

Performance of Ability Estimation Methods
for Writing Assessments under Conditions
of Multidimensionality

Jason L. Meyers

Ahmet Turhan

Steven J. Fitzpatrick

Pearson

Paper presented at the annual meeting of the
National Council on Measurement in Education

May 2010



*Using assessment
and research to
promote learning*

Abstract

An increasing number of large scale assessments contain constructed response items such as essays for the advantages they offer over traditional multiple-choice measures. Writing assessments in particular often contain a mixture of multiple-choice and essay items. These mixed-format assessments pose many technical challenges for psychometricians. This study directly builds upon the Meyers et al. (2009) study by investigating how ability estimation, essay scoring approach, measurement model, and proportion of points allocated to multiple choice items and the essay item on mixed-format assessments interact to recover ability and item parameter estimates under different degrees of multidimensionality. Results indicate that when high levels of multidimensionality are present in the data, using separate estimation of multiple choice and essay ability, analytic scoring procedures, and the 1PL family of models produces more accurate results.

Introduction

Large scale writing assessments typically contain constructed response items such as essay items. Assessments containing these item types are becoming increasingly popular (Aschbacher, 1991) and offer several advantages over traditional multiple-choice measures (Bennett & Ward, 1993). These advantages, however, do not come without a cost. Compared to multiple-choice assessments, essay assessments usually assess a narrower domain of content (Hoover & Bray, 1995) and the test scores do not generalize as well as multiple choice assessments (Nichols, 2005). In addition, essay responses are expensive and time consuming to score (Hardy 1996; Khattri & Sweet, 1996; Parkes, 2000), while multiple choice items can be scored quickly and cheaply thus providing efficient information (Wainer & Thissen, 1993). Writing assessments comprised solely of multiple choice items also lend themselves well to linking designs that allow for equating different forms of the test over time (Ercikan et al. 1998). To take advantage of the benefits that both multiple choice (MC) and essay items provide, writing assessments often contain a mixture of multiple choice and essay items.

Mixed-format assessments pose technical challenges for psychometricians attempting to calibrate, scale, and equate these types of assessments. There are many choices to be made when deciding how to create scores from mixed format assessments. Issues such as how the different item types should be calibrated (together or separately) and how student abilities should be estimated (one overall estimate, or separate estimates based on the essay and the multiple choice items) present many questions, and the impact of different psychometric decisions is not well understood. In addition to these technical concerns, many educators and measurement professionals are concerned that essay items and multiple choice items may measure different

skills, particularly in the assessment of writing (Um et. al, 2007). In fact, there is some evidence to support that essay and multiple choice items measure different constructs (Traub, 1993).

IRT models have been shown to be robust to violations of unidimensionality (Drasgow & Parsons, 1983; Harrison, 1986; Reckase, 1979), but most of the previous research has investigated subjects in which the tests either contain one type of item or in which the various item types are intended to measure the same underlying construct. Much of the research on dimensionality of mixed-format tests has involved subjects other than writing in which multiple short answer items are used instead of an extended essay (Sykes, Hou, Hanson & Zhen, 2004; Bouwmeester & Sijtsma, 2004). In writing assessments, the essays are not only thought to measure different constructs than the multiple choice items, they are expected to. In other words, an argument could be made that some degree of multidimensionality is assumed, if not *desired*.

Many of the psychometric questions surrounding calibration methods for mixed-format assessments have been addressed in recent years. Ercikan et al. (1998) compared calibrating multiple choice and constructed response (CR) items together versus calibrating them separately and found that separate calibrations resulted in higher discrimination parameters and consistently higher information for the CR items. However, the loss of information from simultaneous calibration of the CR items was negligible, and simultaneous calibrations did not lead to higher local item dependence or model fit problems. Um et. al (2007) built upon this research by comparing calibration approaches under different conditions of multidimensionality and found the multiple choice items were relatively unaffected by the calibration method and results for all methods were worse in cases with strong evidence of multidimensionality. Finally, Meyers et. al (2009) extended this line of research yet again to compare both calibration methods and ability estimation procedures under differing degrees of multidimensionality. From a calibration

standpoint, their results indicated that the method chosen had no tangible impact on the ability to recover item or person parameters.

While calibration methods for writing assessments with multidimensional writing tests have been thoroughly studied in recent years, questions still exist regarding the performance of ability estimation procedures in such testing scenarios. Meyers et al. (2009) provided some guidance in understanding the functioning of different ability estimation procedures, but much remains unknown. They found that both person and item parameter estimates were impacted by an interaction of theta estimation procedure and multidimensionality and concluded that even in the presence of multidimensionality, unidimensional IRT models could be applied to writing assessments in such a way as to accurately estimate both person abilities and item parameters.

However, despite the significance of their findings, the generalizability of the results from the Meyers et al. (2009) study is unclear. Their study was not designed in a way which permitted a direct comparison of the performance of the 1PL and 3PL models, it only evaluated one testing scenario with a particularly large number of multiple choices items and it only evaluated the specific situation in which the essay item was holistically scored using a 12 point scale, which represents more score categories than is typically seen on statewide K-12 assessments.

Objectives/Purpose

The purpose of the current study is to directly build upon and extend the work done in the Meyers et al. (2009) study. Specifically, this study facilitates a direct comparison of performance of the 1PL and 3PL models. In addition, this study evaluates the performance of different ability estimation techniques under varying degrees of multidimensionality with writing assessments that:

- contain varying numbers of multiple choice items in relation to essay items, and
- contain essays scored using holistic versus analytic trait-based approaches.

Data and Methods

Data Generation

This study utilized operational item responses from one state's large scale high school End-of-Course writing assessment. These data contained over 35,000 student responses. The assessment is comprised of 30 multiple choice items and one essay item scored on three different components- Composition, Expression, and Mechanics. Each component is scored on a 1-4 scale by two raters, whose scores are then summed, yielding three analytic essay scores ranging from 2-8.

MULTILOG (Thissen, 2000) was used to calibrate the item responses under two different item response theory models. In one set of conditions, the One Parameter Logistic Model (Rasch, 1960) was used to estimate the multiple choice items parameters and the Partial Credit Model (Masters, 1982) was used to estimate the essay item parameters and in the other set of conditions the Three Parameter Logistic Model (Lord & Novick, 1968; Lord, 1980) was fit for the multiple choice items and the Two Parameter Generalized Partial Credit Model (Muraki, 1992) was fit for the essay item. The result of these calibrations was two sets of item parameters, one for the 1PL conditions and one for the 3PL conditions, that would serve as the basis for the simulation study.

For each study condition, 20,000 simulees were generated. For each simulee, two known ability estimates (theta values) were generated from a normal distribution with a mean of 0 and a standard deviation of 1, one representing ability on the multiple choice items and one representing ability on the essay item. Response strings for each simulee were generated using

the simulated ability estimates and the appropriate “true” item parameters (1PL or 3PL, depending on condition) using the SAS macro IRTGEN (Whittaker et. al., 2003).

a. Calibration Procedures.

Because the previous research (e.g., Meyers et. al 2009) indicated that the calibration method has no tangible effect on recovery of either the person or item parameters, in this study all conditions utilized a concurrent/simultaneous calibration approach where all multiple choice and essay item parameters were estimated in one run.

b. Ability Estimation Procedures.

Mimicking what was done in the Meyers et al. (2009) study, two ability estimation procedures were investigated in this paper. In the first procedure, one overall ability estimate was computed for each simulee based on performance on all multiple choice and essay items. In the second procedure, a multiple choice theta and an essay theta were computed separately and the overall theta value was defined as the simple average between the two thetas. MULTILOG (Thissen, 2000) was used for both item and person parameter estimation.

c. Item Response Theory Models.

Two sets of conditions were compared, 1PL and 3PL, such that the generating and estimating models corresponded directly. In the conditions in which the data were generated from the 1PL (Rasch, 1960) and Partial Credit Model (Masters, 1982), those same models were used to estimate item and person parameters. The corresponding models from the 3PL were used to estimate item and person parameters in the conditions using data generated from the 3PL.

d. Degree of Multidimensionality.

To simulate varying degrees of multidimensionality, three conditions of correlation, 0.30, 0.50, and 0.80 were created between known multiple choice and essay thetas. To represent the

baseline condition of pure unidimensionality, conditions in which the multiple choice and essay thetas were perfectly correlated (i.e., one known theta per simulee) were created as well. A review of the correlation between raw multiple choice and essay scores from the writing assessments in four states, including the state analyzed here, indicated that these correlations generally range from about 0.40 to 0.60. The correlation matrices in each condition were generated using the MVN (Uebersax, 2006) software package for generating random multivariate normal numbers.

e. Number of multiple choice items.

Typically, statewide K-12 writing assessments that contain multiple choice items have between 20 and 40 multiple choice items. To reflect these practices and evaluate the impact on the outcomes in this study, three conditions were simulated containing different numbers of multiple choice items : 20, 28, and 36. In order to vary the number of multiple choice items, items were randomly dropped or repeated from the original set of 30. Specifically, in the 20 item conditions, 10 items were randomly selected and excluded from data generation, in the 28 item conditions, two items were randomly excluded, and in the 36 item condition, 6 items were randomly repeated and added to the original set of 30 items.

f. Essay Scoring Approach.

Two basic types of essay scoring approaches are utilized operationally: holistic and analytic. Two sets of conditions were simulated to reflect these operational practices. In the analytic or multi-trait approach, 3 scores (ranging from 2-8) for each essay were generated and estimated, reflecting performance on the different dimensions (Composition, Expression, and Mechanics) utilized by the testing program modeled in this study. In the holistic approach conditions, a “pseudo-holistic” essay score was generated as the average of the three component

scores. The characteristics of the pseudo-holistic essay score were evaluated and compared to those of true holistic essay scores observed in other large-scale K-12 testing programs. The pseudo-holistic essay score was found to behave like a true holistic essay score in terms of score distribution and discrimination. This pseudo-holistic essay score ranged from 2-8.

This simulation study contained 2 (estimation methods) * 2 (IRT models) * 4 (degree of multidimensionality) * 3(number of MC items) * 2 (scoring approaches) = 96 conditions. Figure 1 depicts the various simulation conditions.

Insert Figure 1 Here

Evaluation criteria

100 replications were conducted for each condition using SAS version 9.2 (SAS Institute, 2002). The main outcome of interest in this simulation was the correlation between the estimated and true theta values across simulees and conditions. In addition, two widely used goodness-of-recovery (GOR) indices were calculated to evaluate the accuracy of the item and person parameter recovery (Maris, 1999). The first measure, BIAS, presents the difference between the average estimated parameter value and true parameter value,

$$BIAS(\tau_j) = \frac{\sum_{r=1}^R \hat{\tau}_{jr}}{R} - \tau_j, \tag{1}$$

where τ_j is the true value of parameter j and $\hat{\tau}_{jr}$ is the estimate of parameter j for a replicated data set r ($r=1, \dots, R$).

The second GOR index evaluated was the root mean square difference, RMSD, which is defined as the square root of the average squared differences between the estimated and true parameter values,

$$RMSD(\tau_j) = \sqrt{\frac{\sum_{r=1}^R (\hat{\tau}_{jr} - \tau_j)^2}{R}}, \quad (2)$$

using the same notation as the BIAS equation. GOR measures were aggregated over the replications such that only means of BIAS and RMSD were reported for each condition using the following equation:

$$\sum_{i=1}^I BIAS(\tau_i) / I, \text{ and} \quad (3)$$

$$\sum_{i=1}^I RMSD(\tau_i) / I \quad (4)$$

where R is the number of items or persons depending on the reported parameter and I is the number of replications.

Results

The following pages summarize the results from the current study. Results are broken down into two major categories: ability estimation results and item parameter results. Each section addresses:

- whether the results are consistent with the results of the previous study (Meyers et al. 2009),
- whether the added study conditions impacted the previously observed patterns, and
- any new patterns observed that were not noted in the previous study.

Theta Estimation

Table 1 presents the correlational results of the study. The results are presented as the average of the correlations over the replications between the known generating thetas and the estimated thetas from the various conditions of the study. Table 2 displays the average BIAS in recovering the known generating thetas for each condition, summarized across the 100 iterations. Table 3 presents the corresponding RMSD values for each of the 96 conditions. The BIAS and RMSD results were used to confirm and highlight the results observed in the correlational analyses. As such, results of all three sets of analyses will be discussed in parallel.

Insert Tables 1-3 Here

General Observations

One of the key findings of the previous study was that when combined theta estimation was used, the resulting theta estimate was more strongly related to the known multiple choice theta than it was to either the true essay or true average theta, particularly in cases using the 3PL and when high multidimensionality was present. This pattern was observed in the current study, but appeared to be mitigated by the essay scoring method used. This pattern was much less pronounced when analytic scoring was used.

Multiple choice theta estimates

Several conclusions can be drawn with regard to the ability to recover the correlation between multiple choice theta estimates and their known values. In general, multiple choice thetas were recovered fairly well, but not quite as well as in the Meyers et al. (2009) study. On average, the correlation between generating and estimated multiple choice thetas was 0.909 and ranged from 0.686 to 0.974. In the previous study, these values ranged from 0.905 to 0.930.

These correlations were impacted by essay scoring approach. Conditions in which an analytic scoring approach was used had a lower mean correlation (0.900) and were more variable (0.686 - 0.974) than conditions using holistic scoring (mean of 0.917 and range of 0.873 to 0.942). In addition, these correlations improved slightly as the number of multiple choice items increased.

