3-2008

Scoring Multiple Choice Items: A Comparison of IRT and Classical Polytomous and Dichotomous Methods

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Recommended Citation
DeMars, C. E. (2008, March). Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Scoring Multiple Choice Items: A Comparison of IRT and Classical Polytomous and Dichotomous Methods

Christine E. DeMars

James Madison University

Author Note

Correspondence concerning this manuscript should be addressed to Christine DeMars, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg VA 22807. E-mail: demarsce@jmu.edu

Abstract

Four methods of scoring multiple-choice items were compared: Dichotomous classical (number-correct), polytomous classical (classical optimal scaling – COS), dichotomous IRT (3 parameter logistic – 3PL), and polytomous IRT (nominal response – NR). Data were generated to follow either a nominal response model or a non-parametric model, based on empirical data. The polytomous models, which weighted the distractors differentially, yielded small increases in reliability compared to their dichotomous counterparts. The polytomous IRT estimates were less biased than the dichotomous IRT estimates for lower scores. The classical polytomous scores were as reliable, sometimes more reliable, than the IRT polytomous scores. This was encouraging because the classical scores are easier to calculate and explain to users.

Scoring Multiple Choice Items: A Comparison of IRT and Classical Polytomous and Dichotomous Methods

Multiple choice items are often scored dichotomously by treating one option choice as correct and treating the distractors as equally wrong. In item response theory (IRT), the test can then be scored using the one-, two-, or three-parameter logistic model (1PL, 2PL or 3PL). In classical test theory, one point can be given for each correct answer and the test score can be the sum or mean of these points. These approaches do not take into account *which* incorrect distractor was selected by an examinee who failed to choose the most correct answer. Item response theory approaches to modeling each distractor individually include the nominal response (NR) model (Bock, 1972) and multiple-choice (MC) modifications of the NR model which take guessing into account (Samejima, 1981, Section X.7; Thissen & Steinberg, 1984). The NR and MC models require large samples and can be difficult to estimate for response categories with few respondents (De Ayala & Sava-Bolesta, 1999; DeMars, 2003; Thissen, Steinberg, & Fitzpatrick, 1989). Sympson and Haladyna (1988) developed a simple, non-parametric method of determining scoring weights for each category, which they termed *polyweighting*. Polyweighting is similar to Guttman's (1941) method of weighting categories to maximize internal consistency, which is a special case of generalized optimal scaling (McDonald, 1983). The purpose of this study is to compare the accuracy and precision of dichotomous and polytomous scoring using both IRT and classical models.

Scoring Models

The most common method of scoring items in classical test theory is number-correct (or percent-correct). This procedure gives all incorrect responses a score of zero. Alternatively, when the option score is the mean score of the examinees who chose that option, and the examinee score is the mean of the option scores selected, item-total correlations (Guttman, 1941) and coefficient alpha reliability (Haladyna & Kramer, 2005; Lord, 1958) are maximized. This scoring method has been called *classical optimal scaling* (COS) or *dual scaling* (McDonald, 1983; Warrens, de Gruijter, & Heiser, 2007), as well as *polyweighting* (Sympson & Haladyna, 1988) or *polyscoring* (Haladyna, 1990). The label *COS* will be used in what follows because *polyweighting* and *polyscoring* could easily be confused with other types of polytomous scoring such as the polytomous IRT models.

Sympson and Haladyna (1988) detailed a simple procedure to obtain the response options' different weights. In Sympson and Haladyna's algorithm, each examinee's initial score is the number-correct score. Based on these scores, the response option score (including the correct option) is calculated as the mean percentile rank of the examinees who chose that option. Total scores can then be re-computed with these new weights, followed by re-computation of the option scores, continuing until there is little change in the examinee or option scores. This procedure is often followed using the percent-correct scores or z-score in place of the percentile ranks (Haladyna, 1990; Haladyna & Kramer, 2005; Hendrickson, 1971; Warrens et al. 2007), in which case the option and examinee scores are equivalent to Guttman's (1941) procedure. Crehan and Haladyna (1994) advocated using the percentile rank because the category weight depends on the difficulty of the other test items when the total score is used in estimating the weights. Within a given test form, both methods should produce similar weights because of the monotonic relationship between total score and percentile rank.

In IRT, the 1PL, 2PL and 3PL models treat all incorrect options as a single category. In Bock's (1972) nominal response (NR) model and Thissen and Steinberg's (1984) multiple choice models, the probability of each option is modeled separately, without imposing an a-priori ordering to the options as the graded response (Samejima, 1969) and partial credit (Masters, 1982) models do. The NR model is:

$$P_{ij}(\theta) = \frac{e^{c_{ij} + a_{ij}\theta}}{\sum_{h=1}^{m_i} e^{c_{ih} + a_{ih}\theta}}, \quad (1)$$

where $P_{ij}(\theta)$ is the probability of an examinee choosing option j of item i given the examinee's proficiency θ, a_{ij} and c_{ij} are the item parameters for option j of item i, and m_i is the number of options for item i. The *a* parameters are related to the category discriminations, with positive values associated with categories that are generally used more as the trait level increases. The keyed answer should ideally have the highest *a*. Samejima (1981) modified the NR model to include a lower asymptote, equal to $1/m_i$ for each option. Thissen and Steinberg (1984) made a similar modification, except that the lower asymptote can vary for different options within the item. The lower asymptotes in these models sum to one across options within an item; conceptually, they represent the probability of guessing each option. These models are thus analogous to the 3PL model and would conceptually seem more appropriate for multiple choice items than the NR model.

Previous Findings

Classical Optimal Scoring.

Several studies have examined the reliability of COS. Haladyna (1990) summarized 20 studies and found a median reliability increase, compared to dichotomous scoring, of .042. Crehan and Haladyna (1994, p. 6) summarized the literature in saying: "The general findings are that internal consistency is slightly improved with the use of these linear methods, but correlations with external (validity) criteria do not improve." Frary (1989) reached a similar conclusion about COS and other empirical option-weighting methods. COS tends to make the most difference in reliability at the lower end of the score range (Haladyna & Kramer, 2005). This would be expected because the effects of weighting are greatest for examinees who choose incorrect options (Haladyna & Kramer, 2005). As a result, the overall reliability may not change a great deal, particularly for tests that are easy relative to the average ability.

In Haladyna and Kramer (2005), COS scores on shortened forms of a 400-item test had higher correlations with the full-length test score and pass/fail decisions more consistent with the full-length test than number correct scores. Comparisons were complicated by the fact that the short forms of the test were different for each scoring method, because each scoring method was used in item selection as well as scoring. This increased the realism of the study, but it prevents the separation of differences due to scoring from differences due to test construction.

IRT Methods.

