Profiles* strands and the *ITEL* test. Students deemed to have suspect data were omitted from the data set. The purpose of this run was to derive approximate values for the item parameters and the inter-dimensional correlation values.
2. A second estimation routine much like the first was carried out using the approximate values from Step 1 as starting values. The number of estimation nodes (for the numerical integration) was increased and the convergence criterion of the estimation procedure was tightened. The purpose of this run was to derive precise estimates of the item parameters and the inter-dimensional correlations. The reader is directed to the *ConQuest* manual (Wu, Adams and Wilson 1998) for more information on numerical integration constraints and algorithms.
3. The item parameter estimates from Step 2 were used in a third estimation routine as anchor (fixed) values. A regression model was incorporated into the run. Regression coefficients were estimated for student background variables that explained the most variation in student test performance or were of genuine interest for reporting purposes. These included the father's education level, the frequency of English use outside of school, the number of books at home, the number of English language books at home, gender and year level. The school mean *ITEL* performance was also included due to the high intra-class correlation coefficient amongst schools (see Griffin, Woods, Nadebaum & Tay 2005).
4. For this final step, a restored data set (2873 students) was implemented. This consisted of the original number of students, but the *ESL Profiles* data was included selectively according to model fit and the number of missing responses. The purpose of this run was not to estimate anything other than the student posterior distribution location and spread. All other parameters (items, inter-dimensional correlations, regression coefficients) were fixed and considered accurate owing to the use of cleaned and reduced samples that were still representative of the target population (see *Table 2*).

Population Parameter Estimates from Plausible Values

The procedure used for determining estimates for the population mean and variance under varied imputation conditions was based upon that described by Allen et al (1999).

1. An estimate of the population mean μ_i and population variance σ^2_i was calculated for each of the five sets of plausible values.
2. The sampling variance for each of the estimates described in *Step 1* was calculated. These were denoted $Var(\mu_i)$ and $Var(\sigma^2_i)$ for the mean and variance respectively.
3. The best estimate for the population mean ($\mu.$) and population variance ($\sigma^2.$) was calculated by averaging the five values of μ_i and σ^2_i respectively.
4. The variance of $\mu.$ and σ^2 was then derived in terms of a sampling variance and the uncertainty inherent in the multiple imputation methodology (see Gonzalez, Galia & Li 2004).

Analysis of Rater Consistency by EAPs

An investigation of the discrepancy between student EAPs generated in both the presence and the absence of *ESL Profiles* data was conducted. This was possible by comparing observed *ESL Profiles* posterior distributions with *ESL Profiles* posterior distributions based only on background variables and achievement on the *ITEL* assessment. The student EAPs were aggregated and averaged for each rater. The objective was to establish rules for identifying raters for whom a significant discrepancy existed. It was thought that these discrepancies could indicate rater behaviour or student group performance that was not consistent with the regression model or the student *ITEL* performance. The diagnostic potential of such an investigation seemed attractive, given that systematically large discrepancies might indicate either rater harshness/leniency or unexpectedly high or low student sub-group performance on the *ESL Profiles*. Of course, some student subgroups might be genuinely weak/strong in particular language strands compared to other subgroups with comparable backgrounds and ancillary achievements. Thus large discrepancies do not necessarily detect data that must be wrong. Nonetheless, it was thought that such an analysis might be a useful tool for addressing consequential validity violations and detecting unexpected response patterns.

RESULTS

Suspect Data on the Basis of Mean Person Fit by Rater

To identify raters whose behaviour was not consistent with the scaling model the following multi-step technique was implemented:

1. The mean and standard deviation of person fit values was determined for each rater.
2. Raters were ranked by mean person fit values and diagnostic error plots were produced (see Figure 1).
3. Rater response vectors were inspected from the highest (most under-fitting) rank until the response vectors became clearly interpretable in light of the *ESL Profiles* scoring rubrics.
4. A mathematical rule was formulated to summarise the level of misfit that was to be deemed unacceptable. This rule was applied across *ESL Profiles* dimensions.

This seemed the most valid way forward. Linacre (1990) suggested essentially the same technique (at least for the first three steps) at the individual student level. In this study, it was important to form impressions of misfit at the rater level, as individual student aberrations were frequent and inevitable and were not as pragmatic to pursue.

