ACT Research Report Series 2000-2

Effects of Scale Transformation and Test Termination Rule on the Precision of Ability Estimates in CAT

Qing Yi

Tianyou Wang

Jae-Chun Ban

ACT

January 2000

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Effects of Scale Transformation and Test Termination Rule on the Precision of Ability Estimates in CAT

Qing Yi
Tianyou Wang
Jae-Chun Ban

Abstract

Error indices (bias, standard error of estimation, and root mean square error) obtained on different scales of measurement under different test termination rules in a CAT context were examined. Four ability estimation methods (MLE, WLE, EAP, and MAP), three measurement scales (θ, number correct score, and ACT score scale), and three test termination rules (fixed length, fixed standard error, and target information) were studied. The findings indicate that the amount and direction of bias, standard error of estimation, and root mean square error obtained under different ability estimation methods is influenced both by measurement scale and by test termination rule in a CAT environment. WLE performed the best among the four ability estimation methods on the ACT score scale with a target information termination rule.

Effects of Scale Transformation and Test Termination Rule on the Precision of Ability Estimates in CAT

Computerized adaptive testing (CAT) is designed to construct a unique test for each examinee, so that the test targets the examinee's estimated ability level. Theoretically, CAT has many advantages over paper-and-pencil (P&P) tests. One often mentioned advantage is its measurement efficiency or capability to deliver shorter tests. CAT also provides examinees with the benefits of test on demand and immediate test scoring and reporting. With the recent development in computer technology and psychometric knowledge, the popularity of CAT is increasing. Several high-stake testing programs have implemented CAT versions of P&P tests (Eignor, Way, Stocking, & Steffen, 1993; Sands, Waters, & McBride, 1997), and some others are moving towards using CAT as an alternative test-delivery method (Miller & Davey, 1999). Although CAT has many advantages, there are issues that need to be considered in the application of CAT. For example, the effects of scale transformations and test termination rules on the precision of ability estimation methods have not yet been fully investigated.

Most computerized adaptive tests (CATs) use item selection and scoring algorithms that depend on item response theory (IRT). However, it may be difficult for the general public with limited psychometric knowledge to understand the meaning of the θ scale. Thus, the measurement scales of CATs are often transformed from θ to more familiar scales, such as number correct (NC) score or reported score scales (Stocking, 1996). The test termination rule is another important factor that has to be considered when CATs are implemented. The choice of a stopping rule is affected by several factors, for example, the comparability between P&P tests and CATs, cost (e.g., cost of computer sitting time), measurement efficiency, and among others.

Because most CATs use IRT as the testing model, much of the previous research on CAT has been based on the IRT θ scale. For example, studies that compared different ability

estimation methods in terms of error indices, such as bias, standard error of estimation (SE), and root mean squared error (RMSE) of those methods, most often made the comparison on the θ scale (e.g., Bock & Mislevy, 1982; Crichton, 1981; Wang & Vispoel, 1998; Wang, Hanson, & Lau, 1999; Warm, 1989; Weiss & MacBride, 1984). This tendency is quite natural because much of the basic engine of CAT, such as the item selection algorithm, is based on IRT parameters that are directly related to the θ scale. However, when CATs move to actual implementation, the final reported score scale is rarely a linear transformation of the θ scale. For example, the GRE CAT uses the same score scale as the P&P version of the test, which is a nonlinear transformation of the estimated θ to a NC score and then to a reported score (e.g., Eignor & Schaeffer, 1995; Eignor et al., 1993). One question arising is whether the previous research results regarding the properties of different ability estimation methods or other CAT components (e.g., test termination rule) are scale specific.

A recent study indicated that the error indices obtained from the maximum likelihood estimation (MLE) method may be drastically different on the NC scale than on the θ scale (Yi, 1998). Lord (1980, Chapter 6) provided a theoretical derivation of the effects of scale transformation on the shape of information function. He indicated that "the units in terms of which information is measured depend on the units used to measure ability." (p. 87). In a general IRT setting, SE is directly related to information function, thus Lord's derivation provides an indication on how scale transformation can affect SE. However, it is not clear how scale transformation will affect the precision of different ability estimation methods in a CAT environment. Additionally, the effects of a nonlinear transformation of scale on bias have not been examined. Therefore, a major purpose of the current study was to systematically investigate the effects of nonlinear scale transformation on the properties of ability estimation methods in a CAT context.