Compared to dichotomous IRT models, the NR model tends to increase the information in the ability range where the item is difficult for the examinees. As discussed for COS, scoring the distractors differentially can not make much difference for examinees who mostly choose the keyed answer. Bock (1972) showed this in his illustration of the NR model. De Ayala (1992) found that in computerized adaptive testing (CAT) a target standard error could be met with fewer items using the NR model, compared to the 3PL model, for low-ability examinees. The number of items administered was similar for the NR and 3PL models for examinees of middle or high ability because the information in the item pool was similar for both models in these ability ranges. Thissen (1976) found that the information gain for low-ability examinees did not change the overall score reliability greatly for the test scores he studied, because there was little or no gain in information for the middle- or high-ability examinees. The marginal reliability was .72 using the NR model compared to .70 using the 2PL model. Childs, Dunn, van Barneveld, Jaciw, and McIlroy (2003) obtained greater information for low-ability examinees when errors on a medical examination were coded into three separate types rather than combined into a single category. However, they were not using typical multiple choice items; in the items they studied, examinees chose multiple options from a menu. They had several partial-credit categories for each item which were kept separate regardless of how the error categories were treated; the information gain would likely have been larger if they had compared a model where these partial credit categories were coded with the errors. In Si and Schumacker (2004), ability parameters were recovered with smaller RMSE (which implies more information) using polytomous models than dichotomous models. However, their data were generated with an ordinal model, the generalized partial credit (GPC) model, which may not be representative of most multiple-choice data.

Comparisons of IRT and non-IRT Methods.

Warrens, de Gruijter, and Heiser (2007) showed that for scales composed of dichotomous items, COS scores were quite similar to 2PL scores. For scales composed of ordered-category items, COS scores were similar to graded response scores. Also, they showed how good approximations of the item parameters could be derived from the option weights and the proportion choosing the option. They did not include the nominal response or multiple-choice models. The graded response results may generalize, but lack of an a-priori order to the response options of multiple-choice items may introduce some differences. Also, they did not compare dichotomous to polytomous scoring—the comparisons were between COS and IRT within dichotomous scoring or within ordered-polytomous scoring.

Haladyna and Kramer (2005) compared COS scores and number-correct scores to 3PL and 1 PL scores, but they did not include IRT polytomous models. As described earlier, they created shorter test forms and compared them to scores on the full-length test. Both number-correct scoring and COS scoring had higher correlations and more consistent pass/fail decisions than the 3PL and 1PL methods, perhaps because the maximum likelihood scores were poorly estimated when the raw scores were near zero.

Huynh and Casteel (1987) compared NR maximum-likelihood scores to number-correct scores. They found that pass/fail scores based on nominal scores were nearly identical to pass/fail scores based on raw scores when decisions were determined at the total test (30 or 36 items) level. For shorter subtests, consistency between the scores was greater at higher score levels. When decisions differed, raw score decisions were more consistent with teacher classifications. They did not compare polytomous IRT scores to dichotomous IRT scores, or polytomous non-IRT scores to dichotomous non-IRT scores, or polytomous IRT scores to polytomous non-IRT scores.

Rationale

The present study extends previous work, using the same data sets for both IRT and non-IRT scoring instead of focusing solely on either IRT or non-IRT methods. For the IRT scoring, the 3PL and NR models were used. For the non-IRT scoring, number-correct and COS were used. The research question was: How do the scoring methods compare in terms of bias, standard error, and reliability? Based on the previous literature, greater differences among the models were expected at the lower end of the ability distribution, where the NR model and COS were expected to be more accurate and reliable than the 3PL model and number correct scoring.

Method

Data Simulation

Empirical data were used to estimate realistic item parameters. These item parameters were then used as the true parameters to simulate the data. Data were simulated to fit either the NR model or a nonparametric model. The multiple-choice models of Thissen and Steinberg (1984) or Samejima (1981) might seem conceptually more appropriate for multiple-choice items than the NR model. However, Drasgow, Levine, Tsien, Williams, and Mead (1995) found that the NR model fit their multiple-choice data reasonably well and the multiple-choice models did not improve the fit. Additionally, Thissen, Steinberg, and Fitzpatrick (1989) and Thissen and Steinberg (1984) discussed how using the multiple-choice model for scoring individuals could be problematic. Their concern was that the response curve for the correct answer can be nonmonotonic in the multiple choice model, which would effectively sometimes lead to a lower θ-estimate for choosing the correct answer. Finally, attempts at fitting the multiple-choice models to the empirical data in this study yielded option characteristic curves with unexpected shapes. For these reasons, the NR model was selected as the parameteric IRT model. A nonparametric model was also included to simulate data that did not fit the NR model as well. When data are generated with a particular model, scores based on that model will tend to be most accurate. The nonparametric model would be less likely to advantage the NR scores.

The empirical data set had 20 items, each with 5 responses. The test had been administered as a final exam to 804 first- or second-year college students enrolled in a course required for several different majors. The instructor had attempted to use common errors or misconceptions in the distractors. Response options chosen by fewer than 15 students were coded as ignorable non-response in the item calibration and were not used in the ensuing data simulation. The resulting test had five 3-option items, three 4-option items, and twelve 5-option items.

Nominal Response Data. The NR item parameters were first estimated from the empirical data, using Multilog 7 (Thissen, Chen, & Bock, 2003). Default options were used, except that the number of quadrature points was increased to 20, uniformly spaced from -4 to 4. One very easy item had very low estimated discrimination for the correct answer; before simulating the data, the parameters for this item were replaced with those from the next-easiest item, but with the option curves shifted to the left to make the difficulty of the correct answer more similar to that of the original item. Next, 93,000 simulees were generated, 3000 at each 0.2 θ value from -3 to

3. Fixed θ's were used so that the standard error at each θ value could be estimated. Item responses were generated using these θs and the item parameters estimated earlier for the empirical data.

Nonparametric Data. The nonparametric option characteristic curves were estimated from the empirical data, using TestGraf98 (Ramsay, 2001). The θ values listed above were again used to simulate the NP responses. The smoothing parameter, *h*, was set to .27; experimentation with higher values led to smoother ICCs but somewhat higher reliability of simulated raw scores than was found for the empirical data. The correct answer curve was constrained to be monotonic.