For the purpose of demonstrating the concept of rejecting rater data believed to be systematically improbable, the cut-off of 1.2 was chosen as an upper limit of acceptable mean person fit. This value, although taken from literature based on data exhibiting good model fit, is arbitrary but seemed to work when used in the following way (Linacre & Wright 1994): In order to add some solidarity to this exclusion rule, the lower limit of a plus/minus two standard error confidence interval for the person fit values for each rater was used. This approach offered a reasonable numerical summary of the cut-off level identified using the previously stated four-step procedure. Data not meeting this criterion were arguably implausible and not interpretable given the *ESL Profiles* composition. An example is evidenced in Rater 1 from Table 1. The inability to interpret the data in terms of a cohesive language proficiency description was important for justifying its removal. The exclusion of data by a researcher in any research must be heavily justified or the validity and conduct of the research will come into question (Howe & Moses 1999).

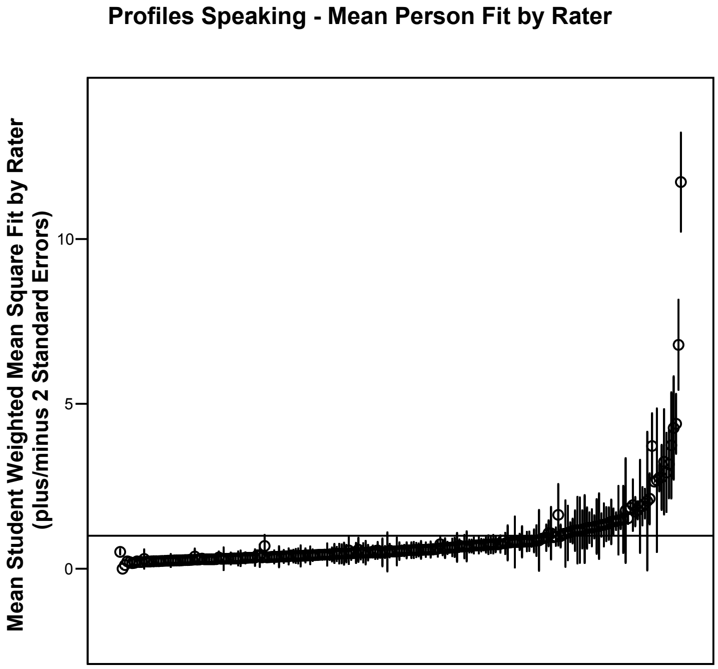

Figure 1 Plot of mean person misfit by rater showing plus/minus two standard deviation intervals

The distinct rise in mean infit values to the right of the person-fit distributions (see Figure 1) raised the question of what type of item score sequences could cause such systematically inflated fit values. Had raters misunderstood the task? Had they submitted protest data? Had they by chance or circumstance rated a subset of the population with an atypical instructional background? A closer look at the kind of data that is flagged by inflated person fit values is offered in Table 1.

Table 1 Response data and infit values for two disparate raters from one school

Student	**Rater**	***ESL Profiles* Speaking Items**									**infit**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
1	**1**	0	1	1	1	2	2	3	0	0	**10.73**
2	**1**	0	1	2	3	2	1	0	0	0	**8.37**
3	**1**	0	0	1	2	1	1	2	3	2	**11.49**
4	**1**	0	0	1	2	2	2	2	3	1	**10.19**
5	**1**	0	0	0	1	2	2	2	3	2	**13.81**
6	**1**	1	0	1	2	2	2	3	3	2	**7.51**
7	**1**	0	0	0	1	2	2	2	3	2	**13.81**
8	**1**	1	0	0	1	1	2	1	2	3	**11.97**
9	**1**	0	0	0	1	1	2	3	3	2	**15.01**
10	**1**	0	0	0	1	1	2	2	3	2	**14.83**
11	**1**	0	0	1	1	2	2	2	2	3	**11.26**
12	**2**	3	3	3	1	1	0	0	0	0	**.55**
13	**2**	3	3	3	2	1	0	0	0	0	**.46**
14	**2**	3	3	3	1	0	0	0	0	0	**.85**
15	**2**	3	3	3	2	1	0	0	0	0	**.46**

In Table 1, Rater 1 had the highest mean person fit for *ESL Profiles Speaking*. The scores that they provided for the nine *ESL Profiles Speaking* items are tabulated together with their colleague's, Rater 2. Comparison of the two raters' responses is a worthwhile exercise. The mean person fit statistic value associated with Rater 1 is 11.69 while that associated with Rater 2 is 0.58.

