Psychological Test and Assessment Modeling, Volume 55, 2013 (2), 162-180

Estimation of composite score classification accuracy using compound probability distributions

Chris Wheadon[1] *& Ian Stockford*[2]

Abstract

Presented is a demonstration of an intuitively simple, flexible and computationally inexpensive approach to estimating classification accuracy indices for composite score scales formed from the aggregation of performance on two or more assessments. This approach uses a two stage application of the polytomous extension of the Lord-Wingersky recursive algorithm and can be driven by any IRT model with desired simplicity or required complexity to best represent the properties of the tests. The approach is demonstrated using operational data from a high stakes mathematics qualification which is formed from two tests administered on distinct occasions. To provide the simplest representation of a test containing both dichotomous and polytomous items, the partial credit model is applied to model behaviour on the two tests. As an extension to this, a testlet model is applied to allow joint calibration of parameters from both tests. This model provides more information to the calibration process at the expense of some added computational complexity. Further to this, the potential application of this approach in the absence of operational data is investigated using a comparison of simulated data to the observed data.

Key words: Classification accuracy, IRT, composite scores

[1] *Correspondence concerning this article should be addressed to:* Chris Wheadon, PhD, Centre for Education Research and Policy, AQA, Stag Hill House, Guildford, Surrey, GU2 7XJ, UK; email: cwheadon@aqa.org.uk

[2] Centre for Education Research and Policy, AQA, Guildford, Surrey, UK

1 Introduction

The purpose of this paper is to present a new method for calculating the classification accuracy of composite scores. Wherever scores are reported as classifications such as pass / fail or grade A to grade E users of those scores have an interest in understanding how accurate those classification decisions are. Classification accuracy approaches provide an estimate of the accuracy of the grading through a comparison of the degree to which observed classifications agree with those based on examinees' true scores (Lee, Hanson, & Brennan, 2002; Livingston & Lewis, 1995). Composite score classification accuracy refers to the accuracy of classification when scores have been scaled or aggregated across multiple assessments (Livingston & Lewis, 1995). There are a number of benefits to understanding the extent of misclassification and the factors that influence it. These include being aware of the potential consequences when designing assessments (or combinations of assessments) used for a qualification, as part of assessment quality control monitoring processes, and also in educating users of qualification results in areas such as the over-interpretation of grades.

Many previous studies have considered classification accuracy for single assessments. Wheadon and Stockford (2011) presented an empirical evaluation of the classification accuracy and consistency of single assessments forming high stakes qualifications in England. This adopted both a Classical Test Theory (CTT) and Item Response Theory (IRT) approach as previously implemented in other assessment contexts by Livingston and Lewis (1995) (CTT) and Lee (2008) (IRT). For application of CTT approaches to classification accuracy see also Breyer and Lewis (1994), Hanson and Brennan (1990), Woodruff and Sawyer (1989), and Peng and Subkoviak (1980). In greater depth, Verstralen and Verhelst (1991) have investigated the consequences of applying different IRT based measurement models for item calibration and accuracy calculation in an item banking scheme, with Lee, Hanson, and Brennan (2002), Wang, Kolen, and Harris (2000), and Bramley and Dhawan (2010), considering further IRT based approaches at the test level.

Regarding articulations of classification accuracy at the composite score level, He (2009) considers the extensions available for more conventional reliability indicators; however, composite score classification accuracy has only been considered in a limited number of studies. Van Rijn, Verstralen, and Béguin (2009), Douglas and Mislevy (2010) and Chester (2003) looked at the consequences of different decision rules applied to classify candidates based on composite scores including consideration of the validity of the rules dependent on the content and aims of the assessment. Issues to be addressed when considering composite score data sets are the hierarchical and multidimensional nature of the items across separate assessments. Multidimensional IRT models (Reckase, 1997) offer a potential solution to the management of multiple assessment multiple trait scenarios; however, the additional model complexity introduced renders operationalization of these approaches challenging. Multidimensional IRT models allow the efficiency of assessment to be improved as estimations of performance on related constructs can draw strength from each other, as shown, for example, by Frey and Seitz (2011), but where an assessment is required, due to validity constraints, to sample from a given number of dimensions there is less to be gained from a complex modelling solution applied post-hoc.

This study seeks to demonstrate a simple, robust and intuitive analytical solution to estimating composite score classification accuracy. This approach applies no constraints on the simplicity or complexity of the model used to represent the constituent tests. The proposed approach is demonstrated on an operational data set and its application using simulated data is discussed to provide a preliminary insight into the appropriateness for use in instances of (partially) absent data.

2 Method

2.1 The Data

The operational data selected for consideration is that arising from an examination undertaken in England, specifically a GCSE Mathematics examination sat in summer 2011. This particular GCSE Mathematics qualification is composed of two tests sat on different occasions within a relatively short period of time (eight days in this instance). Both tests have a maximum mark of 100 and are composed of a mixture of dichotomous and polytomous items, as outlined in table 2.

Table 1:
Frequency of items with the quoted number of response categories

	Number of Items with Score Category *K*					Total Number of Items	Maximu m Mark
	K = 2	*K* = 3	*K* = 4	*K* = 5	*K* = 6		
Test 1	12	16	10	4	2	44	100
Test 2	13	6	13	9	0	41	100

When multiple tests are aggregated to qualification level, various methods can be applied to scale the test scores. For simplicity the approach used in this study was to sum candidates’ scores on each test and consider the accuracy of grading against these qualification level cut-scores. The approach can be easily extended, however, to complex non-linear scaling and aggregation approaches.

The number of candidates with valid marks entered for both tests was 17,957, however, for the purposes of practicality when fitting the models described below, a sub-set of 1,000 candidates was drawn at random from this population.