Analyses

The simulated data sets were scored with NR and 3PL MAP (modal a-posteriori) scores using a standard normal prior. Bayesian scores were selected due to the problems Haladyna and Kramer (2005) and Huynh and Casteel (1987) encountered when estimating ML scores for low-ability examinees or short subtests. The COS scores were calculated using Sympson and Haladyna's, (1988) procedures. Number-correct scores were also calculated. The MAP and COS scores were calculated for the NR data using three sets of parameters or weights: the true parameters (NR-data, true-parameter-scoring) and parameters estimated from a calibration sample of 2500 or 250 simulees (NR-data, estimated-parameter-scoring). True item parameters would not be known with non-simulated data; the true-parameter-scoring was used to explore the differences between the scoring models without contamination from item parameter estimation. The estimated-parameter-scoring was used to provide more realistic results. The calibration sample of 250 is far smaller than is typically recommended for the NR model; it was included to see how the scoring methods compared with very small samples. Finally, the MAP and COS scores were calculated for the NP data using two sets of parameters or weights, again estimated from a calibration sample of 2500 or 250 simulees.

For the NR scoring, the true item parameters were simply the parameters used to generate the data. For the 3PL scoring, the true item parameters were defined as the parameters which provided the closest fit between the 3PL item response function and the true NR option response function for the correct answer. To calculate these item parameters, the proportion of correct answers was calculated for each of 50 quadrature points spaced equally from -3 to 3. These proportions, weighted by the proportion of the standard normal distribution at each quadrature point, and the corresponding theta values were used in MULTILOG 7 to conduct a "fixed theta" analysis using the 3PL model. No real data were simulated; the proportions were strictly model-based. The resulting item parameters were defined as the true 3PL parameters. The true option weights for the COS scoring were defined as the predicted mean percentile rank of those who would choose each option, based on the true NR item parameters and a normally distributed population. Again, these were model-based and no simulees were used in their calculation. Finally, for the number-correct scoring the option weights were 1 for the keyed answer and 0 for any other option.

The COS score and number-correct score metrics were not comparable with each other or with the NR or 3PL metrics. These scores were transformed to Z-scores. While the true θs were uniformally distributed to obtain an accurate standard error at each individual θ, in calculating the constants for the transformation, the simulees were weighted based on a normal distribution density to yield the Z-scores that would have resulted from a normal θ-distribution. Note that the scores were not normalized; the weights to form the normal distribution were based on the quadrature point values used to generate the data. The COS and number-correct scores had a non-linear relationships with the θ scale, so the COS and number-correct Z-scores were *not* normally distributed. Figure 1 shows the θ scale along with the number-correct "true-score" scale and COS "true-score" scale. The number-correct true scores were calculated at each quadrature point in the typical way: the sum of the probability of correct response, with the probabilities based on the generating item parameters. The true scores were then converted to Z-scores as described above. The COS true scores were calculated at each quadrature point by multiplying each option weight by the probability of that option, based on the item parameters, and summing across options and across items. Again, these true scores were then converted to Z-scores. Figure 1 shows that both number-correct and COS scores are spread apart at the higher θ levels, and COS scores are squeezed together at the lower θ levels. Scores that are normally distributed in the θ metric would be positively skewed in the COS metric; consistent with this, Haladyna and Kramer (2005) noted that COS scoring tended to result in positively skewed distributions. This non-linear relationship among the scales is not unique to the NR model. Wright and Stone (1979, Chapter 1) discussed this issue in the context of the Rasch model and it can apply to any comparisons between IRT and classical metrics.

Next, estimated item parameters were used to re-score the same response strings. The scoring sample was not well-suited for item parameter estimation because the simulees were uniformly distributed. Instead, for item parameter estimation, samples of 2500 or 250 simulees were drawn from a normal (0, 1) population. Response strings were generated for these samples as they were for the NR-data scoring samples, and the nominal response parameters, 3PL parameters, and COS option weights were estimated. The IRT parameters were estimated using Multilog 7. For both models, 20 quadrature points evenly spaced between -4 and 4 were used for the population distribution. For the 3PL model, a prior of (-1.4, .5) was used for the log of the asymptote, a prior of (1.36, .5) was used for the slopes, and a prior of (0, 2) was used for the difficulties. Otherwise, program defaults were used. COS weights were estimated from the calibration sample using Sympson and Haladyna's (1988) procedures.

These estimated parameters and weights were used to re-score the response strings previously scored with the true parameters, and the COS scores were converted to Z-scores (calculating the transformation parameters based on the true COS scores, weighted by the standard-normal density of the θ). Thus, estimation error in the item parameters/weights due to sampling was incorporated within the variance of the θ estimates for each true θ. No re-scoring was necessary for the number-correct scoring because the option weights were constant.

The responses generated to fit the nonparametric model (NP-data) were scored using methods similarly to the NR-data, estimated-parameter-scoring. Calibration samples of 2500 or 250 simulees were drawn from a normal (0, 1) population. Response strings were generated based on the NP model. Based on these response strings, the nominal response parameters, 3PL parameters, and COS option weights were estimated. The NP-data scoring sample was then scored with these parameters and weights as well as with number-correct scoring.

Results

Estimation accuracy was evaluated based on bias, standard error, and reliability. For the NR and 3PL scores, bias was defined as $Bias_q = \sum_{r=1}^{3000} \frac{1}{3000}\left(\hat{\theta}_{qr} - \theta_q\right)$, where $\hat{\theta}_{qr}$ is the estimated score for simulee *q* in replication *r*, and θ_q is the generating score for simulee *q* ($\hat{\theta}_{qr}$ is used as a generic symbol for the estimated score and includes the NC and COS estimates, not just the IRT-based estimates). Bias for the NR-data, true-parameter-scoring is shown in the top of Figure 2. The NR and 3PL MAP scores were positively biased at the low end of the θ range and negatively biased at the high end, as would be expected for Bayes estimates. At the low end of the θ scale, the 3PL scores were more biased than the NR scores. Because the bias is proportional to reliability, this likely indicates that the 3PL scores are less reliable in this range. The COS and number-correct scores were negatively biased at both extremes, due to differences in the metric as seen earlier in Figure 1. These scores were nearly unbiased estimates of the true scores in the COS or number-correct metric, but the true scores in these metrics were biased relative to θ.

Bias was also calculated for these data using estimating scoring parameters or weights. These values are displayed in the left half of Figure 3. The number-correct scores were of course unchanged but are repeated for easier comparison. The patterns are the same as those using the true scoring parameters, except that the bias in the IRT MAP scores increased in absolute value. This bias was also greater with the smaller calibration sample size. Again, this would be expected because the absolute value of the bias in MAP scores is inversely related to reliability.

Bias for the NP data is illustrated in the right half of Figure 3. The COS scores, and the number-correct scores for higher θ's, were less biased than they were for the NR data because the true-score units were more comparable to the θ units using the NP data. The IRT scores were more biased, in absolute value, than the NR data IRT scores.