A test termination rule decides when a CAT administration stops. Fixed length CAT administers the same length test to every examinee. Variable length CATs terminate a test according to certain stopping criteria. Wang et al. (1999) indicated that test termination rules (i.e., a variable vs. a fixed test length termination rule) might seriously affect the error indices of certain ability estimation methods in a CAT environment. In detail, Warm's (1989) weighted likelihood estimation (WLE) method was originally proposed to reduce the bias of MLE. However, Wang et al. found that, while under a fixed test length termination rule the WLE method was effective in reducing bias, under a fixed SE (i.e., standard deviation of estimates) termination rule it did not do so. Recently target information has been proposed as a test termination rule to achieve comparability between P&P tests and CATs (Miller & Davey, 1999). The target information can be obtained from a P&P test, and CATs terminate when the information obtained from administered items exceeds the target. However, the effects of this test termination rule on the precision of ability estimation methods are unknown. Thus, a second important purpose of this study was to investigate how termination rules can affect the performance of different ability estimation methods.

Ability Estimation Methods

The effects of scale transformations and test termination rules on the precision of ability estimation methods were evaluated on four methods. These four ability estimation methods were: MLE, WLE, expected a posterior (EAP), and maximum a posterior (MAP). A brief description of each method is presented in the following section.

Maximum Likelihood Estimation (MLE)

The MLE procedure is a widely used method for estimating an examinee's ability ($\hat{\theta}$) that maximizes the likelihood function for a particular response pattern. The likelihood function is denoted as

$$L(\mathbf{u} \mid \theta) = \prod_{i=1}^{n} P_i(u_i \mid \theta), \quad (1)$$

where $P_i(u_i \mid \theta)$ is the probability of getting response u_i on item i given an examinee's latent ability θ, and n is the number of items. $P_i(u_i \mid \theta)$ can be obtained from an item response function (e.g., three-parameter logistic IRT model).

Advantages of MLE are that they tend to be consistent and efficient as well as asymptotically normally distributed (Hambleton & Swaminathan, 1985). Due in part to these advantages, MLE is often used in a CAT context.

The asymptotic bias of MLE for a fixed test length in the three-parameter logistic (3-PL) IRT model is (Lord, 1983)

$$Bias(MLE(\theta)) = \frac{D}{I^2} \sum_{i=1}^{n} a_i I_i (\phi_i - \frac{1}{2}), \quad (2)$$

where $\phi_i \equiv \frac{P_i - c_i}{1 - c_i}$, $I_i = \frac{P_i'^2}{P_i Q_i}$, $P_i' = \frac{dP_i}{d\theta}$, $Q_i = 1 - P_i$, and $I = \sum_i I_i$ is the total test information. This bias function indicates that bias will be positive when an examinee's ability level is higher than the average item difficulty level; otherwise, bias will be negative. It also suggests that the bias will be close to zero when all items are targeted at an examinee's ability level, which means that under a CAT context bias should be minimal if an item pool contains a sufficiently large number of acceptable items at all ability levels.

Weighted Likelihood Estimation (WLE)

Warm (1989) proposed the WLE procedure for the 3-PL IRT model. The WLE method was designed to reduce the bias of the MLE method. The WLE estimate of θ, θ^*, is defined as the solution to

$$\sum_{i=1}^{n} \frac{(u_i - P_i)P_i'}{P_i Q_i} + \frac{d \ln w(\theta)}{d\theta} = 0, \tag{3}$$

where

$$\frac{d \ln w(\theta)}{d\theta} = - Bias(MLE(\theta)) * I. \tag{4}$$

Based on Lord's (1986) derivation, Warm (1989) showed that

$$Bias(\theta^*) = Bias(WLE(\theta)) = Bias(MLE(\theta)) + \frac{\frac{d \ln w(\theta)}{d\theta}}{I} = 0. \tag{5}$$

Thus, WLE is asymptotically unbiased for a fixed sample size. Warm conducted simulation studies to demonstrate that WLE was relatively unbiased for both conventional P&P tests and for CATs. However, the simulated CATs Warm conducted used a hypothetical item pool, for which the bias of MLE was relatively small and there was little room left for improvement.