2.2 Measurement models

In order to estimate misclassification rates for test data it is necessary to select a theoretical model to represent candidate behaviour. To provide a probabilistic representation of candidate performance at the item level, IRT models have been applied here as described in the following sections.

2.2.1 Partial Credit Model

The partial credit model (PCM) (Masters, 1982) is an extension of the dichotomous Rasch model (Rasch, 1960) and allows representation of the probability of a candidate achieving a certain score category on a given item. For candidate *i*, with ability θ_i, responding to item *j*, which has K_j available score categories, this model can be expressed as:

$$\Pr\left(X_j = x_k \mid \theta_i, x_k > 0\right) = \frac{\exp\left\{\sum_{l=1}^{k}\left(\theta_i - \beta_{jl}\right)\right\}}{1 + \sum_{m=1}^{K_j}\exp\left\{\sum_{l=1}^{m}\left(\theta_i - \beta_{jl}\right)\right\}} \tag{1}$$

$$\Pr\left(X_j = 0 \mid \theta_i\right) = \frac{1}{1 + \sum_{m=1}^{K_j}\exp\left\{\sum_{l=1}^{m}\left(\theta_i - \beta_{jl}\right)\right\}}$$

where X_j is the candidate's achieved score on the item, x_k is the item level score available in category *k*, and β_{jk} is the k^{th} threshold location of item *j*. To provide the item parameters, β_{jk}, and the person parameters, θ_i, conditional maximum likelihood (CML) estimation can be used (Mair, Hatzinger, & Maier, 2010) resulting in a single most likely value of θ for each candidate and β for each item category.

As this model only specifies a single ability parameter for each candidate (as opposed to multidimensional IRT models where candidate ability is represented by a vector of abilities) this contains the implicit assumption that the items composing the test are measuring a single dimension. This represents the simplest IRT model to describe polytomous items and is used as the base model in this study with separate parameter estimation being performed for the two tests.

To simulate data using this model a simplified approach is taken. Rather than applying the full PCM, all items are modelled as being dichotomous therefore mirroring the Rasch model with both item and person parameters are drawn from a normal distribution with a mean of 0 and a standard deviation of 1.

2.2.2 The Testlet Model

High stakes qualifications are frequently composed of assessments in different modes or assessing diverse skills or content areas. Therefore, different assessments are measuring (or attempting to measure) a number of traits, hence, candidates' true scores are also likely to differ between assessments (but are likely to be positively correlated for any qualification that can make reasonable claims of validity).

A model which accommodates these differences in linked abilities is the testlet model. This model facilitates the analysis of a population of items which can be grouped into sub-populations due to some common property. Each sub-population of items which share this common property forms a testlet. This grouping of items was initially proposed in the context of computer adaptive testing by Wainer and Kiely (1987) to investigate whether the assumption of local independence of items was being compromised.

Within the testlet model, candidates are estimated ability parameters, θ_i, for the combined population of items along with a modifier of this ability for each testlet, termed a testlet propensity. For polytomous items, this model is provided by the re-expression of the equation provided by Li, Li, and Wang (2010) and unitisation of testlet and item discrimination parameters as:

$$\Pr\left(X_j = x_k \mid \theta_i\right) = \frac{\exp\left\{\sum_{l=1}^{k}\left(\theta_i - \beta_{jl} + \gamma_{id(j)}\right)\right\}}{\sum_{m=1}^{K_j}\exp\left\{\sum_{l=1}^{m}\left(\theta_i - \beta_{jl} + \gamma_{id(j)}\right)\right\}} \tag{2}$$

where $\gamma_{id(j)}$ is candidate *i*'s testlet propensity for testlet *d* containing item *j*.

To accommodate the increased model complexity compared with the PCM, the item, person and testlet parameters are estimated in a Bayesian framework via a Markov Chain Monte Carlo (MCMC) approach using Gibbs sampling. In contrast to providing a single set of most likely item and person parameters as is the case for CML, this numerical approach is executed a number of times resulting in a population of possible item, person and testlet parameters (Fox, 2010). Multiple runs of the Bayesian parameter estimation will, therefore, provide an indication of stability related to model fit.

2.3 Estimation of classification accuracy

Rates of grade misclassification are usually expressed as the inverse measure, classification accuracy. The classification accuracy for an individual candidate is defined as the probability that their observed score falls in the same grade classification as his or her true ability. In an IRT framework a candidate's test level true score (reflecting true ability) is defined as the sum of his or her expected item level scores, such that on a given test composed of J items, candidate *i* has a test true score, τ_i, defined as:

$$\tau_i = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\Pr\left(X_j = x_k \mid \theta_i\right)x_k = \sum_{j=1}^{J}E\left(X_j \mid \theta_i\right) \tag{3}$$

Two approaches can be used to estimate the candidate level classification accuracy, as described in the following sub-sections.

2.3.1 Numerical estimation of classification accuracy

Since IRT models provide a probabilistic representation of candidate behaviour it is trivial, although computationally expensive, to estimate classification accuracy statistics numerically. This can be achieved at the candidate level by performing a Monte Carlo simulation to generate candidates' 'observed' scores (at the test level, composite level, or both) based upon the combination of item and person parameters. The frequency of candidates' observed scores occurring in the same grade classification as his or her true score can then be summed and expressed as a proportion of the simulations run to provide an estimate of the candidate level classification accuracy.

Whilst this numerical approach is computationally expensive it is congruent with estimation in a Bayesian framework such as that applied for parameter estimation under the testlet model. Indeed, such an approach is proposed by Wainer, Bradlow, and Wang (2007) for estimation of classification accuracy at the aggregated testlet level which is equivalent to composite score when defining testlets in the manner described here. The numerical approach is difficult to use, however, when the scores from testlets are scaled before they are aggregated.