Several additional study factors impacted the ability to recover multiple choice thetas. In general, correlations were higher with generating values and GOR indices were smaller when the 3PL was used instead of the 1PL. Correlations also slightly increased (and GOR indices decreased) as multidimensionality decreased. However, the impact of multidimensionality was found to depend on several other factors. First, the degree of multidimensionality interacted with the theta estimation procedure used (see Figure 2) such that separate theta estimation procedures were more accurate than combined theta estimation procedures, except in the case of pure unidimensionality. Secondly, degree of multidimensionality interacted with the essay scoring approach (see Figure 3) such that holistic scoring approaches were more accurate in recovering the multiple choice thetas than analytic approaches, except in the case of pure unidimensionality. Finally, an interaction was noted between ability estimation and essay scoring method. When analytic scoring was used, separate estimation produced more accurate results for multiple choice thetas, but when holistic scoring was used there was a slight advantage in favor of combined estimation (see Figure 4). This pattern was more pronounced in conditions involving fewer multiple choice items.

Essay theta estimates

In general, essay theta estimates correlated more highly with their known values as multidimensionality decreased. Similarly, the magnitude of the GOR indices was inversely related to degree of multidimensionality. However, as was the case with multiple choice thetas,

this pattern depended on several other study factors. Similarly to the multiple choice thetas, the impact of multidimensionality depended on the theta estimation procedure (see Figure 5). In cases with extreme multidimensionality ($r=0.30$), separate estimation of multiple choice and essay thetas yielded more accurate results, but in cases of low multidimensionality, combined theta estimation was more accurate. The impact of multidimensionality also depended on the essay scoring approach used (see Figure 6). Analytic scoring was more accurate than holistic scoring, but this advantage decreased as unidimensionality was approached. Finally, the impact of multidimensionality was influenced by the IRT model used (see Figure 7). The 1PL performed better than the 3PL in cases of extreme multidimensionality ($r=-.30$), but the pattern reversed as multidimensionality decreased.

The impact of the IRT model used was also related to two other factors: the theta estimation procedure (see Figure 8) and the essay scoring approach (see Figure 9). The 3PL model performed better than the 1PL model when paired with separate theta estimation, but worse when paired with combined theta estimation. In addition, the 3PL performed better than the 1PL with the analytic scoring approach, but worse than the 1PL with the holistic scoring approach. Also, the impact of theta estimation procedure was found to vary based on the essay scoring approach used (see Figure 10). Separate theta estimation was more accurate in recovering essay thetas than combined theta estimation in analytic scoring conditions but less accurate in holistic scoring conditions when using RMSD as the unit of analysis. It should be noted that when analyzing the patterns based on the recovery of the correlations between true and estimated essay thetas, that separate theta estimation always performed better than combined theta estimation, but the difference was significantly less in holistic scoring conditions ($r=0.655$ vs. 0.634) than in analytic scoring conditions ($r= 0.916$ vs. 0.793). Lastly, the number

of multiple choice items on the assessment had no tangible impact on the recovery of the essay theta estimates.

Average theta estimates

As was the case in the Meyers et al. (2009) study, results evaluating the ability to recover average theta across item types were encouraging. When combined theta estimation was used, the average correlations between the theta estimate and the generating theta average ranged from 0.799 to 0.968. When separate theta estimation was used, the average correlations between the average generating thetas and the average estimated thetas ranged from 0.743 to 0.969. By comparison, in the previous study, these values ranged from 0.757 to 0.934 and 0.788 to 0.894, respectively. Average correlations between theta average estimates and their generating values increased as multidimensionality decreased, but this varied based upon the IRT model fitted, the theta estimation procedure used, and the essay scoring approach implemented.

In cases of high multidimensionality, the 1PL model produced more accurate results, but the 3PL model performed better as multidimensionality decreased (see Figure 11). In addition, as observed in Meyers et. al (2009), when levels of multidimensionality were high, separate theta estimation was more accurate than combined theta estimation (see Figure 12), but this pattern reversed when unidimensionality was approached. This pattern held regardless of the number of multiple choice items on the assessment or the scoring approach used. The analytic essay scoring approach was better able to recover average theta values than holistic scoring approaches (see Figure 13) , and the difference was much greater when the simulated multidimensionality was high.

Interactions were also found between IRT model and theta estimation procedure (Figure 14), and IRT Model and essay scoring approach (Figure 15). The 3PL slightly outperformed the

1PL when separate estimation procedures were used versus combined theta estimation procedures and when analytic scoring procedures were used as opposed to holistic scoring approaches.

Item Parameter Estimation

Multiple Choice Items

Tables 4 and 5 compare the different conditions in terms of the goodness-of-recovery of the multiple choice item parameters. Table 4 presents BIAS statistics and Table 5 presents the corresponding RMSD values. It should be noted that theta estimation procedure was not included in these analyses, because theta estimation procedure (separate vs. combined) when viewed as an independent factor does not directly impact estimation of the item parameters. In other words, two study conditions that only differed in the theta estimation procedure utilized produced identical results for item parameter recovery.

As was the case in the Meyers et al. (2009) study, these tables indicate that multidimensionality has little observed impact on recovery of the multiple choice item parameters; there is a small improvement as multidimensionality decreases, but the impact is negligible. In addition, the number of multiple choice items in the dataset had little appreciable impact on recovery of the multiple choice item parameters. The most revealing observation from these tables is an interaction between IRT model and essay scoring method (see Figure 16) Analytic scoring methods resulted in less accurate multiple choice item parameters, but particularly when paired with the 3PL model. The combination of analytic scoring and 3PL produced the worst recovery of the multiple choice item parameters. On the other hand, the best results were found when holistic scoring and the 1PL were used.

Essay Item

Tables 6 and 7 present the BIAS and RMSD values for the recovery of the essay item parameters in the holistic scoring conditions, and Tables 8-13 present the corresponding values for the analytic scoring conditions. Note that in the analytic scoring conditions, a single essay was scored on three different dimensions and was thus treated psychometrically as three separate essays. As found in Meyers et al. (2009), the essay item parameters were much more heavily affected by the conditions of the study than the multiple choice item parameters. It is evident from these tables that both the average step values and the individual step difficulties are strongly affected by the degree of multidimensionality present in the data. In cases with high levels of multidimensionality, these estimates were severely discrepant. Furthermore, in the 3PL conditions, the a parameter is severely underestimated in the 0.30 and 0.50 correlation conditions. However, as expected, the estimated values (both a 's and b 's) approached the generating values as multidimensionality decreased. As shown previously (Meyers et al. 2009), the conditions using the 3PL appeared to be more susceptible to multidimensionality in the data. In general, the 1PL was more accurate in recovery of both average b 's and step values, but the advantage decreased steadily as multidimensionality decreased (see Figure 17). Overall, analytic scoring procedures were more accurate than holistic scoring procedures in recovery of both average b values and step difficulties, but there was an interaction with measurement model. The most accurate essay parameter recovery occurred with the pairing of analytic scoring and the 1PL and the least accurate occurred with the pairing of holistic scoring and the 3PL (Figure 18).

Discussion and Conclusion

This simulation study was conducted as a direct extension to the Meyers et al. (2009) work and evaluated the impact on recovery of person and item parameters of several additional

writing assessment characteristics. In the Meyers et al. (2009) study multiple choice items dominated the construct being measured, accounting for nearly 98 percent of the items. To evaluate whether assessments containing different mixes of multiple choice and constructed response items responded differently to the conditions of the study, the proportion of multiple choice items to essay items was varied in the current study. In addition, in the writing assessment modeled in the 2009 study represented a somewhat atypical case. In that assessment program, two raters holistically score the essay on a six point scale and the ratings are summed, resulting in a scale ranging from 2-12. Typically, essay items contain fewer than 11 score categories, which presents estimation problems in general. Furthermore, many writing assessments contain essays that are scored analytically rather than holistically. It was unclear whether the results from the 2009 study would apply to different models of writing assessment. The current study evaluated how different ability estimation techniques, measurement models, essay scoring approaches, and distributions of item types impact theta and item parameter recovery under different conditions of dimensionality. This study was used as a way of assessing the generalizability of the results from the previous study.

As was the case with the 2009 study, the results from this study should be encouraging to testing and measurement professionals working with writing assessments. The results of this study suggest that even when writing assessments contain items measuring two separate constructs, they can be adequately handled through the use of unidimensional IRT models. Specifically, if as is done with most writing assessments, the goal is to estimate and report one, overall writing assessment score, this study suggests that either of two methods can be successfully used. By estimating either one ability based on performance on the combination of multiple choice and essay items, or by estimating separate abilities and then averaging them,

examinees' underlying "writing" ability can be recovered fairly accurately without the need for multidimensional IRT models. With that said, this study provides some clear guidance for practitioners deciding how to handle writing assessment psychometrically.

One of the main findings from the previous study was that estimation procedure interacted with multidimensionality such that separate theta estimation was more accurate when multidimensionality was high, but combined theta estimation was more accurate when the data approached unidimensionality. That result generalized to this study and that relationship did not vary based on the scoring approach implemented or the number of multiple choice items.

The previous study also concluded that thetas based on multiple choice responses were recovered much better than either essay thetas or thetas based on averaging the multiple choice and essay thetas and that when using combined theta estimation, the theta estimate was more strongly related to the multiple choice generating value than the essay generating value. While this pattern was observed in the current study, using analytic scoring seems to mitigate these relationships.

In general, varying the number of multiple choice items had little impact on the results in this study. However, adding conditions that involved analytic scoring of essay items impacted the results substantially.

Based on the results of this study, it appears that when high levels of multidimensionality are observed or expected, that following general recommendations are in order:

- Separate estimation procedures are recommended over combined theta estimation techniques
- Analytic essay scoring tends to better recover the writing component driving essay performance, and

- 1PL and PCM models are preferable to the 3PL family of models if the skills being measured by the multiple choice and essay component are not thought to be highly correlated.

It is hoped that the results of this study will add to the growing body of existing research on and provide state departments and measurement professionals some guidance in choosing the most appropriate psychometric methods for their mixed-format writing assessments.

References

- Aschbacher, P. R. (1991). Performance assessment: state activity, interest, and concerns. *Applied Measurement in Education, 4*, 275–288.
- Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum.
- Bouwmeester, S., & Sitjtsma, K. (2004). Measuring the ability of transitive reasoning, using product and strategy information. *Psychometrika, 69*(1) 123-146.
- Dragow, F., & Parsons, C. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.
- Ercikan, K., Schwarz, R.D., Julian, M.W., Burket, G.R., Weber, M.M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*(2) 137-154.
- Hardy, R. (1996). Performance Assessment: Examining the costs. In M.B. Kane & R. Mitchell (Eds.) *Implementing Performance Assessment (pp. 107-117)*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Harrison, D.A. (1986). Robustness of IRT Parameter Estimation to Violations of the Unidimensionality Assumption. *Journal of Educational Statistics, 11*(2), 91-115
- Hoover, H.D. & Bray, G. (1996, April). *The research and development phase: Can a performance assessment be cost effective?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

- Khatti, N. & Sweet, D. (1996). Assessment reform: Promises and challenges. In M.B. Kane & R. Mitchell (Eds.) *Implementing Performance Assessment (pp. 1-21)*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison Wesley.
- Lord, F. (1980). *Application of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Maris, E. (1999). Estimating multiple classification latent class models, *Psychometrika*, 64 (2), 187-212.
- Meyers, J.L, Turhan, A., & Fitzpatrick, S.F. (2009). *Interaction of Calibration Procedure and Ability Estimation Method for Writing Assessments under Conditions of Multidimensionality*. Paper presented at the annual conference of the National Council on Measurement in Education, San Diego.
- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16 (2), 159-176.
- Nichols, P.(2005). Evidence for the interpretation and use of scores from an automated essay scorer. *Pearson White Paper*.
- Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. *Educational Policy Analysis Archives*, 8(16), 1-15.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

- Sykes, R., Hou, L., Hanson, B., & Zhen, W. (2002). *Multidimensionality and the Equating of a Mixed-Format Math Examination*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.
- Thissen, D. M. (2000). *MULTILOG: Item analysis and scoring with multiple category response models (version 7.0.1)*. Mooresville, IN: Scientific Software, Inc.
- Traub, R.E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R.E. Bennett and W.C. Ward (Eds.) *Construction versus choice in cognitive measurement*. (pp. 29-44). Hillsdale, NJ: Erlbaum.
- Uebersax JS. MVN program for random multivariate normal numbers. 2006. Available at the *Statistical Methods for Rater Agreement* website: <http://ourworld.compuserve.com/homepages/jsuebersax> . Accessed: 11/05/2008.
- Um, K., Kim, D., & Turhan, A. (2007). *An evaluation of calibration/linking approaches to mixed format writing assessments*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Whittaker, T.A , Fitzpatrick, S.J., Williams, N.J., & Dodd, B.G. (2003) IRTGEN: A SAS Macro Program to Generate Known Trait Scores and Item Responses for Commonly Used Item Response Theory Models. *Applied Psychological Measurement* 27(4), 299-300.