This is an extreme case, where Rater 1 appears to have scored the assessment in reverse. The data of Rater 2 is much more plausible, given the hierarchical ordering of proficiencies with increasing item number. Investigation of additional response strings quickly led to the observation that *reasonable* data do not result in inflated person fit statistics. Data that are *reasonable* might appear similar to that exhibited by Rater 2 in Table 1.

It is apparent from Figure 1 that a great deal of the individual person fit values must have been below one. This is not ideal from the perspective of a *Rasch* measurement purist (Smith 1990). However, the origin of this observation was easily explainable in terms of the level of determinacy built into the *ESL Profiles*. This concern did not prevent an attempted formulation of data classification rules in this study. Schulz (1990) states that there is a need '…*to show that misfit, or more specifically misfitting ratings, do not seriously invalidate most uses of the Rasch model, but are diagnostically productive…'*. Indeed misfit was used productively as a diagnostic tool throughout this study. The priority here was to construct a number of methods to promote data quality control. The lack of fit to the underlying model for this data set, once understood, was of reduced concern. It was however acknowledged as a threat to the validity of the results in this study.

Suspect Data on the Basis of Minimal Response Information

Figure 2 shows values for the student posterior distribution predictive reliability defined in (12). The latent variance used to generate the curve was the estimated latent variance for the *ESL Profiles* Speaking dimension (which was equal to 4.88 logits). The curves were analogous for the other two *ESL Profiles* dimensions.

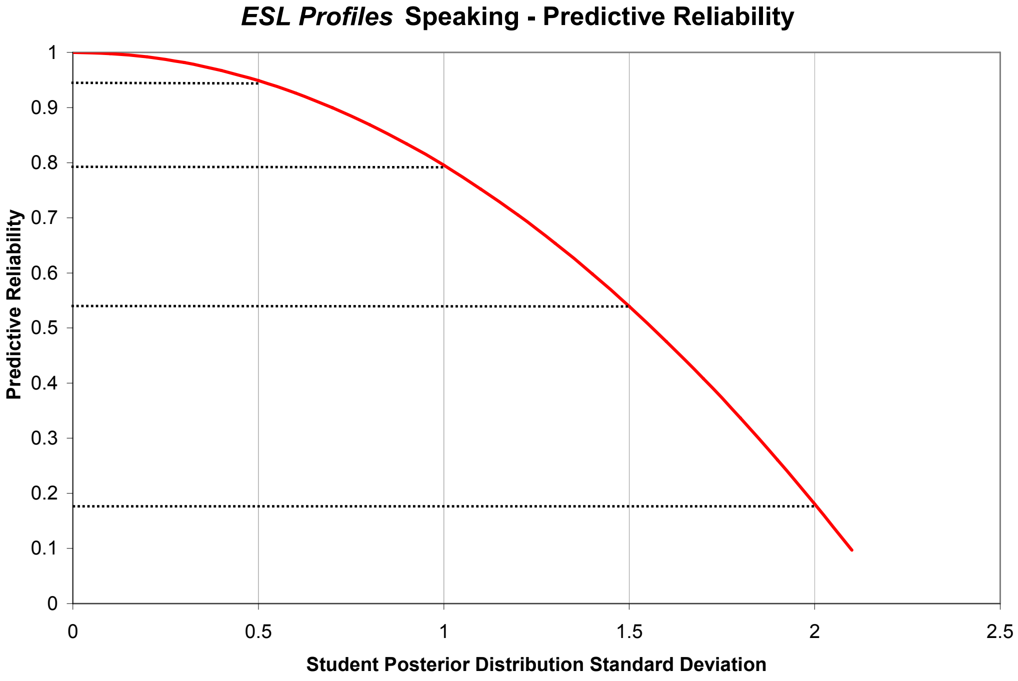

Figure 2 Plot of predictive reliability versus student posterior distribution spread