2.3.2 Analytical estimation of classification accuracy

As an alternative to the numerical approach, the probabilistic models can be extended to determine analytically the probability that a candidate with given θ_i will achieve each score on the test. Lord and Wingersky (1984) propose a recursive approach to calculating the probability that a candidate will achieve each score. This can be extended to manage polytomous items summarised as:

$$\Pr\left(Y_J \mid \theta_i\right) = \sum_{x=0}^{x_{K_J}} \Pr\left(Y_{J-1} = Y_J - x \mid \theta_i\right) \Pr\left(X_J = x \mid \theta_i\right), \qquad J > 1 \quad (4)$$

$$= \Pr\left(X_J = Y_J \mid \theta_i\right), \qquad J = 1 \quad (5)$$

where Y_J is the candidate's test level score on a test of length *J*, and the probability of a test score of zero is given by $\Pr\left(Y_J = 0 \mid \theta_i\right) = \prod_{j=1}^{J} \Pr\left(X_j = 0 \mid \theta_i\right)$. Since this recursive method relies on knowledge of the probability of achieving a test score of $Y_{i\bullet} - x$ it is therefore efficient for determining the value of $\Pr\left(Y_{i\bullet} \mid \theta_i\right)$ for all scores. Collection of these values for all possible values of $Y_{i\bullet}$ provides a conditional sum score probability distribution.

As proposed by Lee (2008), for a given test, the probability that a candidate is correctly classified can then be determined by integrating this conditional sum score probability distribution between the cut-scores that surround the candidate's true score. The candidate level classification accuracy, I_{CA}, is therefore defined as:

$$I_{CA} = \Pr\left(C\left(Y_{i\bullet}\right) = C\left(\tau_i\right)\right) = \sum_{y=Y_{L,i}}^{Y_{U,i}-1} \Pr\left(Y_{i\bullet} = y\right) \qquad (6)$$

where $C\left(a\right)$ is the grade classification based on a test score of a, and $Y_{U,i}$ and $Y_{L,i}$ are the cut-scores above and below the candidate's true score, respectively.

Proposed here is the extension of this approach to combine test level conditional sum score probability distributions, using the Lord-Wingersky recursive algorithm, to provide a conditional sum score probability distribution based on composite scores which can be applied to equation 6 using the qualification level cut-scores. This provides an estimate of a candidate's qualification level classification accuracy. To achieve this, equations 4 and 5 can be re-expressed as:

$$\Pr\left(Z_{N_T} \mid \mathbf{\Theta}_i\right)=\sum_{y=0}^{\hat{Y}_{N_T}}\Pr\left(Z_{N_T-1}=Z_{N_T}-y \mid \mathbf{\Theta}_i\right)\Pr\left(Y_{N_T}=y \mid \mathbf{\Theta}_i\right), \qquad N_T > 1 \quad (7)$$

$$=\Pr\left(Y_J{}'=y \mid \mathbf{\Theta}_i\right), \qquad N_T = 1 \quad (8)$$

where Z_{N_T} is the candidate level composite score arising from aggregation of N_T tests, $\hat{Y}_a$ is the maximum scaled test score on test *a*, $\mathbf{\Theta}_i$ is the vector of test level ability parameters for candidate *i* and $\Pr\left(Z_{N_T}=0 \mid \mathbf{\Theta}_i\right)=\prod_{n=1}^{N_T}\Pr\left(Y_n=0 \mid \mathbf{\Theta}_i\right)$.

Using this analytical approach it is relatively easy, once probabilities have been derived at the test level, to incorporate linear and non-linear scaling into the process of estimating classification accuracy at the composite score level.

For clarity, the steps of the proposed procedure are:

1. Fit an appropriate IRT model to the tests (be that separate estimation at the test level using the PCM, joint parameter estimation using the testlet model, or otherwise).
2. Apply the Lord-Wingersky recursive algorithm at the test level (equations 4 and 5) for each candidate based on his or her probability distributions of scoring each category on each item. This results in a test level conditional sum score probability distribution for each candidate.
3. If of interest, apply equation 6 to these conditional sum score probability distributions, to determine I_{CA} at the test level.
4. Scale the test level conditional sum score probability distribution using the required transformation (be that linear or non-linear).
5. Reapply the Lord-Wingersky recursive algorithm at the qualification level (equations 7 and 8) for each candidate based on their scaled test level conditional sum score probability distributions. This provides a qualification level conditional sum score probability distribution for each candidate.
6. Apply equation 6 to this conditional sum score probability distribution using the qualification level cut-scores to determine I_{CA} at the qualification level.

It should be noted use of the Lord-Wingersky algorithm depends on the assumption that the conditional distributions of the item scores are independent of each other. This assumption will not hold for composite scores when models have been fitted separately to each test: the conditional probability of achieving a score on one test given a score on the other test would give a more accurate estimation of the likelihood of classification accuracy than the marginal probabilities used in the process described here.

2.3.3 Models fitted

Three models are fitted:

1. The partial credit model is fitted to each test separately, the Wingersky-Lord algorithm is applied to the model parameters derived from each test separately, and then combined once again using the Wingersky-Lord algorithm. This is the simplest application of the procedure suggested here, but is subject to the limitations concerning conditional independence as described above resulting in degradation of the classification accuracy estimation.
2. The partial credit model is fitted to the combined sets of items from both tests and the Wingersky-Lord algorithm applied to the single set of model parameters. While this approach would not be generally recommended it allows the degradation of the estimation from loss of conditional independence in the separate estimation procedure to be evaluated and is valid here due to the highly correlated nature of the test scores composing the composite score.
3. The testlet model is fitted across both tests, with each test representing one testlet. The person, item and testlet parameters are then used to define separate conditional sum score distributions for each test. The two sets of probabilities are then combined using the Wingersky-Lord algorithm. This approach allows the estimation of the model parameters to benefit from joint estimation across both tests. This approach should yield a more accurate estimation than the separate estimation of partial credit models and does not suffer the loss of conditional independence to which model 1 is subjected.