		Holistic			
		1 PL		3 PL	
# of MC items		Combined	Separate	Combined	Separate
20		0.30	0.30	0.30	0.30
		0.50	0.50	0.50	0.50
		0.80	0.80	0.80	0.80
		1.00	1.00	1.00	1.00
28		0.30	0.30	0.30	0.30
		0.50	0.50	0.50	0.50
		0.80	0.80	0.80	0.80
		1.00	1.00	1.00	1.00
36		0.30	0.30	0.30	0.30
		0.50	0.50	0.50	0.50
		0.80	0.80	0.80	0.80
		1.00	1.00	1.00	1.00

		Analytic			
		1 PL		3 PL	
# of MC items		Combined	Separate	Combined	Separate
20		0.30	0.30	0.30	0.30
		0.50	0.50	0.50	0.50
		0.80	0.80	0.80	0.80
		1.00	1.00	1.00	1.00
28		0.30	0.30	0.30	0.30
		0.50	0.50	0.50	0.50
		0.80	0.80	0.80	0.80
		1.00	1.00	1.00	1.00
36		0.30	0.30	0.30	0.30
		0.50	0.50	0.50	0.50
		0.80	0.80	0.80	0.80
		1.00	1.00	1.00	1.00

Figure 1. Simulation study design

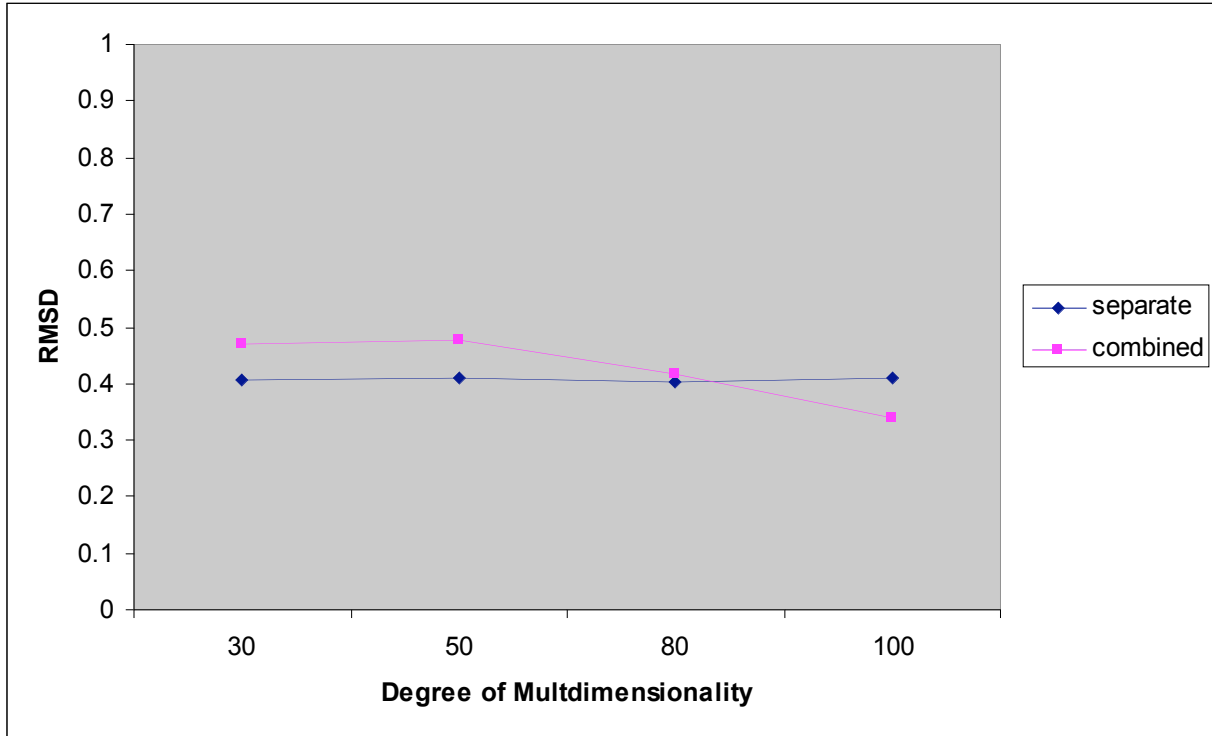


Figure 2. Relationship between theta estimation method and degree of multidimensionality in estimation of multiple choice thetas

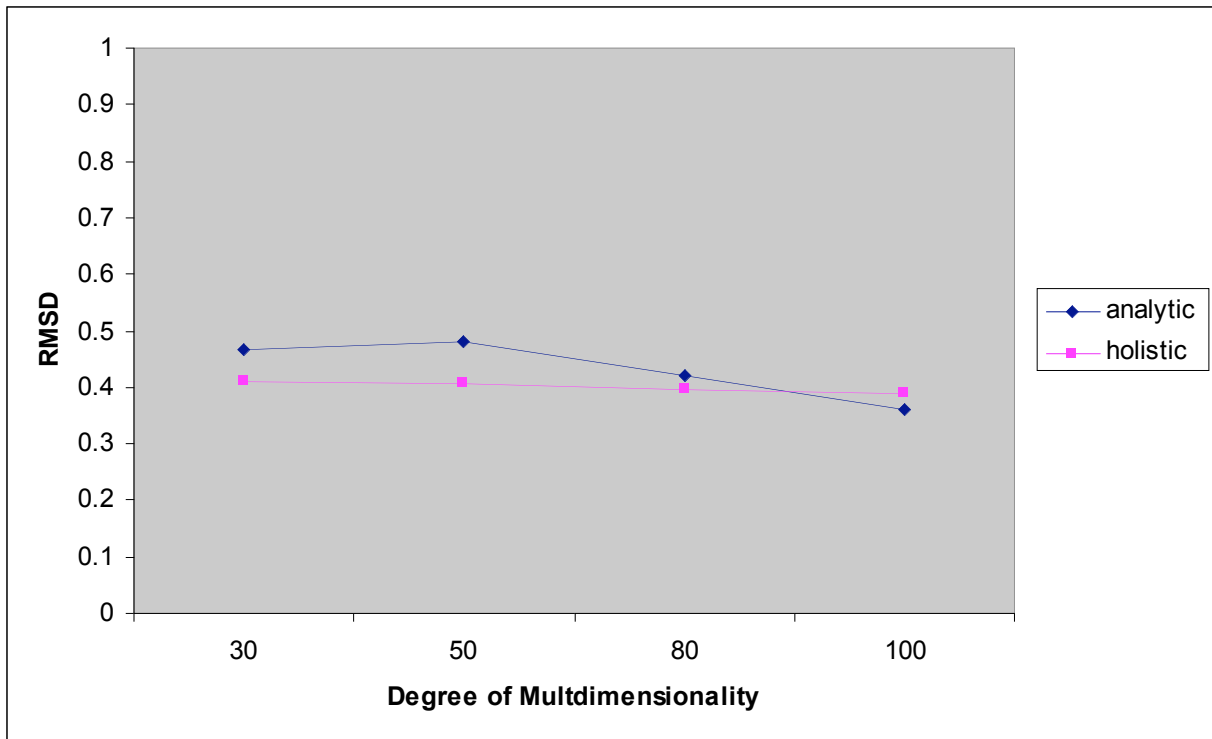


Figure 3. Relationship between essay scoring approach and degree of multidimensionality in estimation of multiple choice thetas

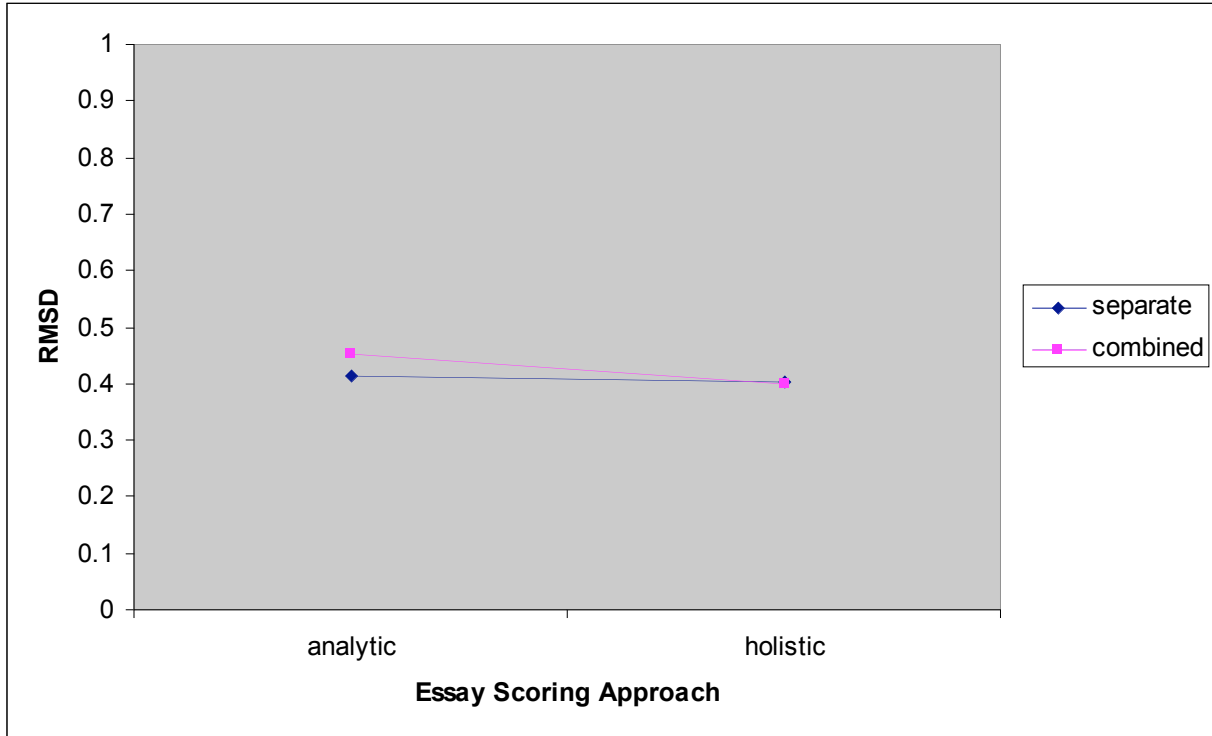


Figure 4. Relationship between essay scoring approach and theta estimation method in estimation of multiple choice thetas

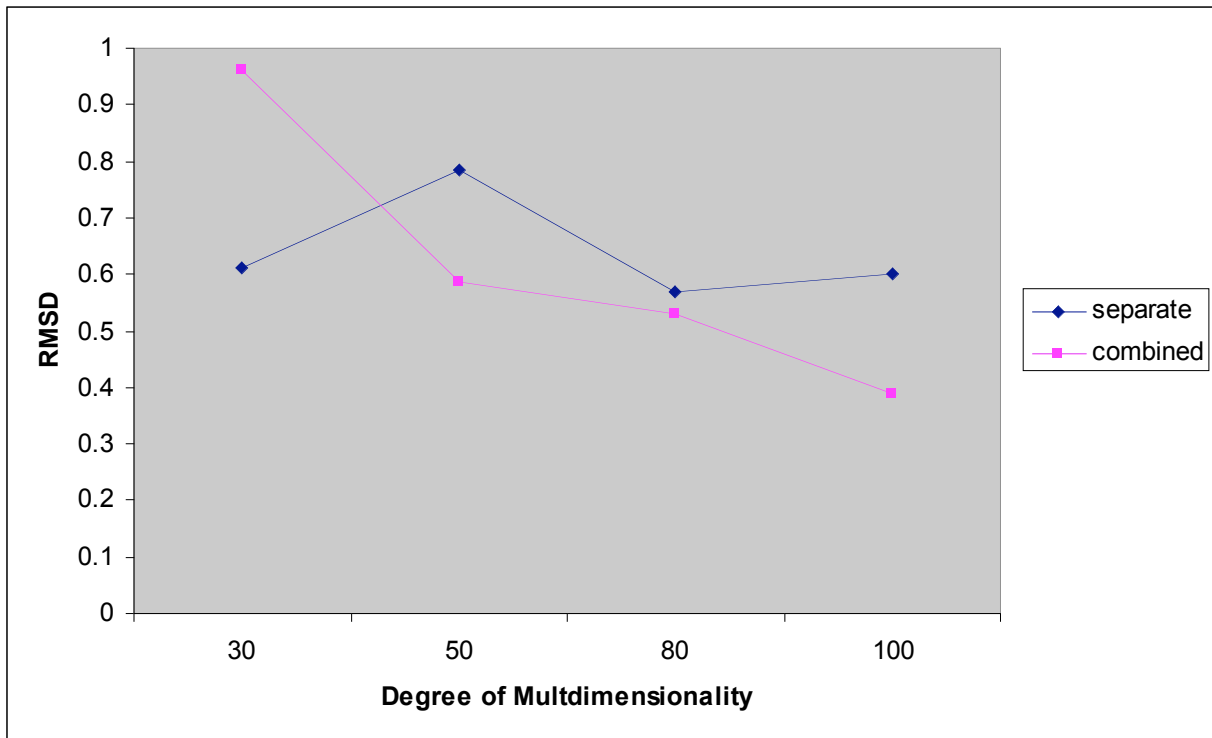


Figure 5. Relationship between theta estimation method and degree of multidimensionality in estimation of essay thetas

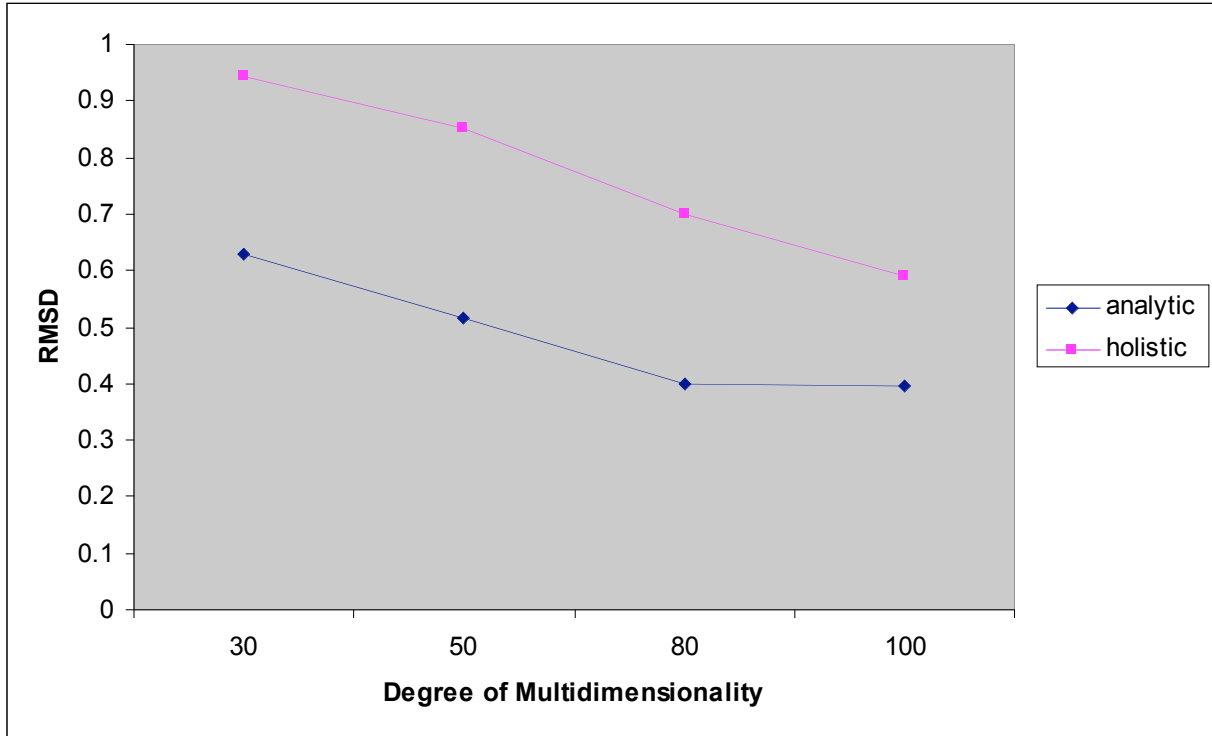


Figure 6. Relationship between essay scoring approach and degree of multidimensionality in estimation of essay thetas