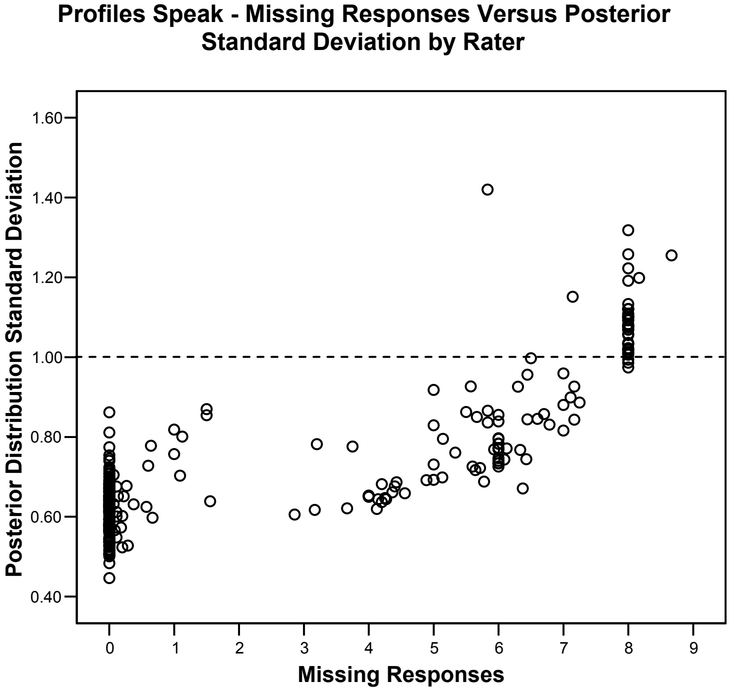

Figure 3 Scatter plot of posterior distribution spread as a function of missing responses

Graphically representing the posterior distribution spread as a function of missing responses at the rater level seemed like a productive way to quantify and classify the precision of rater judgements. Figure 3 illustrates the findings. It can be seen that the majority of raters submit on average at least two responses, but a non-trivial number averaged only one response. The majority of raters produced a mean posterior distribution spread of less than one. This is equivalent to a predictive reliability of better than 0.8. This level of average predictive reliability was almost always attained by raters who on average submitted scores for at least two items per dimension. Given that 0.8 was considered a reasonable figure for a minimally acceptable predictive reliability (though this could be challenged), the second data exclusion rule seemed appropriate.

A summary of the proportion of the observed data impacted by the two data exclusion rules is presented in Table 2. The three data sets shown were those subjected to comparative analyses throughout the study. *Subset 1* and *Subset 2* were used to determine the best possible estimates of imputation parameters (such as inter-dimensional correlations, regression coefficients, item parameter estimates) to be incorporated into the final plausible value multiple imputations.

Table 2 Summary of the three samples used throughout the analysis

Sample Name	Data Exclusion Criteria	Number of Students	Proportion of Original Data
Observed Data	None	2873	1.00
Subset 1	Lower bound of ±2 standard deviation confidence interval of rater mean person fit > 1.2	2235	0.78
Subset 2	As per Subset 1, but also students for which less than two responses were scored for each Profiles dimension	1922	0.67

Posterior Distributions and Population Parameter Estimates

The plausible values methodology described in section *Population Parameter Estimates from Plausible Values* was used to estimate the mean and standard deviation of the latent trait distribution for each of the *ESL Profiles* dimensions and the *ITEL* dimension. The results of applying this methodology to three different data sets representing the same population are presented in Table 3. The purpose of this procedure was to identify how much suspected bias had been removed (and hopefully not introduced) by the data cleaning and imputation process.

Table 3 Population estimates resulting from different proportions of retained *ESL Profiles* data

	No *ESL Profiles* Data				**All *ESL Profiles* Data**				**Composite**			
Parameter	Mean θ		Std Dev θ		Mean θ		Std Dev θ		Mean θ		Std Dev θ	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Speak	-1.73	0.13	2.25	0.14	-1.82	0.13	2.17	0.13	-1.74	0.13	2.20	0.13
Read	-0.06	0.13	2.16	0.14	-0.19	0.12	2.10	0.12	-0.09	0.12	2.12	0.12
Write	-0.81	0.13	2.16	0.13	-0.91	0.12	2.11	0.12	-0.83	0.12	2.12	0.12
ITEL	-0.93	0.12	2.06	0.12	-0.96	0.12	2.05	0.12	-0.95	0.12	2.03	0.12

The data set referred to as *Composite* consisted of *ESL Profiles* data only for students that met the data retention criteria used in the construction of data *Subset 2*. It was hoped that the *Composite* data set would retain enough information to enable valid estimates of the mean and variance of the ability while removing the influence of any misleading and unreliable data.