2.3.4 Classification accuracy summary statistics

Regardless of both the approach used to estimate the classification accuracy and level of the hierarchy at which it is expressed, I_{CA} is available at the individual candidate level. Collectively, this provides a rich representation of how classification accuracy varies with different candidate properties (usually plotted against candidate true score). However, for many applications such as routine quality monitoring, the definition of a single summary statistic is potentially beneficial for manageability and interpretability.

The summary statistic applied here is that proposed by Lee (2008) which takes the mean of the candidate level classification accuracies. This statistic can be interpreted as the probability that a candidate selected at random from the cohort will be accurately classified. Whilst this provides an intuitive measure it should be borne in mind if using this measure for quality monitoring purposes that this measure is heavily dependent on the distribution of candidates across the mark range. This measure reflects as much about the properties of the cohort as it does about the underlying assessment (Wheadon & Stockford, 2011).

2.4 Software

The analyses described throughout this work have been implemented in R (R Development Core Team, 2011). The PCM model and accompanying CML is implemented using the eRm package (Mair, Hatzinger, & Maier, 2010). The MCMC and Gibbs sampling procedure applied when estimating parameters under the testlet model was performed using JAGS (Plummer, 2012) accessed via the R2jags (Su & Yajima, 2011) R package when analysing the operational data. Due to its improved handling of missing data, this estimation used WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) for the production and estimation of simulated data due to its robustness in the presence of missing data at the expense of a degree of computational speed. The testlet model specification was taken from Curtis (2010). In order to support further research in this area the authors have developed the R package *classify* (Wheadon & Stockford, 2012).

3 Results

3.1 Descriptive statistics

Before considering the classification accuracy estimates it is important to examine the descriptive statistics to provide some context for the later analyses. The test level descriptive statistics for the operational data are presented in Table 2. The Cronbach's alpha and average item to test correlation are high suggesting each test comprises a coherent scale. Both tests show only a slight positive skew with mean marks around 50% suggesting that the tests are appropriately targeted at the cohort. The correlation between candidates' scores on the tests is high, which would suggest that a single trait is being measured across both tests, and that the hierarchical structure within the data has minimal effect. Indeed, the disattenuated correlation, which approaches a value of 1, is highly suggestive of a single dimension being assessed. In spite of this apparent unidimensionality, there is still, at least, a strong theoretical case for the use of a testlet model in this specific case as the tests are sets of items that are designed to be administered separately and are likely to be taught as coherent courses in separation. The consequences of fitting the testlet model to this highly coherent data set are evaluated in the next section.

Table 2:
Descriptive Statistics

	Mean	SD	Max	Cronbach's Alpha	Skew	Kurtosis	Average Item to Test Correlation	Inter-Unit Correlation
Test 1	56.88	20.07	100	0.93	0.10	-0.84	0.22	0.92
Test 2	54.94	22.82	100	0.94	0.06	-1.10	0.26	0.92

3.2 The testlet effect

To evaluate the magnitude of the testlet effect for the cohort as a whole, the ratio of the variation in abilities can be compared to the variation introduced by the testlet effect defined as:

$$\frac{\sigma_\theta^2}{\sigma_\theta^2 + \gamma_{d(j)}^2} \tag{9}$$

where $\gamma_{d(j)}^2$ is the variance associated with the testlet parameter. For the operational data set presented, this value is 0.11 suggesting a low level of local dependence within the testlets. Further, around 11% of candidates have a gamma value that does not intersect with zero within one standard deviation (Figure 1). While there does therefore appear to be a measurable testlet effect it would seem unlikely that the testlet model would perform considerably better than a model which neglects this hierarchy for the purposes of estimating classification accuracy in this instance. Additional value is, however, added under the testlet model since information is combined from both tests during parameter estimation. The analogous non-hierarchical approach is to apply the PCM model to all items combined across both tests (as previously specified as model 2). This approach is applied in this case to examine the degradation due to loss of conditional dependence; but it is only potentially viable with this highly coherent data set as it would violate assumptions of the model in the majority of cases.