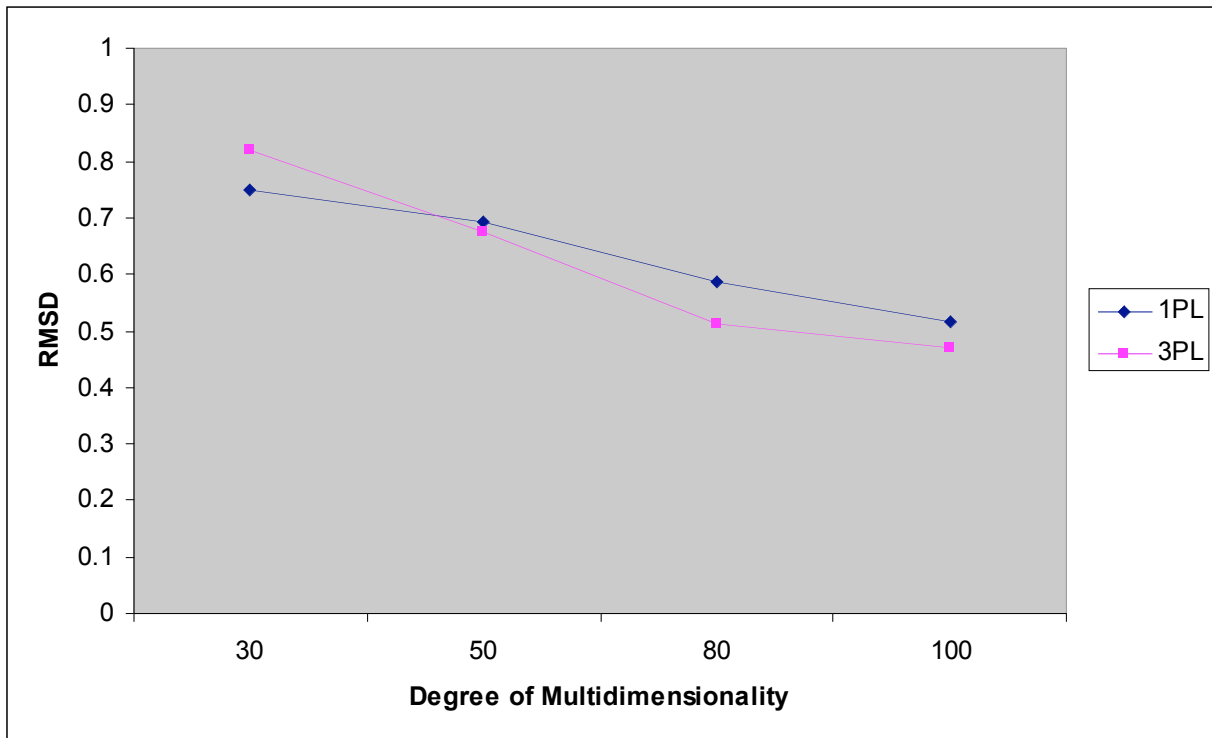


Figure 7. Relationship between IRT model and degree of multidimensionality in estimation of essay thetas

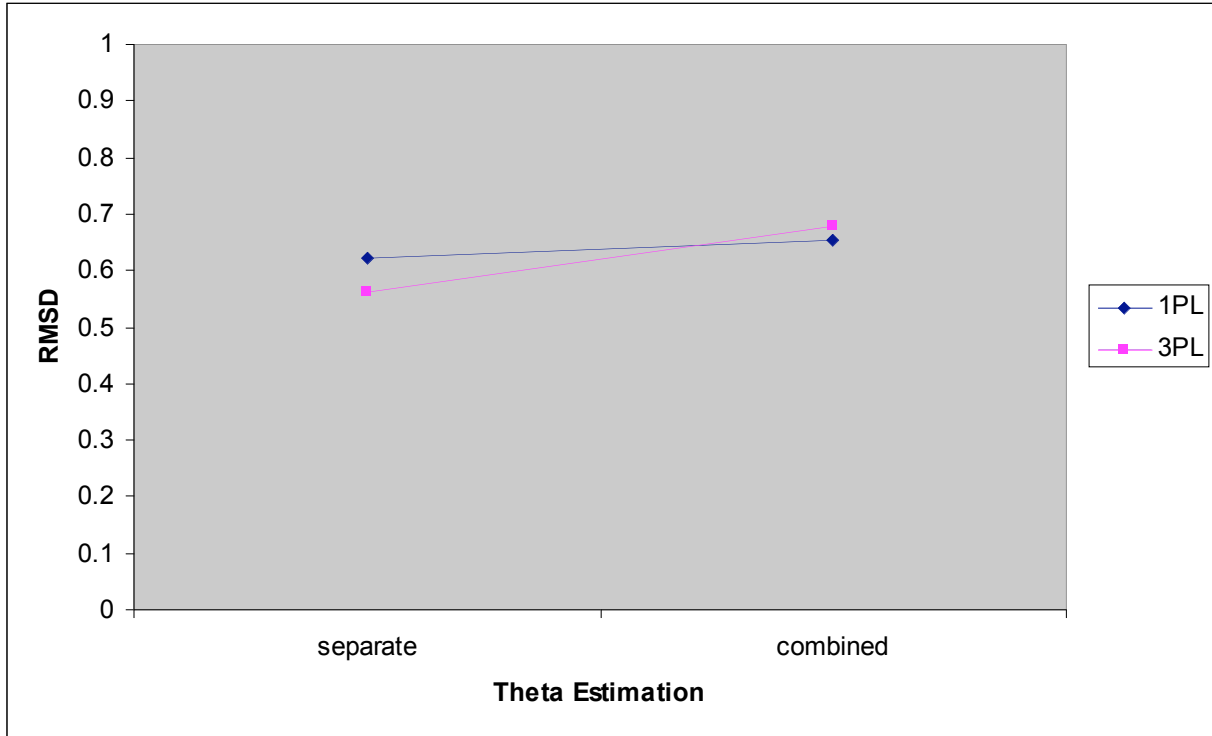


Figure 8. Relationship between IRT model and theta estimation method in estimation of essay thetas

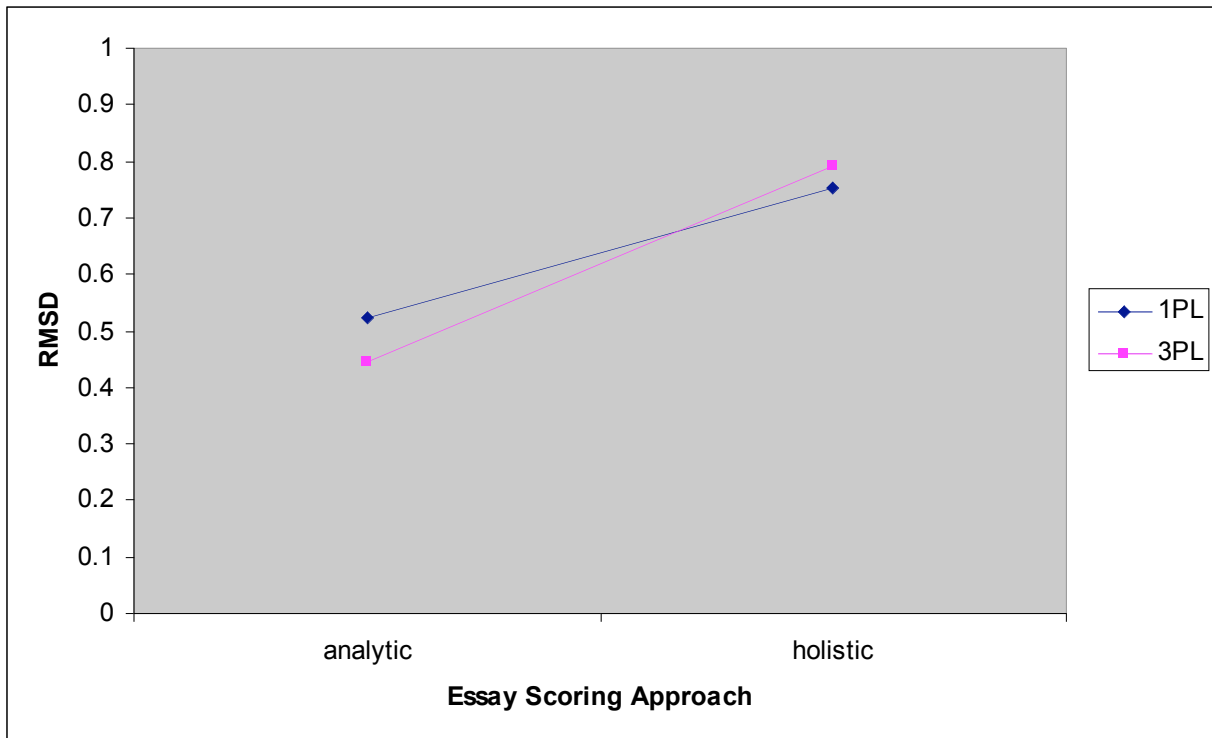


Figure 9. Relationship between IRT model and essay scoring approach in estimation of essay thetas

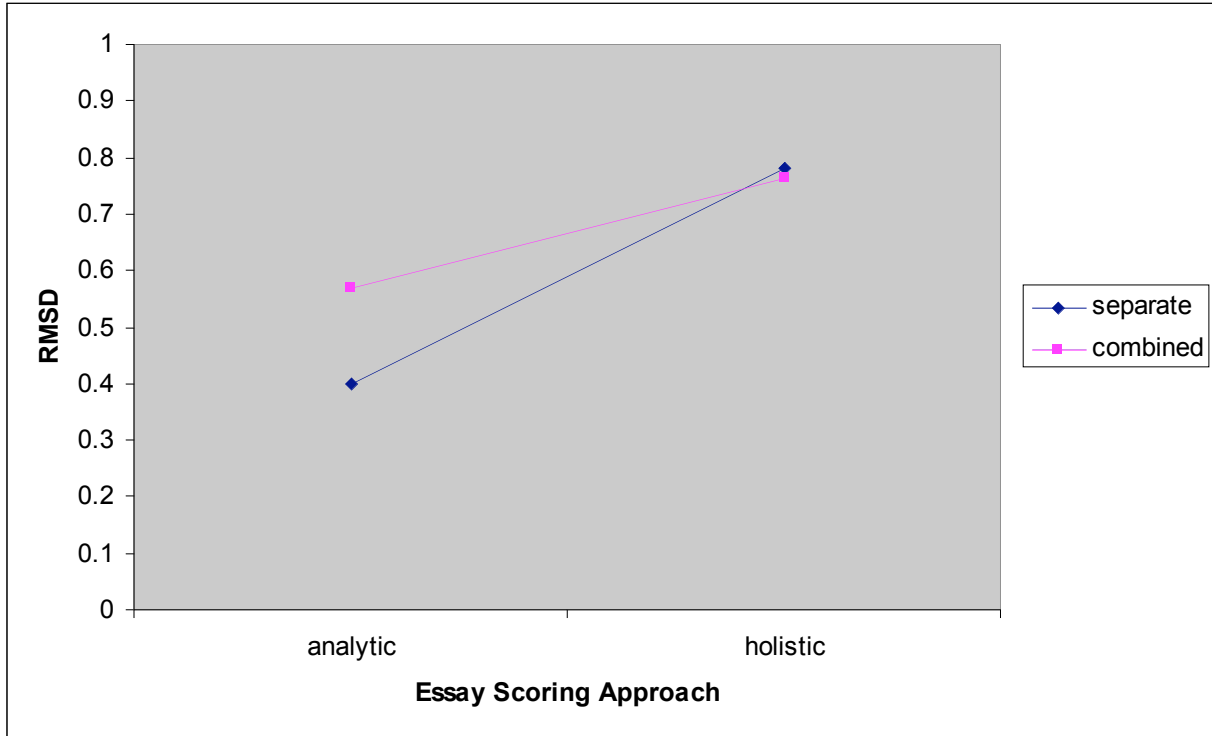


Figure 10. Relationship between essay scoring approach and theta estimation method in estimation of essay thetas