The standard error at quadrature point q was defined as: $SE_q = \sqrt{\sum_{r=1}^{3000} \frac{1}{3000}\left(\hat{\theta}_{qr} - \bar{\hat{\theta}}_{q.}\right)^2}$,

where $\hat{\theta}_{qr}$ is the estimated score for simulee q in replication r, and $\bar{\hat{\theta}}_{q.}$ is the mean score for simulee q. Note that $\bar{\hat{\theta}}_{q.}$ is not equal to the generating θ if there is any bias and thus this definition of SE is not necessarily the same as the RMSE. These standard errors are graphed in the lower half of Figure 2 for the NR-data, true-parameter-scoring. Comparing these standard errors is complicated by the non-linear relationships among the scales. Looking back at Figure 1, at the low-end of the scale the COS units are smaller than the number-correct or θ units, and at the high end of the scale the COS and number-correct units are considerably larger than the θ units. In Figure 3, the COS standard errors appear to be larger than the standard errors for the other scores for low θs, but this may be due to the smaller units for the COS scores. Similarly, at the high end of the scale the COS standard errors appear smaller, but this may be due to the larger units. In addition to unit differences, the bias of the NR and 3PL scores yields smaller variance in score estimates, and thus the smaller standard errors for these scores may be larger relative to the score variance. The standard errors were higher for the 3PL scores than the NR scores, as would be expected if the NR model added any information.

The standard errors using estimated scoring parameters for the NR and NP data are shown in Figure 4. The COS standard errors for low θs were somewhat higher using estimated parameters, especially for the smaller calibration sample. The IRT standard errors did not seem to depend on the calibration sample size. However, because the bias increased in absolute value for the smaller samples, the total error (RMSE) would be greater for the smaller calibration sample.

Reliability, unlike standard error, does not depend on the scale of the measurement units and thus is helpful for comparing the scores in this context. Reliability was estimated in several ways for this study: (a) 1 - the ratio of mean empirical squared standard error (calculated as described above) to observed score variance; (b) squared correlation with generating θ; (c) squared correlation with true COS score or true number-correct score corresponding to the generating θ, calculated as described in the Method section for Figure 1; (d) coefficient alpha; and (e) 1 - model-based error variance, with the model-based error calculated from the MAP information function and integrated over the θ distribution (integration approximated on 50 quadrature points from -3 to 3). Methods *c* and *d* were used only for the COS and number-correct scores, and method *e* was used only for the NR and 3PL scores. Note that methods *a* and *e* use different equations because the standard error is different; in method *a*, it is the observed standard deviation around the mean $\hat{\theta}$ conditional on the true θ; in method *e*, the model-based standard error is an estimate of the standard deviation of the true θ conditional on $\hat{\theta}$ (for Bayesian scores, these are not the same). For all methods, each quadrature point was weighted by its density in a normal distribution; without weighting, reliability would have been higher because the uniform distribution used in generating the data has greater variance than a standard normal distribution.

Reliability was calculated separately for true θ's above and below zero because the differences between the dichotomous and polytomous scoring methods were expected to be larger for lower scores. To make these reliability estimates comparable to those from the full range, the estimates were corrected to estimate what the reliability would have been if the distribution of scores above zero had been the mirror image of the distribution below zero (or conversely when estimating reliability for the upper half of the distribution). For the information-based reliability, this simply meant using a constant true score variance of 1. For the methods based on the ratio of squared empirical standard error to observed score variance or correlations between estimated and true scores (methods *a*, *b*, and *c*), a constant was added to each estimated score so that the expected value of the score estimate would be zero when the true score was zero. Then a twin was created for each simulee by taking the negative of the simulee's estimated and true score, and the estimates were calculated using the original and negative twin simulees. The final method, coefficient alpha, was calculated by estimating coefficient alpha in the half sample and then adjusting it based on the ratio of the score variance in the half sample to the score variance in the half-sample combined with negative twins.

The reliability estimates are shown in Tables 1-5. Notice that columns 2 and 3 are not necessarily identical because of the non-linear relationship between θ and the number-correct or COS true score. If one considers the number-correct or COS score to be an estimate of θ, the correlation between the score estimate and θ is the more appropriate measure of reliability. If one considers the number-correct or COS score to be an estimate of the expected value of that score across parallel forms, the usual CTT perspective, the correlation between the score estimate and θ transformed to the number-correct or COS metric is the more appropriate measure. When the two differed, the former estimate was generally lower and the latter estimate was nearly equivalent to the reliability estimate based on the empirical standard error and empirical standard deviation of the scores. Coefficient alpha generally estimated this latter index well.

Using the other reliability estimates as a criterion, the reliability estimate based on the information function predicted the reliability well using the true parameters, or, for the NR scores, the larger calibration sample. For the 3PL model, for the larger calibration sample this reliability was an underestimate for the lower scores and an overestimate for the higher scores. The guessing parameter used to generate the NR data was zero, but the estimated parameter used in the scoring ranged from about .10 to .25. A non-zero c will (in this case, falsely) decrease the information for lower scores. Finally, for the smallest samples, using the information function overestimated the reliability for both 3PL and NR scores.

Reliability for the COS scores was higher than reliability for the number-correct scores, with the exception of the NP data, small calibration sample. The difference was greater for the NR data than for the NP data. The difference was *slightly* greater for low scores then for high scores; in fact, it was reversed for higher scores in the NP data, small calibration sample. The weights for the COS scoring may have been poorly estimated in this sample.

For the IRT methods, the reliability estimates were somewhat larger for the NR scores compared to the 3PL scores, again with the exception of the NP data, small calibration sample. The difference was generally smaller than the reliability difference for the COS and number-correct scores, but again there were greater differences for the NR data than for the NP data.

Differences were about the same for low scores and high scores, consistent with the finding that the empirical standard error was stable throughout the θ range, especially for the true-parameter scoring. For the NP data with the smallest calibration sample, the 3PL scores were slightly less biased and had slightly smaller standard errors than the NR scores and thus were actually more reliable. The simpler 3PL model parameters may have been estimated better than the NR model parameters in this condition.

The reliability of the 3PL scores was either the same or higher than the reliability of the number-correct scores. The reliability of the NR scores was either the same or slightly *lower* than the reliability of the COS scores. This result was unexpected, particularly for the NR data when the generating item parameters were used in the scoring. Any advantage to COS scoring was expected for the NP data, but the slight difference between COS and NR scoring seemed to be about the same whether the data followed the parametric model or not. Additionally, reduced sample size for calibration impacted both polytomous models.

Conclusions

Consistent with previous research, the polytomous scoring methods yielded more reliable scores estimates. The expectation of greater differences for lower θs was weakly supported for the classical scores and not supported for the IRT scores. This may be because for this test the reliability was lower for the higher θs and thus scores in this range could benefit from any additional information. The difference between polytomous and dichotomous scoring was larger for the classical scores (COS and number correct) than for the IRT scores, perhaps because the 3PL, unlike the number-correct scoring, utilized varying weights for the items (the NR model effectively adds varying weights for the distractors). The COS and number-correct scoring may have differed more because number-correct scoring uses neither varying weights for the items nor varying weights for the options.