Unfortunately, owing to the relatively large standard error (S.E.) associated with the population parameter estimates and the fact that each of the above three samples are identical, it was not possible to easily confirm the significance of differences. However, a trend prevailed suggesting that the removal of all suspect cases resulted in estimates more consistent with those resulting from the background variables and *ITEL* performance only. This was an important result. The stability of the estimates of the standard deviation suggested that the characteristic variability within the samples remained fairly unchanged. This might be marked as evidence that the proportion of information removed from the sample had not been too severe. A key question, and one that is difficult to answer, is whether bias had been removed or whether information relating to ability variation had been lost. The answer to this question relies largely on the strength of the justification for removing the suspect data in the first place. Also, it remained that while trends could be inferred, they could not be statistically verified.

Diagnostic Charts Using Mean EAP Shift by Rater

Figure 4 presents a form of diagnostic plot produced as part of this study. The purpose was to plot the mean shift in student EAP caused by each rater's *ESL Profiles* scoring. Importantly, this statistic had both a magnitude and a direction of interest. Large values were of key interest, but their interpretation depended on their sign. Large negative discrepancies might indicate either rater harshness or unexpectedly low student sub-group performance. Large positive discrepancies might indicate rater leniency or unexpectedly high student sub-group performance.

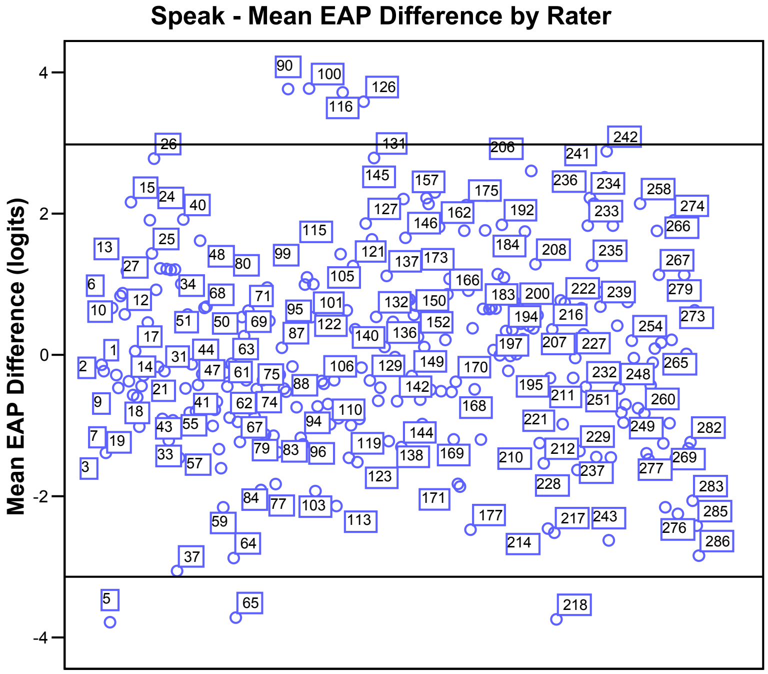

Figure 4 Scatter plot of average impact of rater judgements on student EAP

Determination of a suitable cut-off for the mean EAP difference by rater presented an interesting measurement challenge. It would certainly be possible to develop statistical rules that take into account both the regression model and the *ITEL* performance. However, it was thought that perhaps a more meaningful approach would involve classifying the mean EAP difference in terms of the corresponding step in language proficiency. Perhaps even combining the two aforementioned approaches could be informative.

CONCLUSION

It seems apparent that there is the potential to further develop these data quality measurement techniques. Simulation studies might be used to further enhance the effectiveness of monitoring the mean EAP difference for each dimension in turn. Simulated data would be a good place to start, given that numerous inter-dimensional correlation values, test lengths and regression models could be investigated systematically.

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23.

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M.O. Martin & D. L. Kelly (Eds.), *TIMSS technical report volume II: Implementation and analysis* (pp. 111-145). Chestnut Hill, MA: Boston College.

Adams, R. and Khoo, S.T. (1995) Quest: an interactive item analysis program. Melbourne: Australian Council for Educational Research.

Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). *The NAEP 1996 Technical Report,* NCES 1999-452, by. Washington, DC: National Center for Education Statistics.