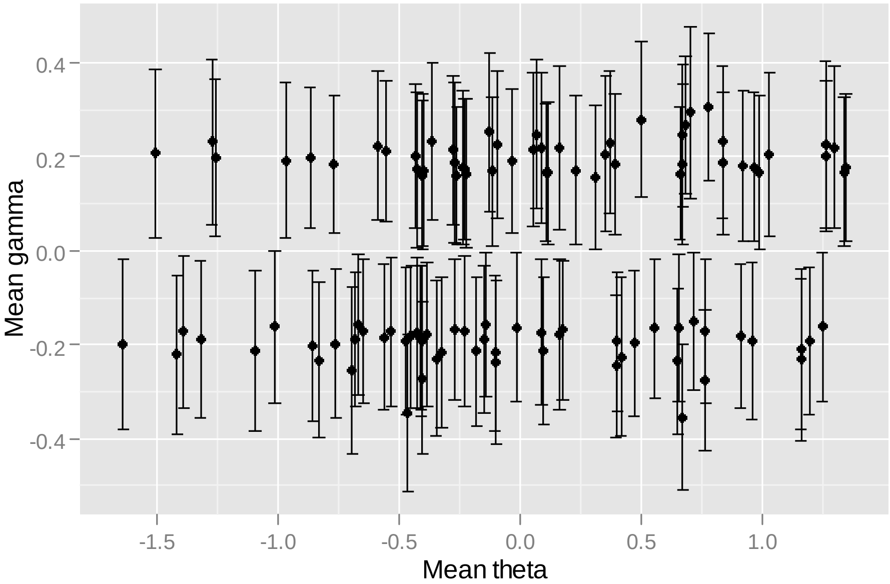


Figure 1:
Mean and standard deviation of gamma values

3.3 Model fit of observed data

To further evaluate the appropriateness of the applied models it is necessary to establish how well the models fit the data. As measures of classification accuracy are based on the cumulative information yielded by all item information, and the distribution of score probabilities compared to the grade boundaries, the most important measure of fit appears to be the comparison of the observed and expected score distribution. This predicted observed score distribution is defined as the composite conditional sum score distributions, provided by equations 7 and 8, summed across all candidates (Hanson & Béguin, 2002).

Therefore, the IRT models were fitted to the observed data and the estimated frequency distributions compared with the observed distribution. As can be seen from Figure 2, the estimated distribution from the PCM intersected the multiple models produced by the Bayesian fit performed under the testlet model. Critically, the expected score distributions under both models appears to follow the observed distribution suggesting good model fit at the test level in both cases.

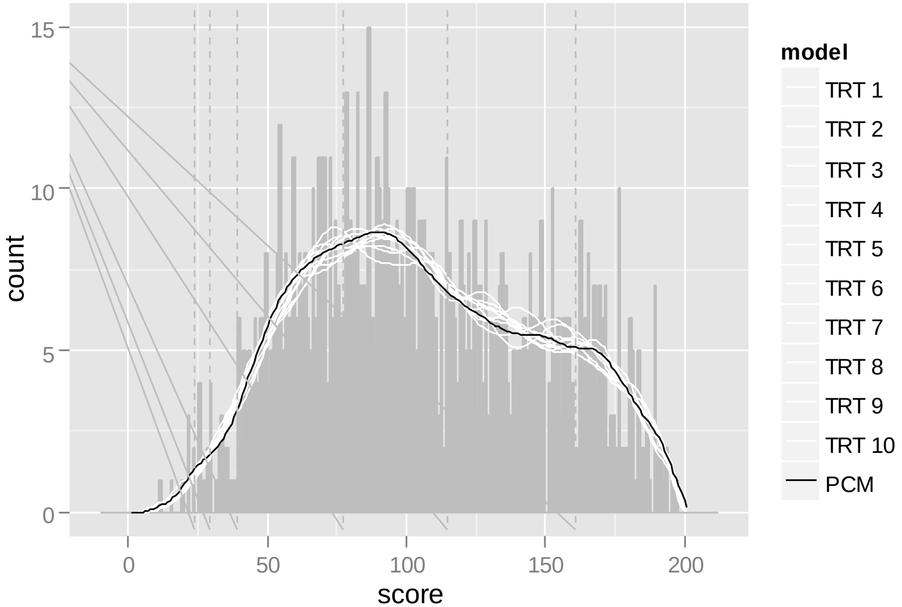


Figure 2:
Observed and expected score distributions with grade boundaries super-imposed

3.4 Classification accuracy

The candidate level composite score classification accuracy values are illustrated in Figure 3 (including data sets using simulated item parameters for later reference). All of the models follow the same typical shape, with the lowest values at the grade boundaries which are fundamentally limited to a maximum of 0.5. As can be seen from Figure 4, the differences between the testlet model and the PCM models are small apart from score points directly around the grade boundaries, with the largest differences occurring around the narrowest boundaries. This is due to any differences between the models being accentuated by the narrow boundaries where sensitivity to misclassification is greatest. Since the testlet propensities are small, the differences between the models are likely to be due to the combination of information across tests due to joint estimation under the testlet model and the constraint of item parameters to a common scale resulting in differing item level fit.

The summary classification accuracy indices for the two individual tests are around 0.79 under both models. Whilst the accuracy with which candidates are classified at the test level is not of primary concern here it is worth noting that these values are higher than any measured in Wheadon and Stockford (2011). From consideration of the descriptive statistics, this is unsurprising given that both tests have considerably higher mean grade boundary separations (13.8 marks and 13.6 marks, respectively) than any considered as part of the previous study.

The composite score classification indices for the different models are presented in Table 3, including the values for the simulated data sets for later reference. Due to a combination of the increase in measurement information provided by multiple tests and the increased separation of subject level grade boundaries (27.4 marks) over those found at test level, the composite score classification index values increase to around 0.85. These tests benefit from long raw mark scales which allow clear differentiation of ability and wide spacing of grade boundaries.

Virtually no difference is apparent between the joint estimation of the partial credit model and the separate estimation of the partial credit model. Differences would be due to the loss of conditional independence between the test scores on the two separate tests when the model parameters are estimated separately. As the tests are highly correlated, the degradation represents a worst case scenario. This shows that the consequences of violating this assumption may be minimal when estimating the classification accuracy summary statistic.