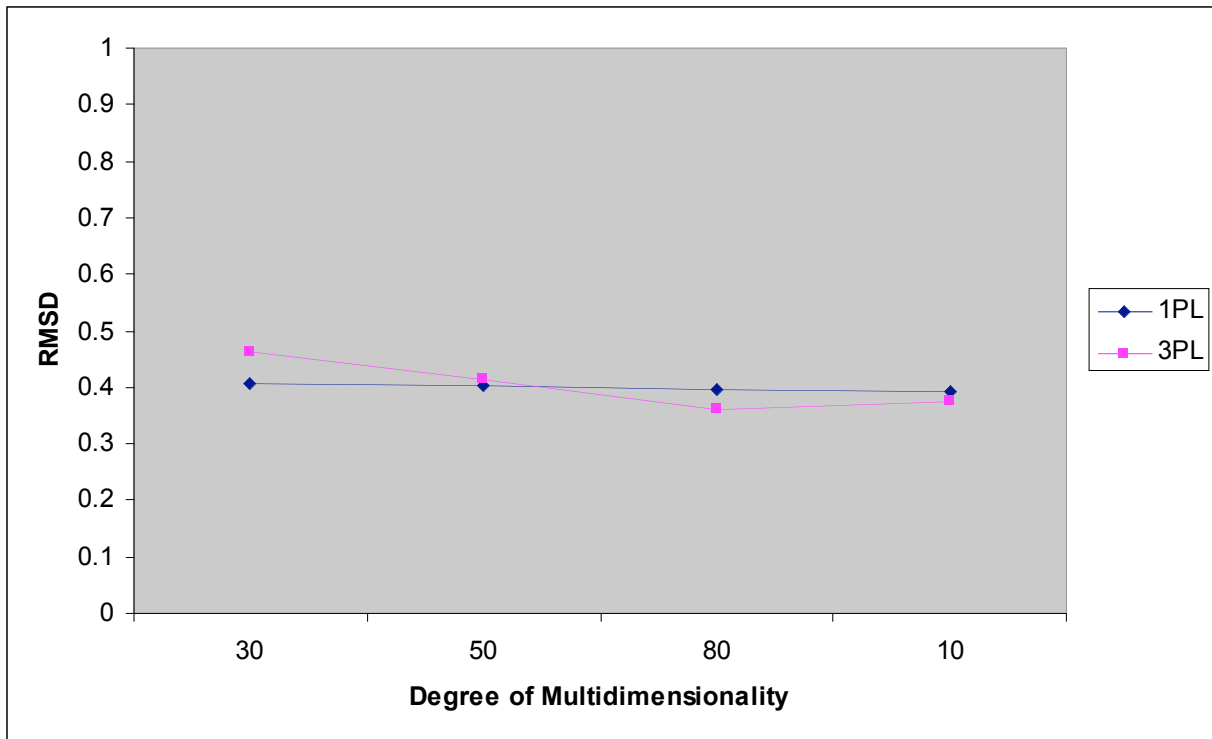


Figure 11. Relationship between IRT model and degree of multidimensionality in estimation of average thetas

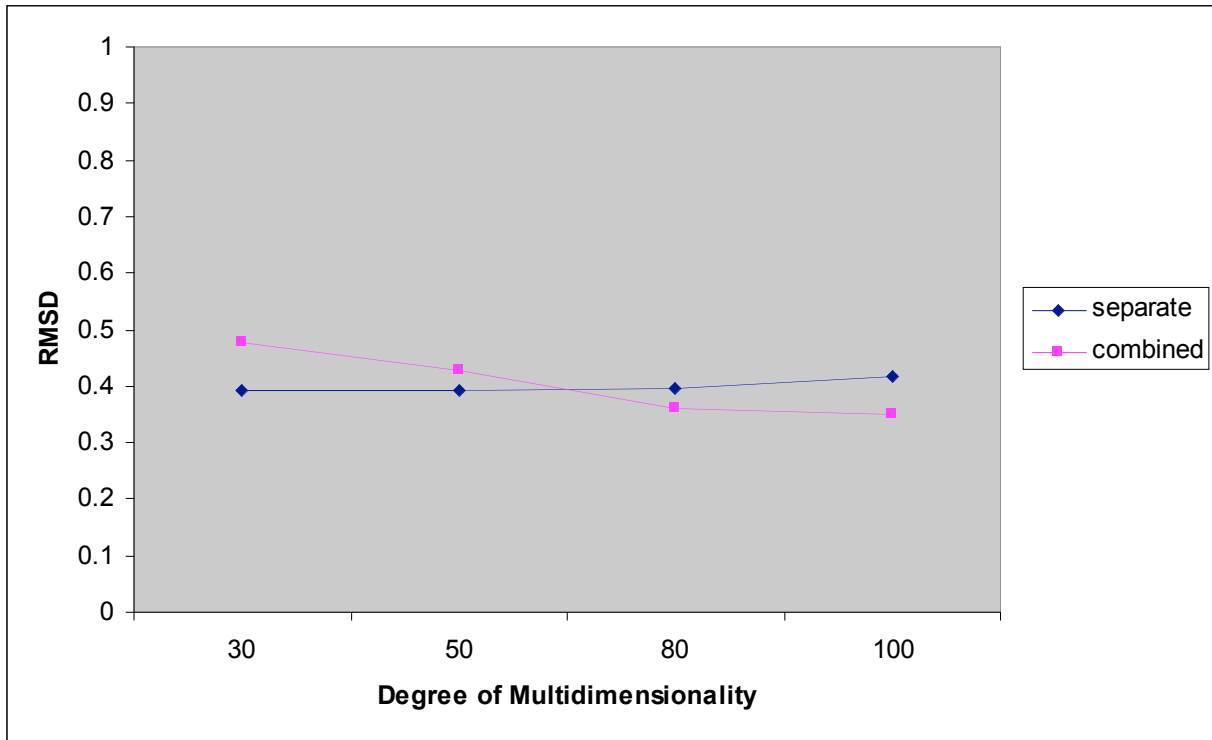


Figure 12. Relationship between theta estimation method and degree of multidimensionality in estimation of average thetas

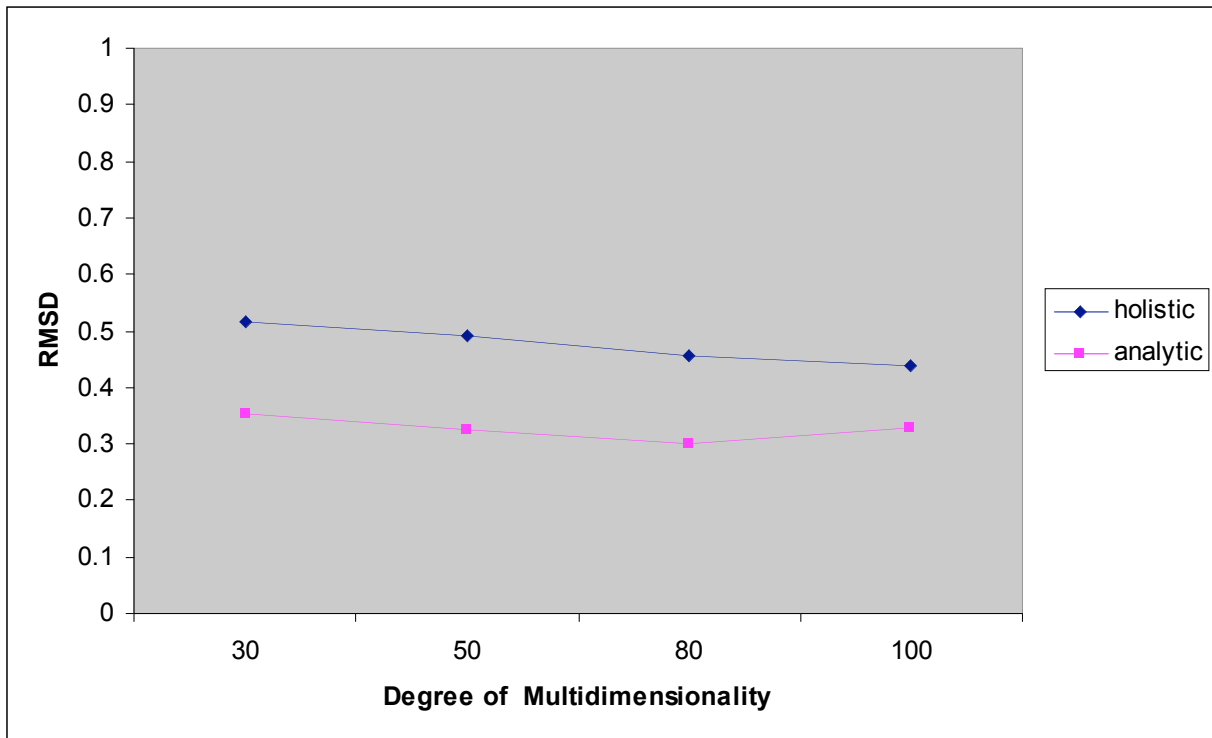


Figure 13. Relationship between essay scoring approach and degree of multidimensionality in estimation of average thetas

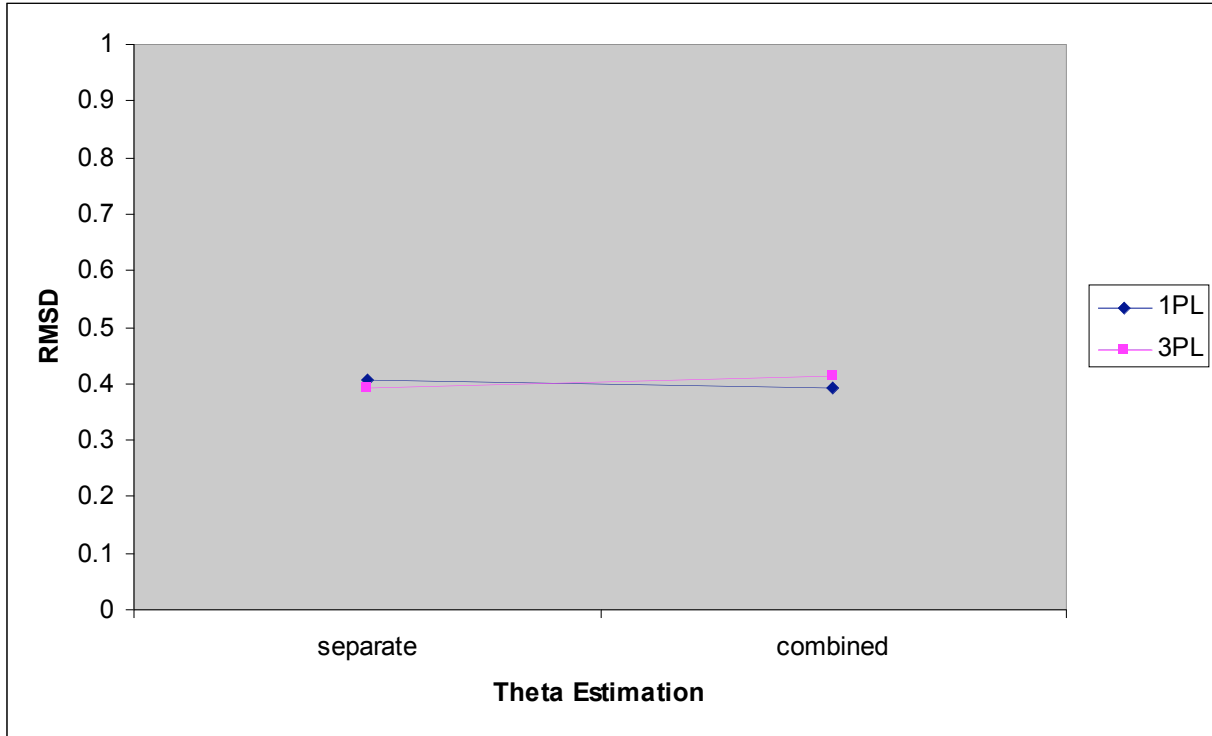


Figure 14. Relationship between theta estimation method and IRT model in estimation of average thetas

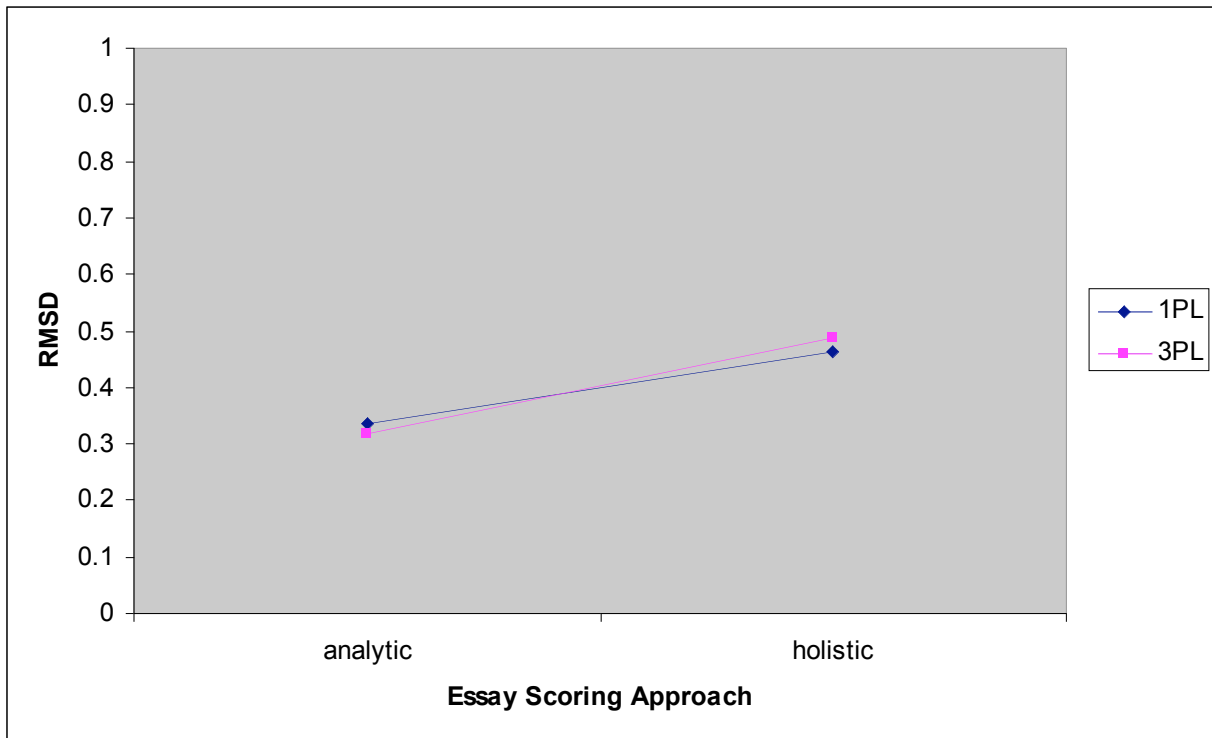


Figure 15. Relationship between essay scoring approach and IRT model in estimation of average thetas