Beaton, A.E., & Johnson, E.G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement*, 26(2), 163-175.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Education Research Journal, 5,* 437-474.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

de la Torre, J and Patz, R.J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioural Statistics; 30:3,295-311.*

Gonzalez, E.J., Galia, J., Li, I. (2004). Chapter 11. Scaling methods and procedures for the TIMSS 2003 Mathematics and Science Scales. In Martin, M. O., Mullis, I.V.S., & Chrostowski, S. J. (Eds.) *TIMSS 2003 Technical Report***.** Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Griffin , P., Tomlinson, B., Martin, L., Adams , R., & Storey, P. (2004). *An interview test of English (ITEL-ed).* Melbourne : Profile Press International.

Griffin, P. and Woods K. (2004). Progress report on the Hong Kong PNET Program. Report submitted to the Hong Kong Education and Manpower Bureau. Melbourne : Assessment Research Centre, University of Melbourne

Griffin, P., Woods, K., Nadebaum, C. and Tay, L. Evaluation of the Hong Kong Primary Native English-Speaking Teacher Scheme, Technical Appendix 2005. Appendix C. Sampling.

Griffin , P., Smith, P. G., & Martin, L. (2003). *Profiles in English as a second language*. Melbourne : Robert Andersen & Associates Pty Ltd.

Harnisch, D.L. & Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18(3), 133-146.

Karabatsos, G. (1997). The Sexual Experiences Survey: Interpretation and validity. *Journal of Outcome Measurement, 1,* 305–328.

Kenneth R. Howe, Michele S. Moses. Ethics in Educational Research
Review of Research in Education, Vol. 24, 1999 (1999) , pp. 21-59

Linacre, J.M. (1990) Where does misfit begin? *Rasch Measurement Transactions*, 1990, 3:4 p.80

Linacre, J.M., & Wright, B.D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Lord, F.M. (1980) *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc: New Jersey.

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

Meijer R.R. and Sitsma K. (2001). Person Fit Statistic - What is Their Purpose?. *Rasch Measurement Transactions*, Fall 2001, 15:2 p. 823.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177-196.

Monseur, C., & Adams, R. (2002, April). The limitation of the plausible values. Paper presented at the International Objective Measurement Workshop, New Orleans, LA.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Reckase, M.D. (2002, September). Contributions of non-cognitive questions to improving the precision of NAEP results. Paper presented at workshop on NAEP background questions. National Assessment Governing Board, Washington, DC. Available at http://nagb.org.

Schulz, E.M. (1990). Functional assessment of fit. *Rasch Measurement Transactions*, 1990, 3:4 p.82

Smith, R. M., Schumacker, R. E., & Bush, J. M. (1995, April). Using item mean squares to evaluate fit to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Smith, R.M. (1990) Theory and practice of fit. Smith RM. *Rasch Measurement Transactions*, 1990, 3:4 p.78

Stocking, M.L. (1999). Item Response Theory. In G. Masters and J. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment*, (pp. 55-63), Pergamon, Netherlands.

Thorndike, R. (1988). Reliability. In J.P. Keeves (ed.). *Educational research methodology and measurement: An international handbook* (pp. 330-343). Oxford, England: Pergamon Press.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1996 Technical Report,* NCES 1999-452, by Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). Washington, DC: National Center for Education Statistics.

Wright, B. D., and Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement, 1*, 83-106.

Wu, M. (2004). Plausible Values. *Rasch Measurement Transactions*, 2004, 18:2 p. 976-978

Wu, M.L., Adams, R.J. & Wilson, M.R. (1998). ACER Conquest: Generalised item response modelling software. Melbourne: ACER Press.

Wu, M. L. (1997). The Development and Application of a Fit Test for Use with Marginal Maximum Likelihood Estimation and Generalised Item Response Models. Unpublished Masters Dissertation. University of Melbourne.

Zoanetti, N. P. (2006). Applications of Item Response Theory to Identify and Account for Suspect Rater Data. Unpublished Masters Dissertation. University of Melbourne.

THE AUTHORS

Mr Nathan Zoanetti is a Research Fellow at the Assessment Research Centre, Faculty of Education, University of Melbourne.

Professor Patrick Griffin is Director of the Assessment Research Centre, Faculty of Education, University of Melbourne.

Professor Ray Adams is an educational measurement expert at the Assessment Research Centre, Faculty of Education, University of Melbourne.