The COS score estimates were at least as reliable as the NR model estimates. COS scores do not require any knowledge of IRT and can be calculated using any general-purpose statistical software. COS scores may be more interpretable for non-psychometricians because each option is literally assigned a weight. No likelihood functions are involved.

A limitation to this study was that the item parameters were based on a single test. It did not seem realistic to draw the item parameters from typical distributions because there has not been enough published research using the nominal response model with multiple choice data to know what typical distributions might be. The amount of information available in the distractors has a large impact on whether or not the polytomous models are more reliable than their dichotomous counterparts, so this is not a trivial limitation.

The smallest calibration sample size was smaller than would be typically used for the 3PL or NR models. Predictably, the absolute value of bias was higher and reliability was lower using the scoring parameters from the small calibration sample. The estimated information function yielded overestimates of the score reliability; in real-data studies where true θs are unknown, this would have created the false appearance that the IRT scores were more reliable than the classical scores. The COS score reliability also decreased for the smaller calibration sample, so 250 seems to be smaller than desirable for estimating the COS weights accurately as well. Further research is needed to formulate sample size recommendations for COS scoring.

The small increases in reliability from using polytomous models may be large enough to be meaningful in some contexts. The empirical data in this example came from a final exam; any increase in reliability would be beneficial, especially considering that the reliability of the scores was fairly low. Reliability would be particularly important near grades of C+/B-, because this course was among a cluster of courses in which these students needed to maintain an average of B or higher to continue in their program. Overall, 38% of these students earned a C+ or lower in the course (not necessarily on the final exam), so the reliability of scores below the mean was important. The utility of a small increase in reliability will vary with the context, and the complications introduced by COS or NR scoring need to be weighed against the potential benefits in a given context.

References

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

Childs, R. A., Dunn, J. L., van Barneveld, C., Jaciw, A. P., & McIlroy, J. H. (2003). Differential weighting of errors on a test of clinical reasoning skills. *Academic Medicine, 78* (10), S62-S64.

Crehan, K. D., & Haladyna, T. M. (1994). *A comparison of three linear polytomous scoring methods*. Las Vegas, NV: University of Nevada-Las Vegas, Counseling and Educational Psychology. (ERIC Document Reproduction Service No. ED 377246)

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16*, 327-343.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3-19.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement, 27*, 275-288.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B, & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79-96.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment: A survey of logical problems and research techniques, with illustrative appliction to problems of vocational selection, school success, marriage, and crime* (pp. 341-348). New York: Social Science Reseach Council.

Haladyna, T. M., & Kramer, G. (2005, April). An empirical investigation of poly-scoring of multiple-choice item responses. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. *Applied Measurement in Education, 3*, 231-244.

Hendrickson, G. F. (1971). The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement, 8*, 291-296.

Huynh, H. & Casteel, J. (1987). The usefulness of the Bock model for scoring with information from incorrect responses. *Journal of Experimental Education, 55*, 131-136.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, *23*, 291-296.

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

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika*, *48*, 377-391.

Ramsay, J. O. (2001). TestGraf98 [Computer software]. Montreal, Quebec: author. Available from http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplements, 17*.

Samejima, F. (1981). *Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple choice item. Final report*. Arlington, VA: Office of Naval Research. (ERIC Document Reproduction Service No. ED 212652)

Si, C.-F., & Schumacker, R. E. (1983). Ability estimation under different item parameterization and scoring methods. *International Journal of Testing*, *4*, 137-181.

Sympson, J. B., & Haladyna, T. M. (1988, April). An evaluation of "polyweighting" in domain-referenced testing. Paper presented in C. E. Davis (Chair), *New developments in polychotomous item scoring and modeling*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 294911)

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501-519.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13*, 201-214.

Thissen, D., Chen, W.-H, & Bock, D. (2003). MULTILOG 7. [Computer software]. Lincolnwood, IL: SSI.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.

Warrens, M. J., de Gruijter, D. N. M., & Heiser, W. J. (2007). A systematic comparison between classical optimal scaling and the two-parameter IRT model. *Applied Psychological Measurement, 31*, 106-120.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Table 1

Reliability Estimates for NR Data Using True Parameters for Scoring

θ Range and Model	Reliability Estimator				
	$1-\frac{\sigma^2_{error}}{\sigma^2_{observed}}$	$r^2_{\hat{\theta},\theta}$ or $r^2_{\hat{Z},\theta}$	$r^2_{\hat{Z},Z}$	Coefficient alpha	Based on Information
Overall					
3PL	.68	.67			.67
NR	.72	.71			.70
NC	.66	.65	.66	.66	
COS	.73	.69	.73	.72	
θ < 0					
3PL	.72	.72			.71
NR	.76	.76			.77
NC	.70	.70	.70	.70	
COS	.77	.76	.77	.76	
θ > 0					
3PL	.61	.61			.63
NR	.65	.65			.64
NC	.60	.59	.60	.61	
COS	.66	.65	.66	.66	

Table 2

Reliability Estimates for NR Data, Scoring based on Calibration Sample of 2500

θ Range and Model	Reliability Estimator $1-\frac{\sigma^2_{error}}{\sigma^2_{observed}}$	$r^2_{\hat{\theta},\theta}$ or $r^2_{\hat{Z},\theta}$	$r^2_{\hat{Z},Z}$	Coefficient alpha	Based on Information
Overall					
3PL	.67	.66			.66
NR	.71	.70			.70
NC	.66	.65	.66	.66	
COS	.72	.69	.72	.72	
θ < 0					
3PL	.72	.72			.66
NR	.76	.76			.77
NC	.70	.70	.70	.70	
COS	.76	.76	.76	.76	
θ > 0					
3PL	.61	.61			.66
NR	.65	.64			.64
NC	.60	.59	.60	.61	
COS	.65	.64	.65	.65	

Table 3

Reliability Estimates for NR Data, Scoring based on Calibration Sample of 250

θ Range and Model	Reliability Estimator $1-\frac{\sigma^2_{error}}{\sigma^2_{observed}}$	$r^2_{\hat{\theta},\theta}$ or $r^2_{\hat{Z},\theta}$	$r^2_{\hat{Z},Z}$	Coefficient alpha	Based on Information
Overall					
3PL	.66	.66			.70
NR	.68	.67			.74
NC	.66	.65	.66	.66	
COS	.70	.67	.70	.69	
θ < 0					
3PL	.71	.71			.68
NR	.73	.73			.81
NC	.70	.70	.70	.70	
COS	.74	.73	.74	.73	
θ > 0					
3PL	.60	.60			.72
NR	.62	.61			.66
NC	.60	.59	.60	.61	
COS	.63	.62	.63	.63	