Table 3:
Classification accuracy under different models

	Operational Data			Simulated β Parameters	
	PCM (Model 1)	PCM with JE (Model 2)	Testlet (Model 3)	Rasch	Testlet
Mean Classification Accuracy	0.853	0.845	0.850	0.878	0.837
SD of Classification Accuracy	-	0.013	0.003	-	0.010

Figure 3:
Classification accuracy under different models

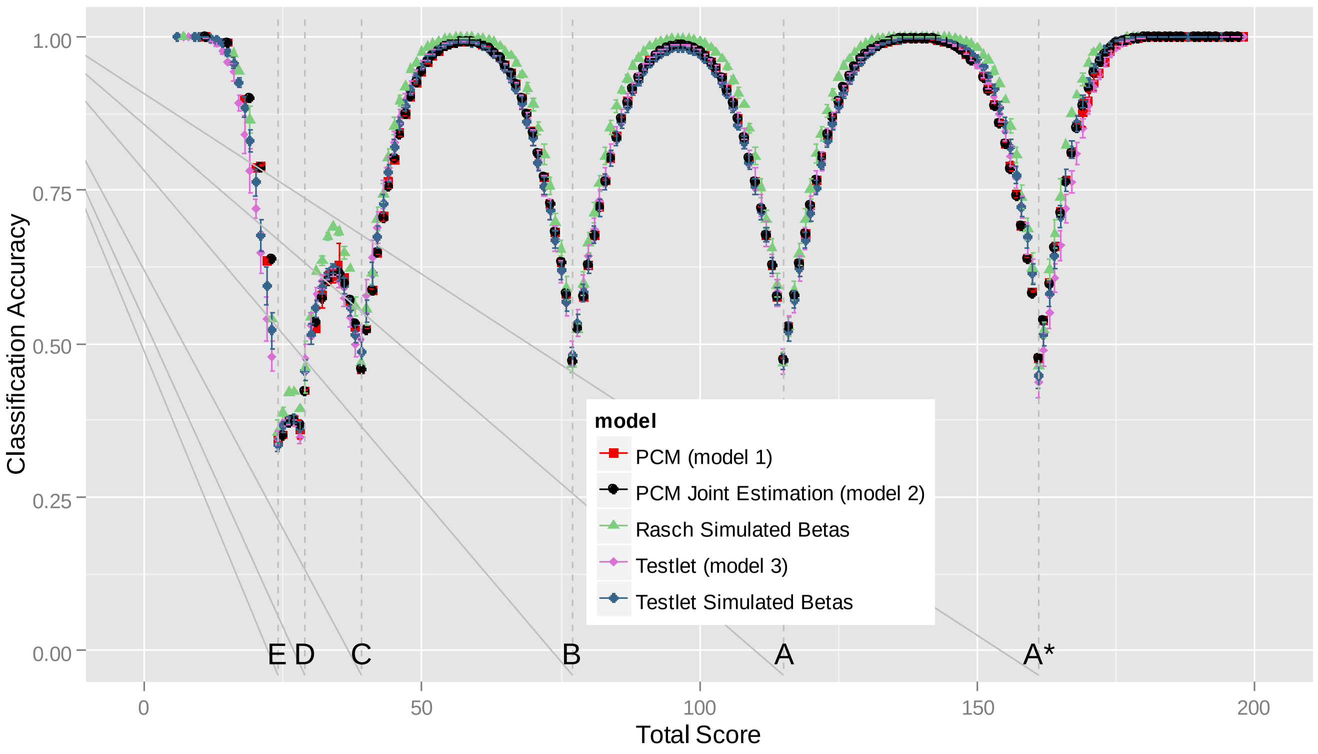
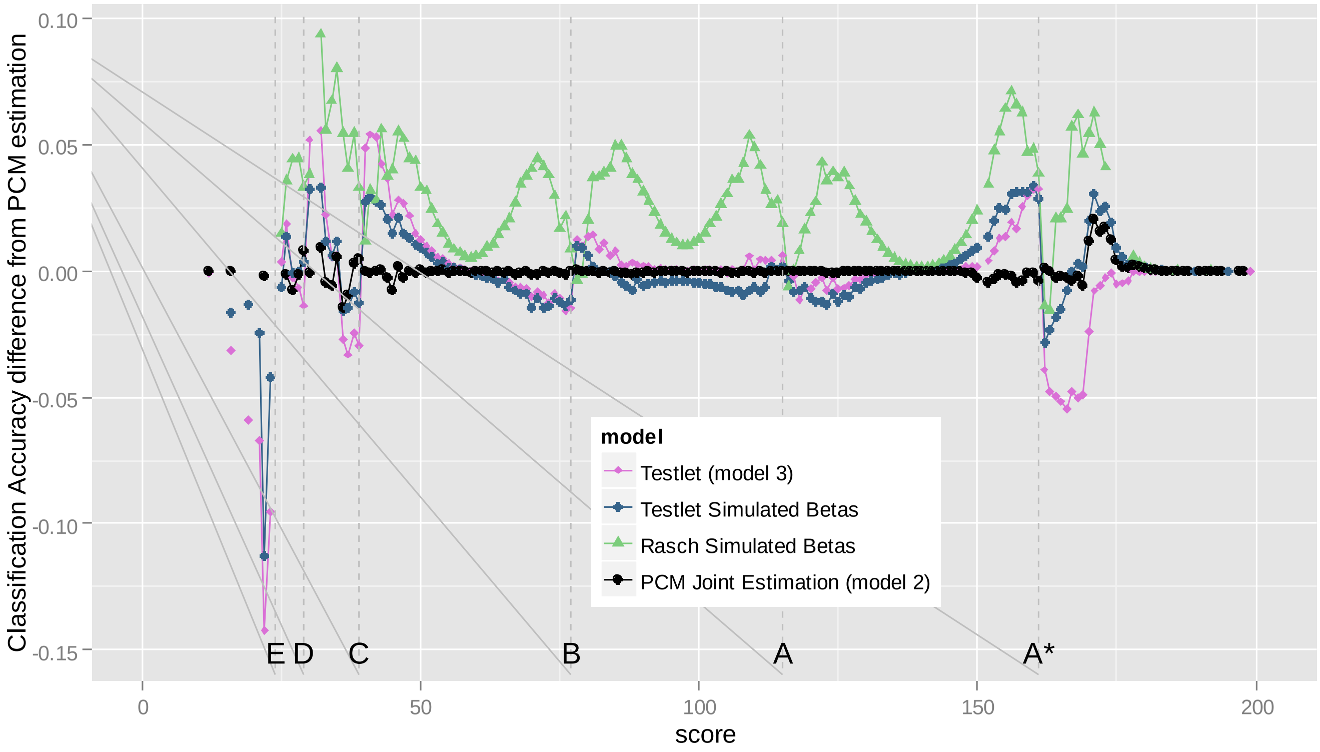