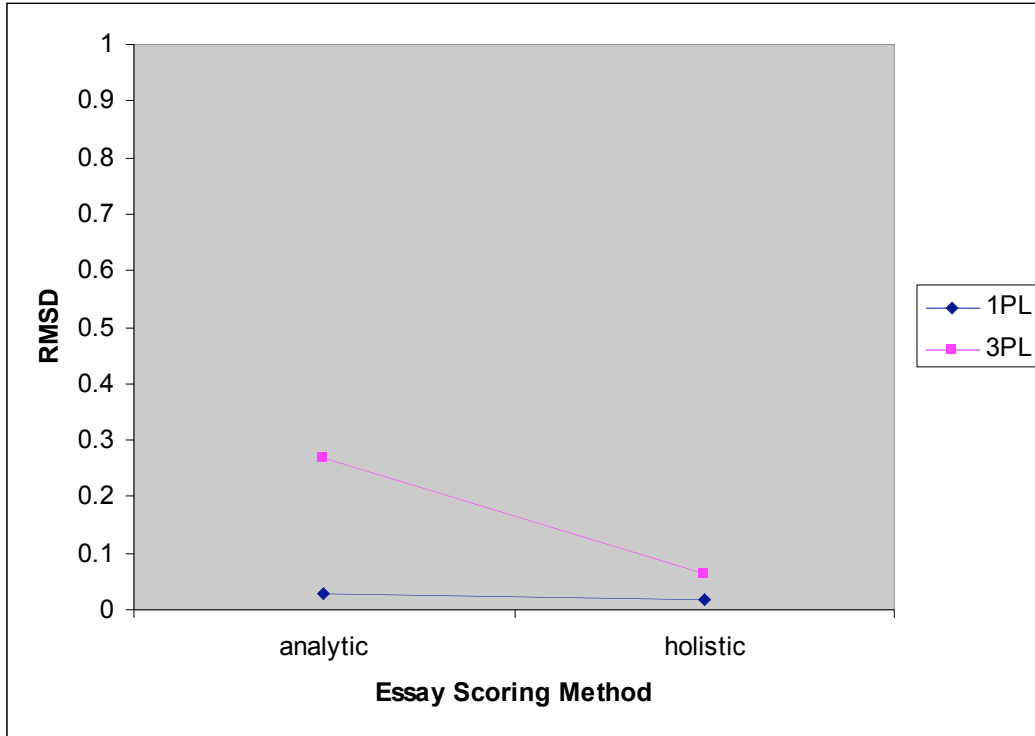


Figure 16. Relationship between essay scoring approach and IRT model when recovering multiple choice b-values

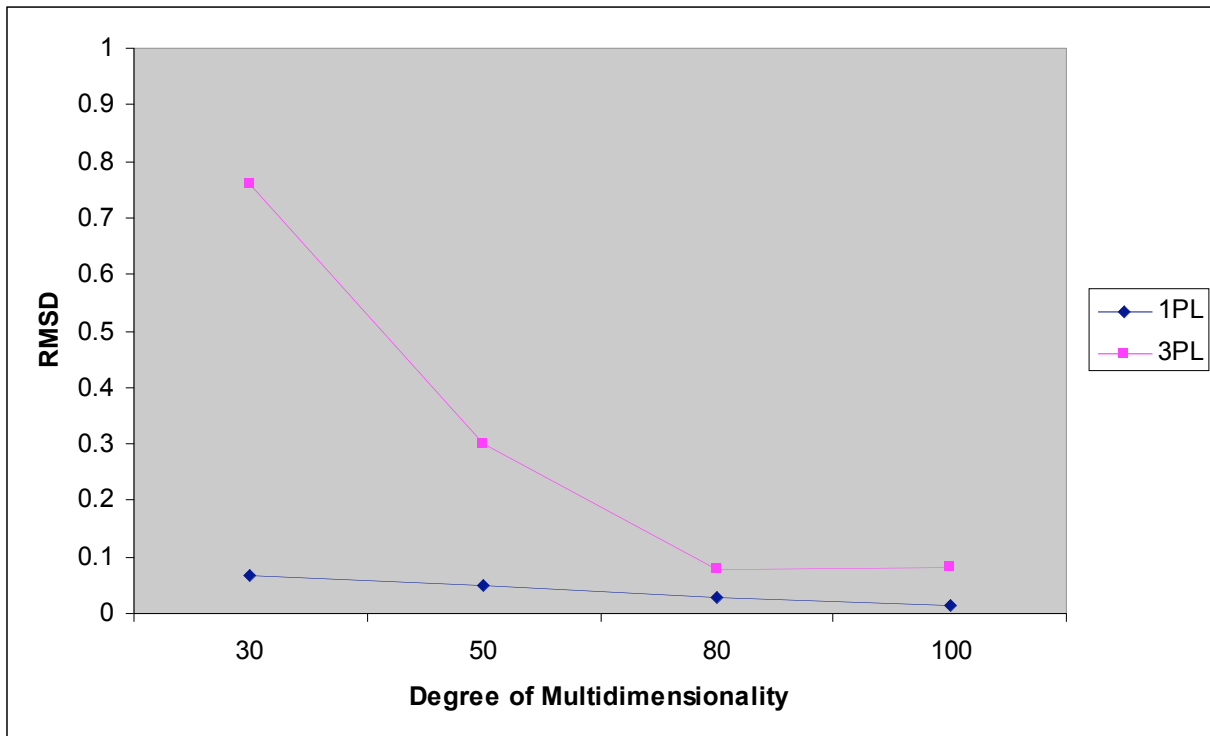


Figure 17. Relationship between IRT model and degree of multidimensionality when recovering average essay step values

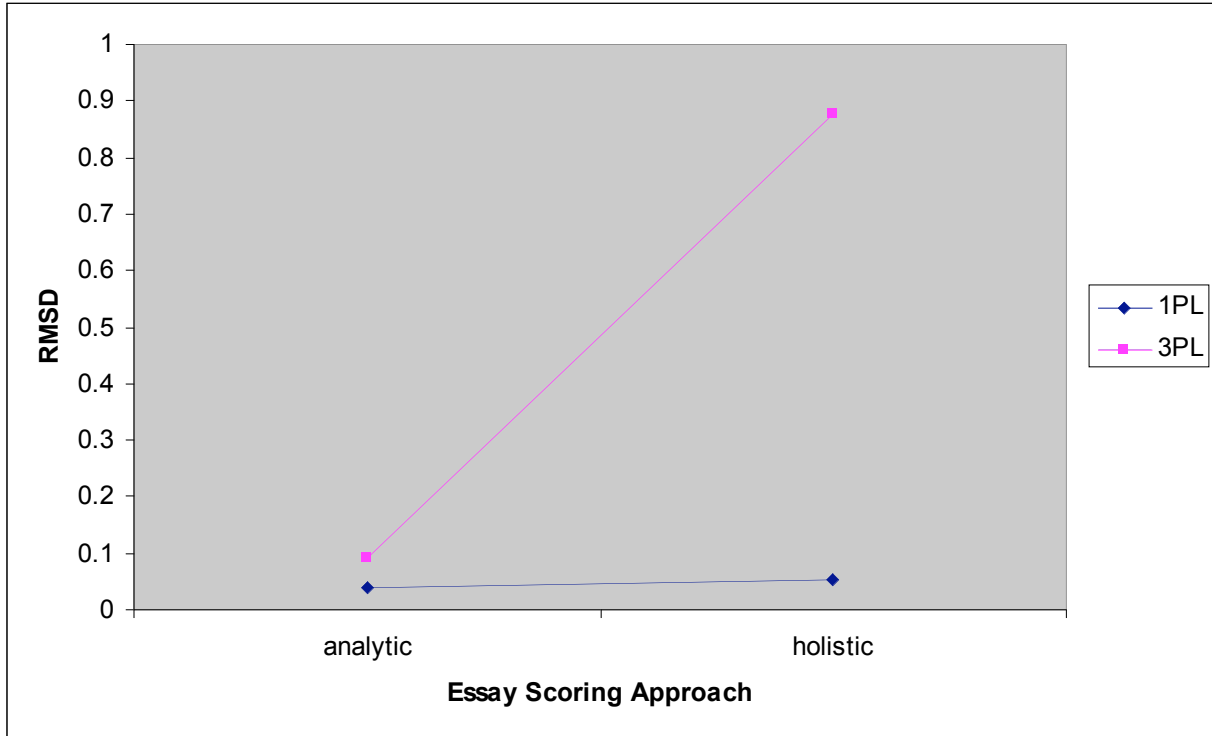


Figure 18. Relationship between IRT model and essay scoring approach when recovering average essay step values

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.086	0.289	0.189	0.110	0.040	-0.024	-0.089
		0.50	xxx	0.060	0.193	0.136	0.080	0.027	-0.015	-0.064
		0.80	xxx	0.026	0.073	0.059	0.034	0.015	-0.003	-0.025
		1.00	xxx	0.004	0.015	0.003	-0.002	0.006	-0.001	0.006
	3PL	0.30	-0.865	-2.490	-3.079	-10.653	-1.677	-4.603	3.344	1.726
		0.50	-0.718	-1.023	-1.024	-4.594	-0.605	-2.040	1.488	0.635
		0.80	-0.394	-0.250	-0.173	-1.205	-0.107	-0.537	0.399	0.121
		1.00	0.002	-0.006	-0.043	0.002	0.003	0.005	0.000	-0.003
28	1PL	0.30	xxx	0.079	0.272	0.163	0.099	0.036	-0.015	-0.077
		0.50	xxx	0.057	0.186	0.125	0.074	0.026	-0.012	-0.055
		0.80	xxx	0.023	0.068	0.055	0.028	0.014	-0.001	-0.024
		1.00	xxx	0.004	-0.007	0.005	0.005	0.007	0.007	0.009
	3PL	0.30	-0.864	-2.475	-3.131	-10.571	-1.646	-4.569	3.340	1.727
		0.50	-0.718	-1.036	-1.086	-4.609	-0.604	-2.039	1.490	0.635
		0.80	-0.395	-0.254	-0.189	-1.210	-0.112	-0.543	0.406	0.123
		1.00	0.001	0.001	-0.003	0.003	-0.001	0.005	0.003	-0.001
36	1PL	0.30	xxx	0.071	0.229	0.153	0.092	0.034	-0.013	-0.066
		0.50	xxx	0.052	0.155	0.124	0.068	0.026	-0.010	-0.049
		0.80	xxx	0.029	0.093	0.053	0.027	0.015	0.000	-0.015
		1.00	xxx	0.005	0.004	0.002	0.004	0.008	0.005	0.006
	3PL	0.30	-0.865	-2.457	-3.018	-10.531	-1.667	-4.593	3.354	1.712
		0.50	-0.719	-1.045	-1.140	-4.614	-0.613	-2.052	1.506	0.644
		0.80	-0.394	-0.251	-0.178	-1.211	-0.105	-0.537	0.400	0.123
		1.00	0.002	0.000	-0.012	0.000	0.001	0.005	0.004	0.002

Table 6. MEAN BIAS for Holistic essay item parameters by condition

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.088	0.317	0.198	0.112	0.045	0.029	0.091
		0.50	xxx	0.063	0.225	0.153	0.084	0.033	0.022	0.069
		0.80	xxx	0.033	0.151	0.094	0.042	0.023	0.017	0.034
		1.00	xxx	0.020	0.131	0.074	0.026	0.019	0.017	0.025
	3PL	0.30	0.865	2.495	3.182	10.666	1.683	4.607	3.347	1.731
		0.50	0.718	1.025	1.106	4.602	0.609	2.042	1.489	0.640
		0.80	0.394	0.252	0.280	1.213	0.113	0.539	0.401	0.126
		1.00	0.013	0.024	0.146	0.070	0.026	0.024	0.019	0.023
28	1PL	0.30	xxx	0.082	0.297	0.178	0.102	0.042	0.022	0.082
		0.50	xxx	0.060	0.218	0.143	0.078	0.032	0.021	0.061
		0.80	xxx	0.032	0.148	0.090	0.038	0.022	0.017	0.034
		1.00	xxx	0.024	0.164	0.078	0.028	0.019	0.017	0.024
	3PL	0.30	0.864	2.479	3.202	10.585	1.653	4.573	3.344	1.732
		0.50	0.718	1.038	1.152	4.616	0.609	2.041	1.491	0.639
		0.80	0.395	0.256	0.276	1.216	0.118	0.544	0.407	0.128
		1.00	0.013	0.022	0.118	0.076	0.028	0.022	0.019	0.022
36	1PL	0.30	xxx	0.075	0.267	0.167	0.096	0.039	0.022	0.070
		0.50	xxx	0.056	0.195	0.141	0.073	0.033	0.019	0.053
		0.80	xxx	0.036	0.165	0.083	0.037	0.024	0.017	0.028
		1.00	xxx	0.019	0.115	0.066	0.029	0.020	0.018	0.023
	3PL	0.30	0.865	2.463	3.130	10.545	1.673	4.598	3.358	1.719
		0.50	0.719	1.047	1.220	4.621	0.618	2.053	1.508	0.647
		0.80	0.394	0.253	0.260	1.217	0.111	0.538	0.402	0.128
		1.00	0.014	0.023	0.127	0.080	0.026	0.022	0.021	0.024