Table 4

Reliability Estimates for NP Data, Scoring based on Calibration Sample of 2500

θ Range and Model	Reliability Estimator				
	$1-\frac{\sigma^2_{error}}{\sigma^2_{observed}}$	$r^2_{\hat{\theta},\theta}$ or $r^2_{\hat{Z},\theta}$	$r^2_{\hat{Z},Z}$	Coefficient alpha	Based on Information
Overall					
3PL	.69	.69			.69
NR	.71	.70			.71
NC	.68	.68	.68	.68	
COS	.71	.70	.71	.70	
θ < 0					
3PL	.70	.70			.66
NR	.72	.72			.76
NC	.69	.69	.69	.69	
COS	.72	.72	.72	.72	
θ > 0					
3PL	.68	.68			.71
NR	.69	.69			.67
NC	.68	.68	.68	.68	
COS	.69	.69	.69	.69	

Table 5

Reliability Estimates for NP Data, Scoring based on Calibration Sample of 250

θ Range and Model	Reliability Estimator				
	$1-\frac{\sigma^2_{error}}{\sigma^2_{observed}}$	$r^2_{\hat{\theta},\theta}$ or $r^2_{\hat{Z},\theta}$	$r^2_{\hat{Z},Z}$	Coefficient alpha	Based on Information
Overall					
3PL	.68	.68			.70
NR	.65	.65			.76
NC	.68	.68	.68	.68	
COS	.67	.66	.67	.67	
θ < 0					
3PL	.70	.70			.65
NR	.67	.67			.81
NC	.69	.69	.69	.69	
COS	.69	.69	.69	.69	
θ > 0					
3PL	.66	.66			.75
NR	.63	.63			.70
NC	.68	.68	.68	.68	
COS	.65	.65	.64	.65	

Figure Captions

Figure 1. Comparison of score metrics.

Figure 2. Bias and standard error in score estimates, using the true item parameters.

Figure 3. Bias in score estimates, using estimated item parameters.

Figure 4. Standard errors of score estimates, using estimated item parameters.

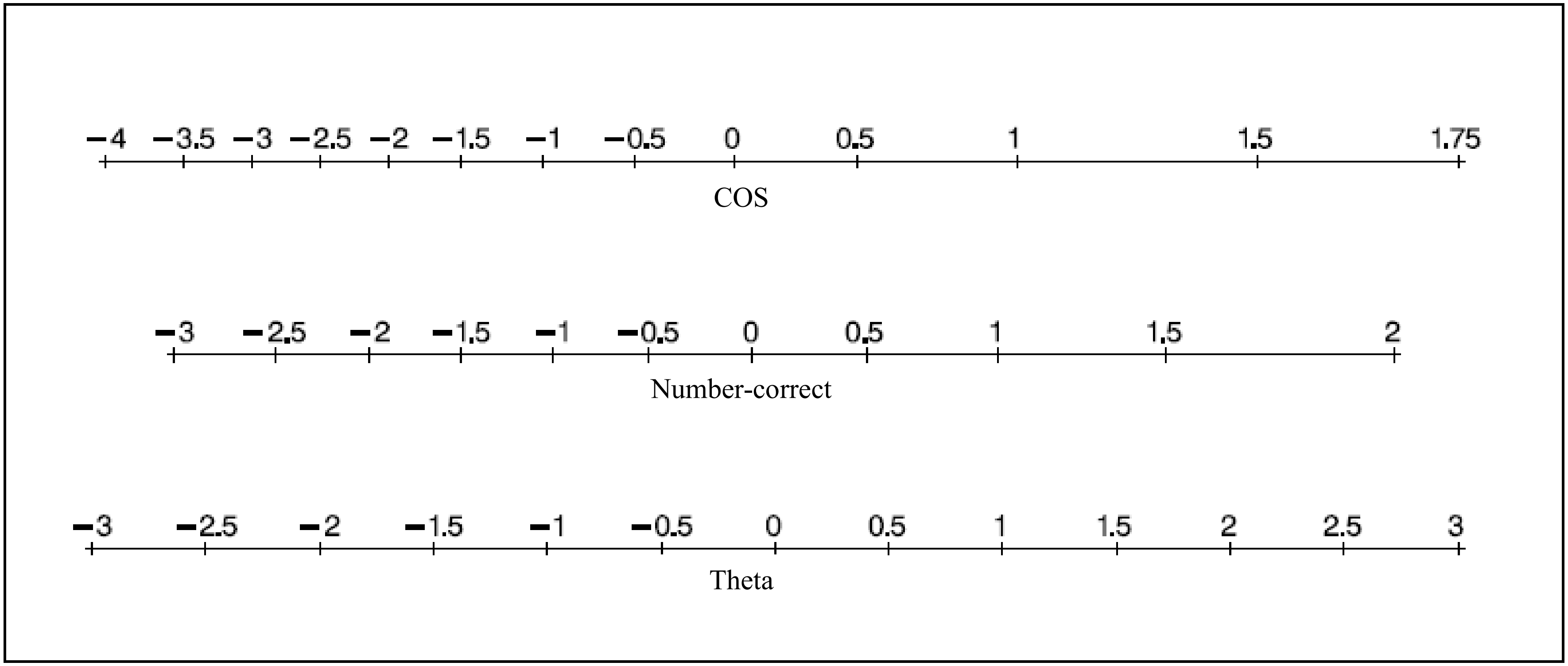


Figure 1. Comparison of score metrics.