Figure 4:
Difference in classification accuracy from PCM estimation



3.5 Simulated data sets

3.5.1 Model fit of simulated data

While models of the operational data showed good fit to the observed score distributions, the entirely simulated models showed a poor fit to both the observed distribution and to each other, as shown in Figure 5. However, since these person and item parameters were drawn from arbitrary (yet reasonable) distributions of parameters, this is not altogether surprising and these differences are likely to be due to the distributions of simulated parameters being poorly matched to the operational data.

To investigate improvements to the accuracy of the simulation, the effects of fixing either the person or item parameters was considered. Since it is more likely that the distribution of person parameters can be estimated from performance elsewhere, the values of θ were constrained to match those arising from the operational data. This gives rise to the estimated composite score distributions given in Figure 6. As expected, this yielded more satisfactory fit to the observed distribution for both models, although the data simulated under the testlet model seems to provide a better fit than the data simulated under the Rasch model.

3.5.2 Classification accuracy for simulated data

In addition to the operational data sets the classification accuracy plots for the simulated data with constrained person parameters are shown in Figures 3 and 4. It should be noted

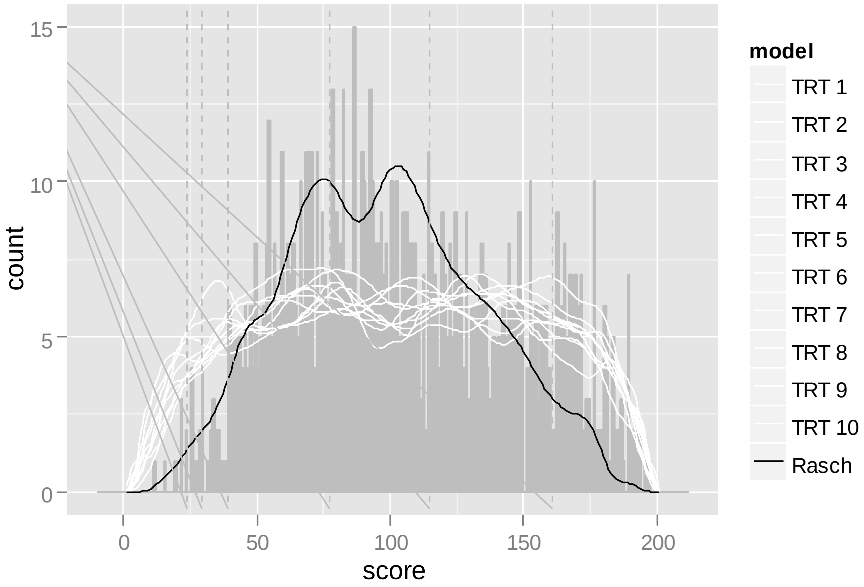


Figure 5:
Observed and simulated expected score distributions

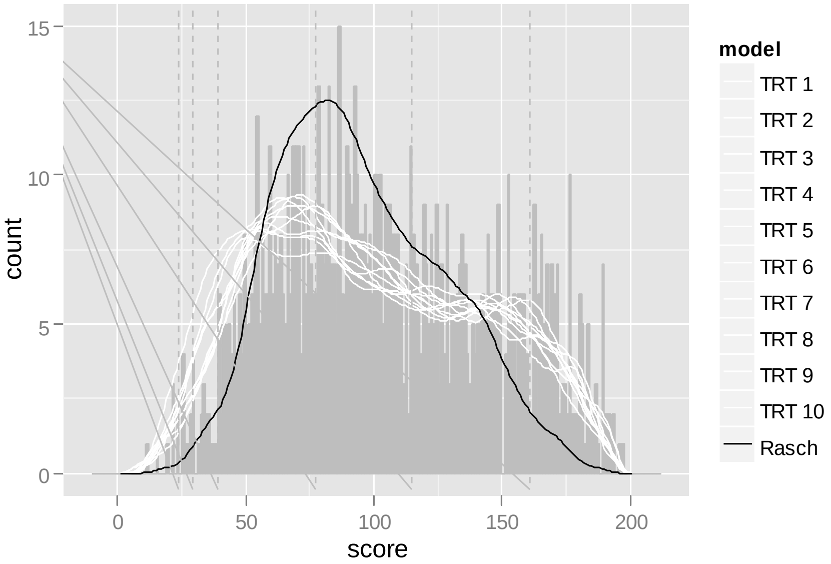


Figure 6:
Observed and simulated expected score distributions with constrained θ and γ values

that, for the testlet model, the shape of the relationship between classification accuracy and composite score is largely the same for the simulated and the observed data. This suggests that the difference in summary statistic (presented in Table 3) largely occurs due to the differing distributions of candidates across the composite mark scale rather than differences in the underlying models.