Table 7. MEAN RMSD for Holistic essay item parameters by condition

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.068	0.375	0.272	0.138	0.005	-0.126	-0.257
		0.50	xxx	0.052	0.288	0.194	0.101	0.007	-0.091	-0.186
		0.80	xxx	0.024	0.116	0.085	0.046	0.007	-0.036	-0.073
		1.00	xxx	0.007	0.008	0.006	0.007	0.007	0.009	0.003
	3PL	0.30	-2.618	-0.086	3.104	-4.276	0.836	-2.075	2.516	-0.622
		0.50	-1.255	-0.034	0.181	-0.484	0.138	-0.128	0.261	-0.171
		0.80	-0.324	0.032	0.154	-0.018	0.098	0.025	0.070	-0.134
		1.00	0.139	0.138	0.329	0.263	0.190	0.133	0.025	-0.115
28	1PL	0.30	xxx	0.068	0.376	0.265	0.135	0.007	-0.127	-0.247
		0.50	xxx	0.052	0.278	0.195	0.102	0.007	-0.092	-0.177
		0.80	xxx	0.027	0.121	0.087	0.050	0.006	-0.029	-0.072
		1.00	xxx	0.006	0.006	0.003	0.007	0.006	0.008	0.009
	3PL	0.30	-2.846	-0.130	4.610	-6.436	1.171	-3.155	3.794	-0.761
		0.50	-2.396	0.041	1.515	-2.078	0.495	-0.978	1.204	-0.406
		0.80	-0.845	-0.001	0.153	-0.225	0.092	-0.079	0.143	-0.088
		1.00	0.027	0.134	0.340	0.237	0.190	0.118	0.039	-0.119
36	1PL	0.30	xxx	0.069	0.375	0.251	0.131	0.009	-0.119	-0.232
		0.50	xxx	0.051	0.270	0.183	0.094	0.005	-0.079	-0.169
		0.80	xxx	0.027	0.116	0.083	0.045	0.008	-0.029	-0.061
		1.00	xxx	0.009	0.008	0.008	0.010	0.009	0.011	0.006
	3PL	0.30	-2.874	-0.151	5.098	-7.190	1.284	-3.529	4.225	-0.792
		0.50	-2.532	-0.043	2.025	-2.684	0.596	-1.306	1.576	-0.466
		0.80	-1.113	-0.007	0.217	-0.348	0.115	-0.134	0.198	-0.093
		1.00	-0.008	0.120	0.326	0.202	0.170	0.102	0.039	-0.123

Table 8. MEAN BIAS for Analytic essay item parameters by condition for Trait 1

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.045	0.346	0.230	0.112	-0.016	-0.140	-0.263
		0.50	xxx	0.036	0.258	0.168	0.085	-0.012	-0.099	-0.185
		0.80	xxx	0.019	0.114	0.069	0.042	-0.005	-0.037	-0.071
		1.00	xxx	0.006	0.006	0.004	0.006	0.007	0.010	0.005
	3PL	0.30	-3.162	0.055	3.705	-3.933	0.571	-2.097	3.004	-0.918
		0.50	-1.761	-0.009	0.299	-0.454	0.139	-0.140	0.318	-0.216
		0.80	-0.766	0.034	0.202	-0.067	0.093	-0.012	0.095	-0.107
		1.00	-0.093	0.123	0.331	0.188	0.182	0.107	0.040	-0.111
28	1PL	0.30	xxx	0.046	0.350	0.226	0.105	-0.016	-0.138	-0.251
		0.50	xxx	0.035	0.253	0.169	0.078	-0.010	-0.100	-0.181
		0.80	xxx	0.020	0.108	0.077	0.036	0.001	-0.036	-0.066
		1.00	xxx	0.008	0.008	0.006	0.005	0.010	0.008	0.008
	3PL	0.30	-3.388	0.083	5.442	-5.848	0.741	-3.141	4.473	-1.170
		0.50	-2.960	0.028	1.857	-1.942	0.377	-1.009	1.470	-0.582
		0.80	-1.297	0.013	0.232	-0.234	0.088	-0.095	0.183	-0.098
		1.00	-0.199	0.115	0.328	0.162	0.174	0.094	0.051	-0.118
36	1PL	0.30	xxx	0.045	0.338	0.214	0.104	-0.011	-0.129	-0.243
		0.50	xxx	0.035	0.244	0.160	0.074	-0.008	-0.091	-0.170
		0.80	xxx	0.019	0.101	0.078	0.033	0.005	-0.034	-0.066
		1.00	xxx	0.009	0.008	0.010	0.009	0.010	0.013	0.007
	3PL	0.30	-3.413	0.087	6.003	-6.521	0.798	-3.484	4.986	-1.262
		0.50	-3.088	0.042	2.402	-2.472	0.423	-1.319	1.892	-0.673
		0.80	-1.590	0.009	0.309	-0.339	0.107	-0.148	0.254	-0.128
		1.00	-0.213	0.104	0.320	0.136	0.160	0.084	0.048	-0.126

Table 9. MEAN BIAS for Analytic essay item parameters by condition for Trait 2

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.070	0.374	0.251	0.136	0.014	-0.111	-0.240
		0.50	xxx	0.053	0.277	0.185	0.097	0.009	-0.081	-0.170
		0.80	xxx	0.025	0.116	0.079	0.045	0.006	-0.030	-0.066
		1.00	xxx	0.007	0.010	0.007	0.004	0.007	0.004	0.010
	3PL	0.30	-2.519	-0.250	1.364	-2.978	0.851	-1.518	1.572	-0.792
		0.50	-1.308	-0.034	0.033	-0.368	0.146	-0.093	0.220	-0.144
		0.80	-0.465	0.033	0.106	-0.008	0.094	0.011	0.072	-0.074
		1.00	0.162	0.150	0.334	0.276	0.188	0.132	0.039	-0.072
28	1PL	0.30	xxx	0.071	0.372	0.251	0.132	0.007	-0.109	-0.228
		0.50	xxx	0.052	0.269	0.185	0.097	0.007	-0.079	-0.165
		0.80	xxx	0.027	0.120	0.082	0.041	0.010	-0.028	-0.062
		1.00	xxx	0.007	0.009	0.006	0.005	0.010	0.007	0.006
	3PL	0.30	-2.723	-0.388	1.950	-4.429	1.168	-2.288	2.333	-1.059
		0.50	-2.317	-0.112	0.687	-1.441	0.485	-0.724	0.787	-0.467
		0.80	-0.919	-0.003	0.072	-0.168	0.094	-0.066	0.116	-0.067
		1.00	0.085	0.147	0.340	0.259	0.184	0.123	0.048	-0.073
36	1PL	0.30	xxx	0.071	0.363	0.241	0.126	0.008	-0.097	-0.216
		0.50	xxx	0.053	0.267	0.176	0.091	0.010	-0.074	-0.153
		0.80	xxx	0.028	0.115	0.079	0.044	0.008	-0.022	-0.055
		1.00	xxx	0.009	0.008	0.010	0.008	0.012	0.006	0.009
	3PL	0.30	-2.747	-0.442	2.115	-4.948	1.288	-2.548	2.588	-1.143
		0.50	-2.438	-0.142	0.936	-1.856	0.578	-0.957	1.001	-0.556
		0.80	-1.152	-0.015	0.090	-0.248	0.113	-0.105	0.153	-0.090
		1.00	0.036	0.134	0.331	0.225	0.166	0.108	0.051	-0.074

Table 10. MEAN BIAS for Analytic essay item parameters by condition for Trait 3

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.069	0.377	0.274	0.140	0.020	0.127	0.259
		0.50	xxx	0.053	0.290	0.196	0.103	0.019	0.094	0.187
		0.80	xxx	0.026	0.122	0.091	0.051	0.020	0.042	0.077
		1.00	xxx	0.010	0.032	0.028	0.022	0.016	0.020	0.021
	3PL	0.30	2.650	0.088	3.260	4.453	0.873	2.178	2.621	0.633
		0.50	1.259	0.035	0.185	0.489	0.139	0.130	0.263	0.173
		0.80	0.354	0.037	0.157	0.067	0.101	0.041	0.073	0.138
		1.00	0.250	0.138	0.331	0.267	0.191	0.135	0.036	0.116
28	1PL	0.30	xxx	0.068	0.378	0.267	0.137	0.022	0.129	0.248
		0.50	xxx	0.053	0.280	0.197	0.104	0.019	0.094	0.178
		0.80	xxx	0.028	0.125	0.091	0.053	0.021	0.035	0.076
		1.00	xxx	0.010	0.038	0.033	0.021	0.018	0.020	0.026
	3PL	0.30	2.847	0.131	4.614	6.440	1.173	3.157	3.796	0.766
		0.50	2.396	0.043	1.518	2.081	0.496	0.980	1.206	0.408
		0.80	0.846	0.007	0.155	0.227	0.093	0.080	0.143	0.089
		1.00	0.176	0.135	0.342	0.245	0.191	0.120	0.049	0.122
36	1PL	0.30	xxx	0.069	0.377	0.252	0.133	0.019	0.121	0.234
		0.50	xxx	0.051	0.272	0.185	0.096	0.019	0.082	0.171
		0.80	xxx	0.028	0.123	0.088	0.049	0.020	0.034	0.066
		1.00	xxx	0.012	0.038	0.030	0.023	0.019	0.023	0.023
	3PL	0.30	2.874	0.152	5.103	7.196	1.286	3.530	4.227	0.798
		0.50	2.532	0.045	2.027	2.686	0.596	1.307	1.578	0.468
		0.80	1.113	0.011	0.219	0.349	0.116	0.134	0.199	0.095
		1.00	0.244	0.120	0.327	0.220	0.172	0.106	0.062	0.128

Table 11. MEAN RMSD for Analytic essay item parameters by condition for Trait 1

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.046	0.347	0.232	0.114	0.025	0.141	0.264
		0.50	xxx	0.036	0.260	0.170	0.087	0.023	0.102	0.187
		0.80	xxx	0.020	0.117	0.076	0.047	0.020	0.042	0.074
		1.00	xxx	0.011	0.033	0.029	0.022	0.020	0.023	0.026
	3PL	0.30	3.190	0.066	3.880	4.094	0.593	2.201	3.131	0.939
		0.50	1.764	0.012	0.302	0.458	0.140	0.142	0.321	0.218
		0.80	0.768	0.037	0.204	0.076	0.095	0.023	0.096	0.108
		1.00	0.183	0.123	0.332	0.191	0.182	0.108	0.045	0.112
28	1PL	0.30	xxx	0.047	0.351	0.228	0.107	0.025	0.140	0.253
		0.50	xxx	0.035	0.255	0.172	0.080	0.023	0.103	0.183
		0.80	xxx	0.021	0.112	0.081	0.043	0.017	0.041	0.071
		1.00	xxx	0.011	0.029	0.027	0.020	0.023	0.022	0.028
	3PL	0.30	3.388	0.084	5.445	5.852	0.745	3.142	4.476	1.173
		0.50	2.960	0.030	1.860	1.945	0.378	1.010	1.472	0.583
		0.80	1.298	0.014	0.233	0.235	0.088	0.096	0.183	0.099
		1.00	0.240	0.116	0.329	0.167	0.175	0.095	0.056	0.119
36	1PL	0.30	xxx	0.046	0.339	0.216	0.106	0.022	0.130	0.245
		0.50	xxx	0.036	0.246	0.163	0.077	0.020	0.093	0.172
		0.80	xxx	0.021	0.106	0.081	0.039	0.021	0.039	0.070
		1.00	xxx	0.012	0.032	0.028	0.021	0.019	0.024	0.025
	3PL	0.30	3.413	0.089	6.008	6.526	0.802	3.486	4.990	1.267
		0.50	3.088	0.043	2.404	2.474	0.425	1.320	1.894	0.675
		0.80	1.590	0.012	0.310	0.340	0.108	0.149	0.254	0.129
		1.00	0.239	0.104	0.321	0.143	0.161	0.085	0.056	0.129

Table 12. MEAN RMSD for Analytic essay item parameters by condition for Trait 2

Number of MC Items	Theta Estimation	Degree of Correlation	a	avg. b.	b ₁	b ₂	b ₃	b ₄	b ₅	b ₆
20	1PL	0.30	xxx	0.071	0.375	0.253	0.137	0.026	0.113	0.241
		0.50	xxx	0.053	0.278	0.187	0.099	0.023	0.083	0.172
		0.80	xxx	0.026	0.121	0.085	0.051	0.021	0.035	0.070
		1.00	xxx	0.012	0.034	0.027	0.022	0.020	0.020	0.020
	3PL	0.30	2.545	0.259	1.442	3.097	0.887	1.595	1.630	0.817
		0.50	1.311	0.035	0.045	0.372	0.147	0.094	0.222	0.146
		0.80	0.469	0.037	0.111	0.038	0.096	0.023	0.073	0.076
		1.00	0.218	0.150	0.336	0.278	0.189	0.133	0.042	0.074
28	1PL	0.30	xxx	0.071	0.374	0.252	0.134	0.021	0.111	0.229
		0.50	xxx	0.053	0.271	0.187	0.100	0.021	0.081	0.167
		0.80	xxx	0.028	0.125	0.086	0.045	0.021	0.034	0.065
		1.00	xxx	0.011	0.035	0.029	0.021	0.023	0.024	0.022
	3PL	0.30	2.723	0.388	1.958	4.433	1.170	2.290	2.335	1.062
		0.50	2.317	0.113	0.690	1.443	0.486	0.726	0.789	0.468
		0.80	0.920	0.007	0.076	0.169	0.095	0.067	0.117	0.069
		1.00	0.159	0.147	0.342	0.262	0.184	0.124	0.053	0.075
36	1PL	0.30	xxx	0.071	0.364	0.242	0.128	0.020	0.099	0.217
		0.50	xxx	0.053	0.269	0.178	0.094	0.022	0.076	0.155
		0.80	xxx	0.029	0.120	0.084	0.049	0.022	0.030	0.059
		1.00	xxx	0.011	0.035	0.031	0.026	0.023	0.021	0.024
	3PL	0.30	2.747	0.442	2.121	4.952	1.291	2.550	2.590	1.147
		0.50	2.438	0.143	0.939	1.858	0.580	0.958	1.001	0.557
		0.80	1.152	0.017	0.094	0.250	0.114	0.106	0.154	0.091
		1.00	0.131	0.135	0.333	0.231	0.167	0.109	0.057	0.078