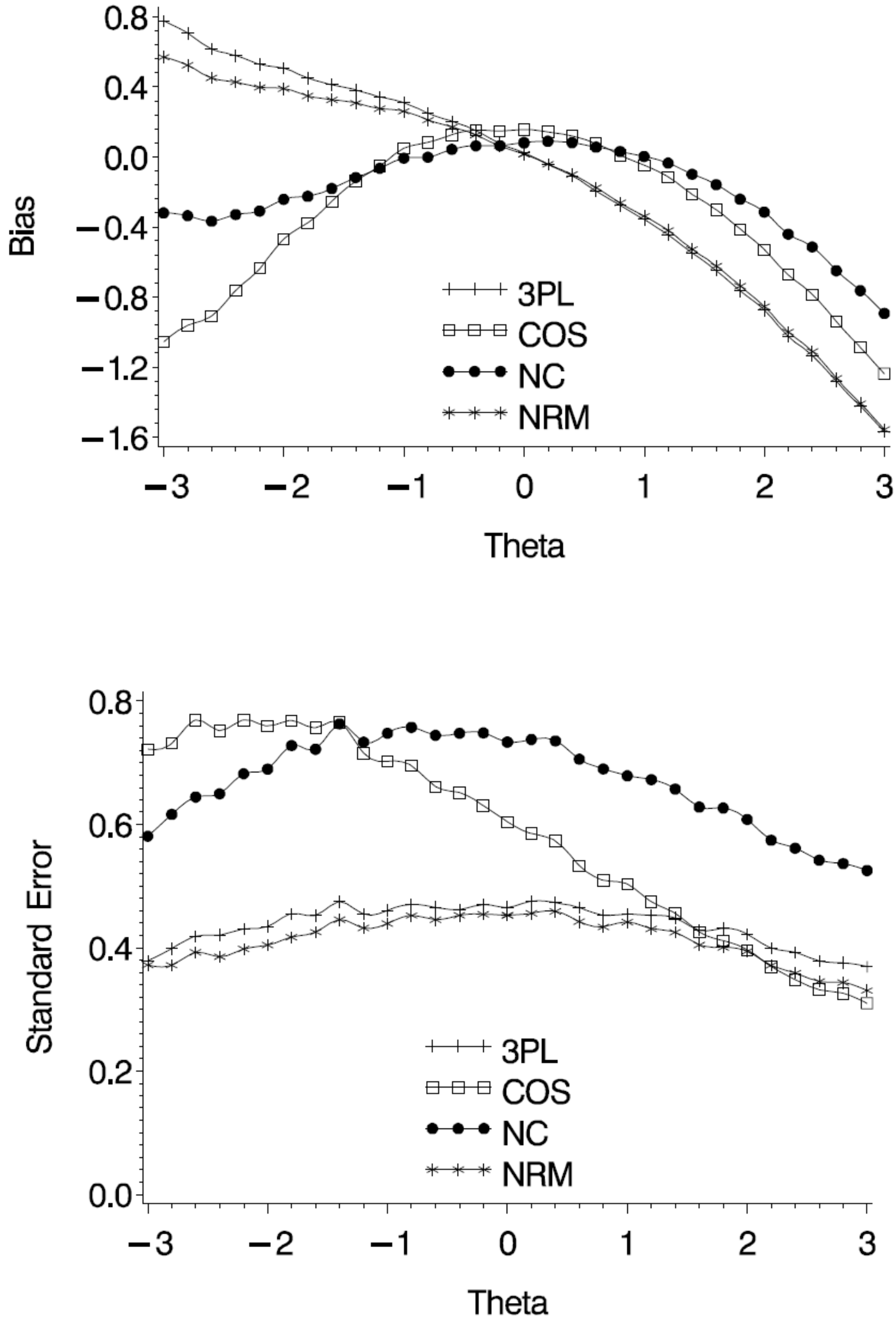
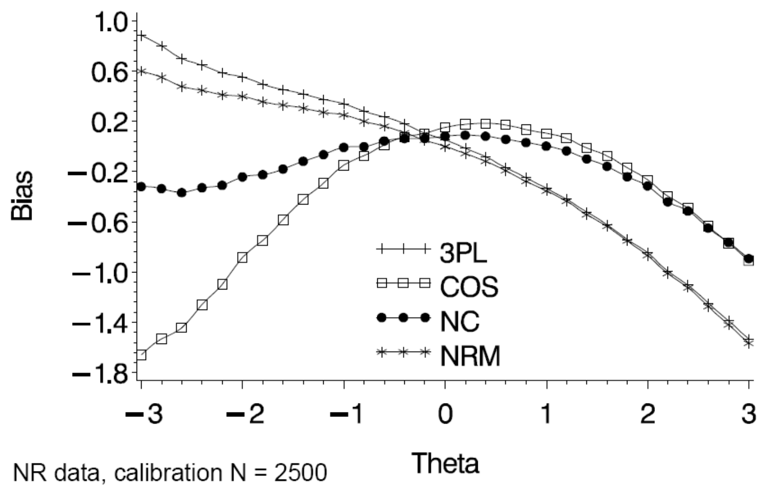
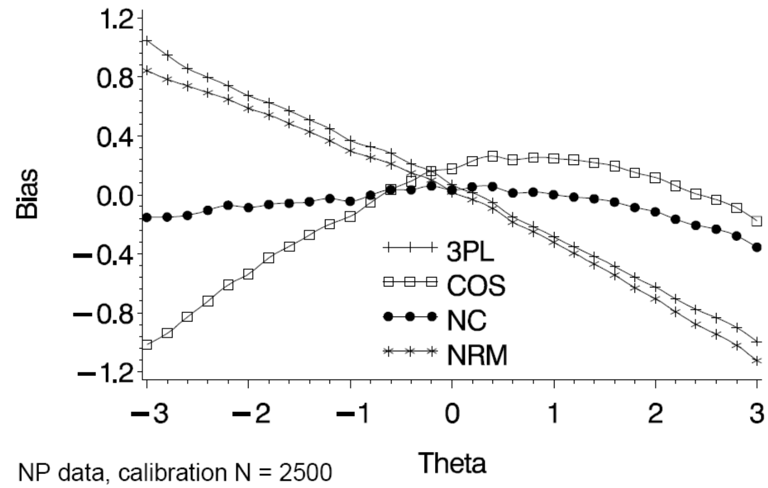


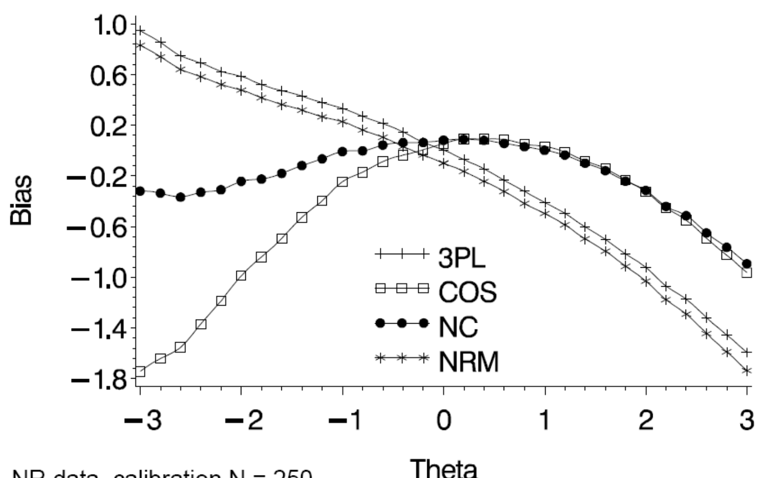
Figure 2. Bias and standard error in score estimates, using the true item parameters.