Figure 3 shows the estimates of classification accuracy are higher for the simulated data set using the Rasch model than the PCM model for all composite score values other than those on the A* boundary. This is because the item parameters are drawn from a normal distribution with a mean of 0 and a standard deviation of 1 and do not reflect the item structure or the parameter values of the observed data. The differences are minimal, however, suggesting that it may be possible to generate reasonable estimations of classification accuracy at individual points on the composite score scale. The accuracy of estimates of summary statistics, however, remains dependent on knowledge of unknown, but not wholly unpredictable, population densities.

4 Discussion

This paper has shown how the work of Lee (2008) on the estimation of classification accuracy measures for single tests can be extended to estimate classification accuracy for scores comprised of discrete tests by using a two stage application of Lord and Wingersky’s (1987) recursion formula. The approach is appealing through its intuitive use of the probability distributions yielded by potentially simple IRT models which replaces the need to model any multidimensionality or hierarchical structure which may exist between tests. Provided the model used to represent each test is locally valid, the test level probability distributions to which the model information is distilled can be applied to the process outlined here. Furthermore, it is relatively easy to apply linear or non-linear scaling to the scores once the probabilities have been derived; the scaling process is problematic for other composite score classification procedures.

The simplicity of the proposed model comes at the expense of some reduction in test information supplied to the fitting process if parameter estimation is undertaken independently at the test level. As the probabilities derived from each test are not independent, there is also likely to be some further degradation of the classification accuracy estimates. However, the results presented here suggest the degradation may be minimal.

Where data is not available, the paper has also shown how simulated item parameters can yield reasonably accurate values of classification accuracy along a score scale. How well item and person information can be predicted more generally, however, is an empirical question that could be worth further investigation in particular contexts.

The present study was limited to a consideration of two highly correlated tests. The more closely correlated the tests the more the estimates of composite score classification accuracy will be degraded due to loss of local independence between the test scores. Further work is required to demonstrate the robustness of the proposed approach with data sets with varying degrees of multidimensional hierarchical effects and comparison to classification accuracy estimates derived from other modelling solutions.

The derivation of a simple, mathematically appealing approach to the calculation of classification accuracy for scores derived from multiple tests opens up a range of further research opportunities. The approach could be used, for example, to determine the relative strengths and weaknesses of different scaling and aggregation schemes. Most importantly, however, the simplicity of the approach means that the estimation of classification accuracy at the composite score level could become a routine part of qualification quality measures as opposed to a research activity in itself.

References

AQA. (2011). *Uniform marks in A-level and GCSE exams and points in the Diploma*. Manchester: Assessment and Qualifications Alliance.

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications*. Coventry, UK: Office of Qualifications and Examinations Regulation.

Breyer, F., & Lewis, C. (1994). *Pass-fail reliability for test with grade boundaries: a simplified method* (ETS Research Rep. No. 94-39). Princeton, NJ: Educational Testing Service.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: a framework for combining measures. *Educational Measurement: Issues and Practice, 22,* 32-41.

Curtis, S. M. (2010). BUGS Code for Item Response Theory. *Journal of Statistical Software, Code Snippets, 36*(1),1-34.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioural Statistics, 35*(3), 280-306.

Fox, J.-P. (2010). *Bayesian Item Response Modeling*. New York: Springer.

Frey, A., & Seitz, N.-N. (2011). Hypothetical Use of Multidimensional Adaptive Testing for the Assessment of Student Achievement in the Programme for International Student Assessment. *Educational and Psychological Measurement*, *71*(3), 503–522.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26,* 3-24.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

He, Q. (2009). *Estimating the reliability of composite scores*. Coventry, UK: Office of Qualifications and Examinations Regulation.

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory.* (No. 27) CASMA Research Report. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistence and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*(4), 412-432.

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (Research Report). Princeton, New Jersey: Educational Testing Service.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on true scores. *Journal of Educational Measurement, 32*(2), 179-197.

Lord, F., & Wingersky, M. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings”. *Applied Psychological Measurement, 8*, 452-461.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10,* 325-337.

Mair, P., Hatzinger, R. & Maier, M. (2010). eRm: Extended Rasch Modeling. R package version 0.13-0, http://CRAN.R-project.org/package=eRm

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Office of Qualifications and Examinations Regulation. (2012). *Ofqual’s Reliability Compendium.* Coventry, UK: Office of Qualifications and Examinations Regulation.

Opposs, D., & He, Q. (2011). *The reliability programme: final report*. Coventry, UK: Office of Qualifications and Examinations Regulation.

Peng, C., & Subkoviak, M. J. (1980). A note on Suynh’s normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

Plummer, M. (2012). Just Another Gibbs Sampler, version 3.2.0 http://mcmc-jags.sourceforge.net/

R Development Core Team. (2012). R: A Language and Environment for Statistical Computing, http://www.R-project.org

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Su, Y-S., & Yajima, M. (2011). R2jags: A Package for Running jags from R, R package version 0.02-15, http://CRAN.R-project.org/package=R2jags

van Rijn, P., Verstralen, H., & Béguin, A. A. (2009). *Classification accuracy of multiple-test based decisions using item response theory.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Verstralen, H. H. F. M., & Verhelst, N. D. (1991). *Decision accuracy in IRT models* (Measurement and Research Department Report 91-7). Arnhem: CITO.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge: Cambridge University Press.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*(2), 141-162.

Wheadon, C., & Stockford, I. (2012). classify: A package for generating IRT-based classification accuracy and consistency statistics, R package version 0.1, http: CRAN.R-project.org/package=classify

Wheadon, C., & Stockford, I. (2011). *Classification accuracy and consistency in GCSE and A level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009*. Coventry, UK: Office of Qualification and Examinations Regulation.

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail half-tests. *Applied Psychological Measurement, 13*, 33-43.