Table 13. MEAN RMSD for Analytic essay item parameters by condition for Trait 3

Number of MC Items	Model	Theta Estimation	Degree of Correlation	a	b	c
20	1PL	Holistic	0.30	xxx	0.003	xxx
			0.50	xxx	0.003	xxx
			0.80	xxx	0.003	xxx
			1.00	xxx	0.003	xxx
		Analytic	0.30	xxx	0.005	xxx
			0.50	xxx	0.004	xxx
			0.80	xxx	0.006	xxx
			1.00	xxx	0.007	xxx
	3PL	Holistic	0.30	0.008	0.012	0.004
			0.50	0.007	0.012	0.004
			0.80	0.009	0.013	0.005
			1.00	0.009	0.012	0.004
Analytic		0.30	0.708	0.292	0.109	
		0.50	0.414	0.696	0.234	
		0.80	-0.062	0.196	0.065	
		1.00	0.179	0.121	0.008	
28	1PL	Holistic	0.30	xxx	0.003	xxx
			0.50	xxx	0.003	xxx
			0.80	xxx	0.005	xxx
			1.00	xxx	0.006	xxx
		Analytic	0.30	xxx	0.008	xxx
			0.50	xxx	0.006	xxx
			0.80	xxx	0.008	xxx
			1.00	xxx	0.007	xxx
	3PL	Holistic	0.30	0.007	0.017	0.005
			0.50	0.008	0.018	0.006
			0.80	0.006	0.018	0.006
			1.00	0.009	0.019	0.006
Analytic		0.30	0.054	0.103	0.039	
		0.50	0.221	0.304	0.113	
		0.80	-0.011	0.195	0.071	
		1.00	0.163	0.120	0.009	
36	1PL	Holistic	0.30	xxx	0.006	xxx
			0.50	xxx	0.006	xxx
			0.80	xxx	0.006	xxx
			1.00	xxx	0.006	xxx
		Analytic	0.30	xxx	0.010	xxx
			0.50	xxx	0.009	xxx
			0.80	xxx	0.009	xxx
			1.00	xxx	0.009	xxx
	3PL	Holistic	0.30	0.054	0.074	0.026
			0.50	0.007	0.021	0.006
			0.80	0.009	0.019	0.006
			1.00	0.009	0.022	0.007
Analytic		0.30	0.039	0.082	0.031	
		0.50	0.110	0.203	0.075	
		0.80	0.034	0.207	0.075	
		1.00	0.152	0.113	0.011	

Table 4. Mean BIAS for multiple choice item parameters by condition

Number of MC Items	Model	Essay Scoring	Degree of Correlation	a	b	c
20	1PL	Holistic	0.30	xxx	0.020	xxx
			0.50	xxx	0.017	xxx
			0.80	xxx	0.015	xxx
			1.00	xxx	0.015	xxx
		Analytic	0.30	xxx	0.039	xxx
			0.50	xxx	0.030	xxx
			0.80	xxx	0.019	xxx
			1.00	xxx	0.016	xxx
	3PL	Holistic	0.30	0.061	0.060	0.023
			0.50	0.065	0.064	0.024
			0.80	0.060	0.058	0.022
			1.00	0.061	0.057	0.021
		Analytic	0.30	2.058	0.435	0.151
			0.50	0.471	0.760	0.255
0.80			0.112	0.225	0.079	
1.00			0.194	0.148	0.020	
28	1PL	Holistic	0.30	xxx	0.022	xxx
			0.50	xxx	0.019	xxx
			0.80	xxx	0.016	xxx
			1.00	xxx	0.016	xxx
		Analytic	0.30	xxx	0.049	xxx
			0.50	xxx	0.036	xxx
			0.80	xxx	0.022	xxx
			1.00	xxx	0.016	xxx
	3PL	Holistic	0.30	0.063	0.072	0.025
			0.50	0.062	0.075	0.026
			0.80	0.061	0.070	0.024
			1.00	0.059	0.070	0.024
		Analytic	0.30	0.089	0.174	0.063
			0.50	0.252	0.419	0.146
0.80			0.079	0.249	0.093	
1.00			0.181	0.167	0.027	
36	1PL	Holistic	0.30	xxx	0.022	xxx
			0.50	xxx	0.019	xxx
			0.80	xxx	0.016	xxx
			1.00	xxx	0.016	xxx
		Analytic	0.30	xxx	0.049	xxx
			0.50	xxx	0.038	xxx
			0.80	xxx	0.022	xxx
			1.00	xxx	0.017	xxx
	3PL	Holistic	0.30	0.008	0.021	0.007
			0.50	0.058	0.073	0.025
			0.80	0.057	0.071	0.025
			1.00	0.055	0.072	0.025
		Analytic	0.30	0.071	0.149	0.053
			0.50	0.136	0.302	0.106
0.80			0.079	0.281	0.101	
1.00			0.171	0.167	0.029	

Table 5. Mean RMSD for multiple choice item parameters by condition

# MC Items	Theta Estimation	Degree of Correlation	Theta Estimate	1PL						3PL					
				True MC Theta	Holistic True Essay Theta	Average Theta	True MC Theta	Analytic True Essay Theta	Average Theta	True MC Theta	Holistic True Essay Theta	Average Theta	True MC Theta	Analytic True Essay Theta	Average Theta
20	Combined	0.3	Combined	0.001	0.001	0.001	-0.001	0.001	0.000	0.027	0.027	0.027	0.047	0.048	0.048
		0.5	Combined	-0.001	0.000	0.000	-0.001	0.001	0.000	0.025	0.026	0.026	0.017	0.016	0.017
		0.8	Combined	0.000	0.000	0.000	0.001	0.000	0.000	0.022	0.023	0.022	0.028	0.028	0.028
		1.0	Combined	-0.001	-0.001	-0.001	0.000	0.000	0.000	0.018	0.018	0.018	0.098	0.098	0.098
	Step	0.3	MC	0.002				0.004		0.027			0.112		
			Essay		-0.008			0.003			-0.001			-0.006	
			Average			-0.003			0.004			0.013			0.053
		0.5	MC	0.001				0.003		0.027			0.231		
			Essay		-0.010				-0.001			-0.004			0.001
			Average			-0.004			0.001			0.012			0.116
		0.8	MC	0.002				0.005		0.027			0.084		
			Average		-0.010				-0.007			-0.009			0.020
1.0	MC	0.001				0.004		0.028			0.113				
	Essay		-0.011				-0.011			-0.011			0.090		
	Average			-0.005			-0.004			0.008			0.101		
28	Combined	0.3	Combined	0.002	0.003	0.002	0.004	0.003	0.004	0.025	0.025	0.025	0.028	0.027	0.027
		0.5	Combined	0.001	0.002	0.002	0.002	0.002	0.002	0.022	0.022	0.022	0.028	0.027	0.027
		0.8	Combined	0.002	0.002	0.002	0.003	0.003	0.003	0.020	0.020	0.020	0.012	0.012	0.012
		1.0	Combined	0.003	0.003	0.003	0.003	0.003	0.003	0.017	0.017	0.017	0.095	0.095	0.095
	Step	0.3	MC	0.003				0.007		0.025			0.037		
			Essay		-0.007			0.004			0.000			-0.008	
			Average			-0.002			0.006			0.013			0.015
		0.5	MC	0.002				0.005		0.023			0.078		
			Essay		-0.009				-0.002			-0.004			0.002
			Average			-0.003			0.002			0.009			0.040
		0.8	MC	0.003				0.006		0.024			0.060		
			Average		-0.009				-0.006			-0.008			0.004
1.0	MC	0.005				0.006		0.024			0.106				
	Essay		-0.009				-0.010			-0.011			0.086		
	Average			-0.002			-0.002			0.006			0.096		
36	Combined	0.3	Combined	0.004	0.004	0.004	0.007	0.007	0.007	0.025	0.025	0.025	0.026	0.026	0.026
		0.5	Combined	0.013	0.009	0.011	0.005	0.005	0.005	0.025	0.024	0.025	0.026	0.026	0.026
		0.8	Combined	0.004	0.004	0.004	0.005	0.005	0.005	0.022	0.022	0.022	0.012	0.012	0.012
		1.0	Combined	0.003	0.004	0.004	0.005	0.005	0.005	0.019	0.019	0.019	0.087	0.087	0.087
	Step	0.3	MC	0.005				0.009		0.025			0.032		
			Essay		-0.008			0.005			-0.001			-0.007	
			Average			-0.001			0.007			0.012			0.012
		0.5	MC	0.005				0.007		0.026			0.050		
			Essay		-0.009				0.000			-0.003			0.000
			Average			-0.002			0.003			0.011			0.025
		0.8	MC	0.004				0.007		0.025			0.056		
			Average		-0.009				-0.006			-0.008			0.002
1.0	MC	0.004				0.007		0.025			0.098				
	Essay		-0.009				-0.009			-0.010			0.076		
	Average			-0.003			-0.001			0.008			0.087		

Table 2. Mean BIAS of theta estimates by condition

# MC Items	Theta Estimation	Degree of Correlation	Theta Estimate	1PL						3PL					
				True MC Theta	Holistic True Essay Theta	Average Theta	True MC Theta	Analytic True Essay Theta	Average Theta	True MC Theta	Holistic True Essay Theta	Average Theta	True MC Theta	Analytic True Essay Theta	Average Theta
20	Combined	0.3	Combined	0.494	0.990	0.512	0.672	0.693	0.341	0.437	1.097	0.588	0.546	0.883	0.442
		0.5	Combined	0.475	0.869	0.490	0.598	0.615	0.343	0.433	0.940	0.534	0.768	0.433	0.372
		0.8	Combined	0.445	0.643	0.453	0.463	0.473	0.345	0.421	0.656	0.451	0.506	0.319	0.281
		1.0	Combined	0.424	0.461	0.431	0.347	0.394	0.358	0.405	0.442	0.412	0.295	0.346	0.305
	Step	0.3	MC		0.457			0.457			0.439			0.488	
			Essay			0.757		0.484			0.881		0.299		
			Average				0.465		0.350			0.524			0.296
		0.5	MC		0.457			0.458			0.439			0.505	
			Essay			0.756		0.478			0.819		0.287		
			Average				0.479		0.357			0.515			0.301
		0.8	MC		0.457			0.457			0.439			0.442	
			Essay			0.756		0.471			0.760		0.295		
1.0	Average				0.499		0.366			0.508			0.294		
	MC		0.457			0.457			0.439			0.463			
	Essay			0.767		0.500			0.760		0.375				
Average				0.514		0.378			0.510			0.355			
28	Combined	0.3	Combined	0.433	1.024	0.517	0.600	0.745	0.328	0.390	1.111	0.586	0.410	0.999	0.482
		0.5	Combined	0.418	0.891	0.484	0.534	0.656	0.328	0.388	0.948	0.524	0.478	0.709	0.341
		0.8	Combined	0.395	0.637	0.425	0.418	0.485	0.324	0.378	0.650	0.427	0.455	0.349	0.254
		1.0	Combined	0.380	0.422	0.389	0.322	0.373	0.334	0.367	0.409	0.376	0.283	0.336	0.294
	Step	0.3	MC		0.404			0.404			0.392			0.395	
			Essay			0.756		0.483			0.881		0.326		
			Average				0.447		0.329			0.509			0.269
		0.5	MC		0.403			0.404			0.392			0.406	
			Essay			0.756		0.478			0.818		0.289		
			Average				0.459		0.335			0.496			0.258
		0.8	MC		0.403			0.403			0.392			0.393	
			Essay			0.757		0.471			0.760		0.287		
1.0	Average				0.476		0.342			0.485			0.262		
	MC		0.403			0.403			0.392			0.418			
	Essay			0.768		0.500			0.760		0.373				
Average				0.489		0.354			0.488			0.332			
36	Combined	0.3	Combined	0.387	1.047	0.523	0.541	0.789	0.329	0.361	1.122	0.587	0.373	1.033	0.503
		0.5	Combined	0.376	0.907	0.481	0.482	0.690	0.323	0.358	0.956	0.521	0.405	0.779	0.369
		0.8	Combined	0.358	0.635	0.407	0.382	0.497	0.310	0.351	0.646	0.413	0.417	0.379	0.243
		1.0	Combined	0.346	0.393	0.357	0.301	0.355	0.313	0.341	0.387	0.351	0.270	0.326	0.282
	Step	0.3	MC		0.364			0.364			0.362			0.363	
			Essay			0.757		0.482			0.881		0.341		
			Average				0.436		0.315			0.500			0.262
		0.5	MC		0.364			0.364			0.362			0.368	
			Essay			0.757		0.477			0.818		0.289		
			Average				0.445		0.319			0.484			0.242
		0.8	MC		0.364			0.364			0.362			0.363	
			Essay			0.756		0.471			0.760		0.286		
1.0	Average				0.459		0.325			0.472			0.248		
	MC		0.363			0.364			0.361			0.386			
	Essay			0.768		0.500			0.760		0.370				
Average				0.471		0.338			0.474			0.315			

Table 3. Mean RMSD of theta estimates by condition