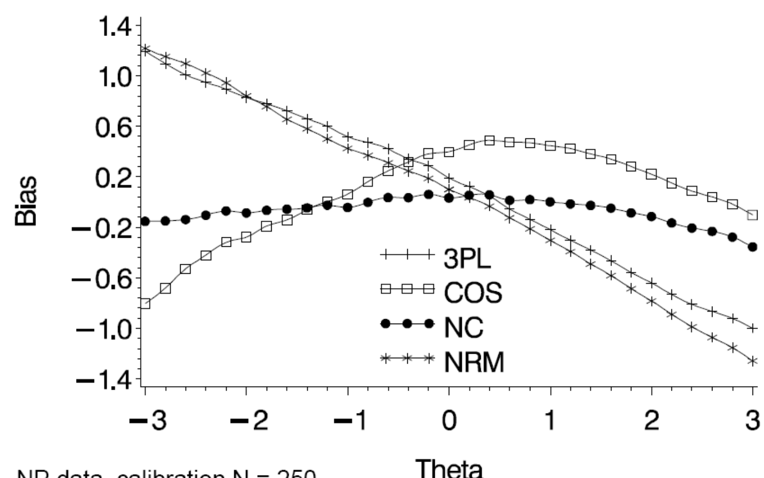
NR data, calibration N = 2500



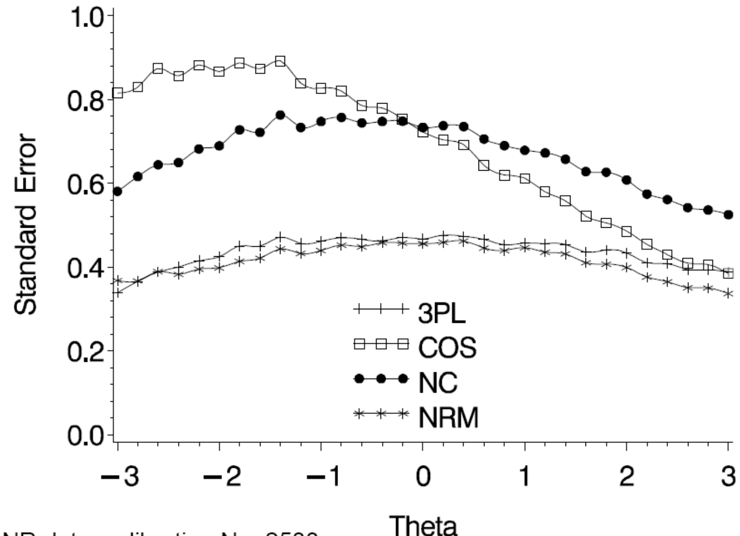
NP data, calibration N = 2500



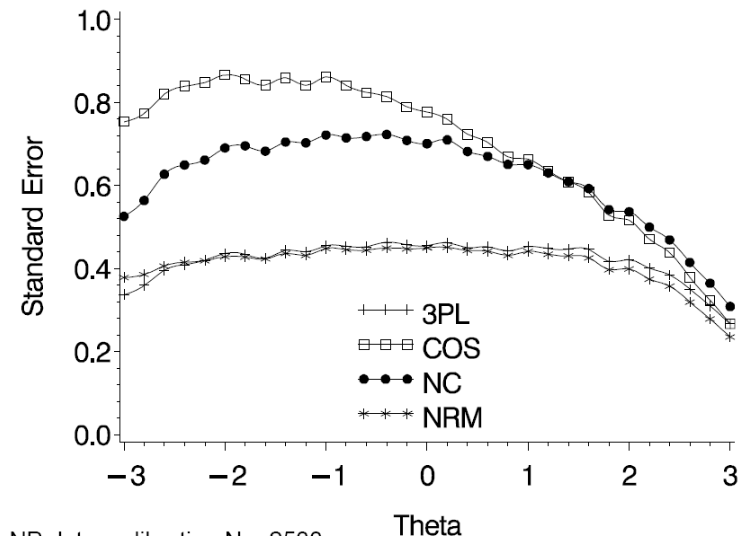
NR data, calibration N = 250



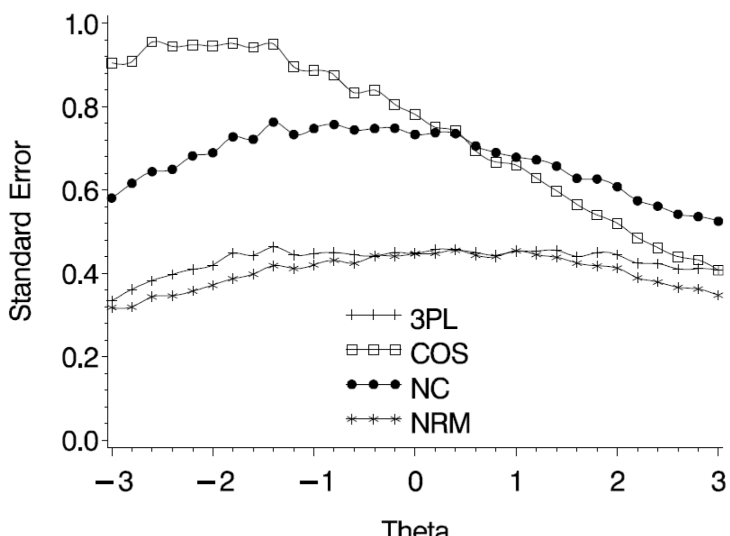
NP data, calibration N = 250



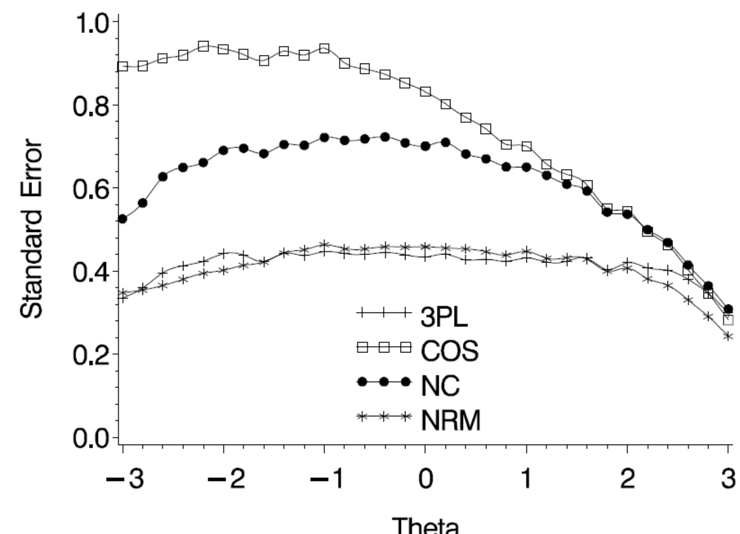
NR data, calibration N = 2500



NP data, calibration N = 2500



NR data, calibration N = 250



NP data, calibration N = 250