Bayesian Ability Estimation Methods

The Bayesian ability estimation approaches incorporate the information on the examinees' ability distribution (i.e., prior distribution) into the ability estimation process. The prior ability distribution is combined with the likelihood function associated with a response pattern to create a posterior distribution. The EAP method finds the mean of the posterior distribution $p(\theta \mid \mathbf{u})$. Given a prior distribution $g(\theta)$, the posterior distribution can be expressed as

$$p(\theta \mid \mathbf{u}) = \frac{L(\mathbf{u} \mid \theta)\, g(\theta)}{\int L(\mathbf{u} \mid \theta)\, g(\theta)\, d\theta}, \tag{6}$$

and the mean of a posterior distribution can be defined as

$$E(\theta \mid \mathbf{u}) = \int_{-\infty}^{\infty} \theta\, p(\theta \mid \mathbf{u})\, d\theta. \tag{7}$$

The integration in the above two equations can be approximated using Gauss-Hermite quadrature (Stroud & Sechrest, 1966).

The MAP method uses the mode rather than the mean of the posterior distribution as the ability estimate. The MAP estimate is the solution of the following equation using an iterative numerical method such as Newton-Raphson procedure:

$$\frac{d \ln L(\mathbf{u} \mid \theta)}{d\theta} + \frac{d \ln g(\theta)}{d\theta} = 0 . \tag{8}$$

Lord (1986) indicated that the asymptotic bias for the MAP is related to the bias function of MLE:

$$Bias(MAP(\theta)) \approx Bias(MLE(\theta)) - \frac{\theta}{I} . \tag{9}$$

This equation indicates that a term linearly (negatively) related to θ is added to the bias of MLE, which means that the bias under the Bayesian methods is pulled toward the middle point of the θ scale.

Previous research on the ability estimation methods (e.g., Wang & Vispoel, 1998; Wang et al., 1999) has showed that the MLE method has lower bias than the Bayesian procedures, but has larger SE and RMSE in a CAT environment. The bias of MLE under a fixed length CAT with a realistic item pool is outward; that is, examinees' abilities are underestimated on the lower end of the θ scale and overestimated on the upper end of the θ scale. Under a fixed SE termination rule (i.e., fixed standard deviation of estimates), MLE was found to have inward bias; that is, examinees' abilities are overestimated on the lower end of the θ scale and underestimated on the upper end of the θ scale.

The bias of the Bayesian estimation methods is inward under both the fixed length and fixed SE termination rules (i.e., fixed posterior standard deviation) on the θ scale, but the

magnitude of the bias is different under different termination rules. Furthermore, Wang et al. (1999) found that WLE has smaller bias and lower SE than MLE under a fixed length CAT, but larger bias than MLE under a fixed SE rule. Wang and Wang (1999) conducted a similar study using a polytomous item pool under the generalized partial credit model and found a similar pattern of comparative precision of the ability estimation methods.

Method

This study had two purposes. One was to empirically investigate the effects of scale transformations on the precision of different ability estimation methods in a CAT environment. The second purpose was to examine the effects of test termination rules on different ability estimation procedures. Computer simulation methods were used in this study. A computer simulation program was developed using the C language. There were no practical constraints, such as item exposure rate control or content balancing, implemented in the simulated CATs to avoid confounding effects.

Item Parameter Calibration for the Item Pool

An item pool containing 420 items from seven forms of the ACT Mathematics Test (ACT, 1997) was used for the CAT simulations. The 3-PL IRT model was chosen as the model for item parameter calibration using the BILOG computer program (Mislevy & Bock, 1990). The calibrated *a*-, *b*-, and *c*-parameters have means of 0.97, 0.18, and 0.15, respectively. The standard deviations for these three parameters are 0.29, 0.97, and 0.05, respectively. The calibrated item parameters were treated as "truth" for item selection, item response generation, and ability estimation in the simulated CAT.

CAT Simulation

The CAT test started with an initial ability estimate of zero for all the simulees. The item selection algorithm was based on the maximum item information. Simulees' true abilities were

set at one of the 21 equally spaced points on the θ scale from -4 to 4 in increments of .4. At each of the θ points, the simulation was replicated 1000 times.

Simulees' CAT item responses were determined using the true item parameters and true ability. Item responses (0/1s) were determined by comparing the probability of a correct response (*P*-value; see Equation 11) based on the 3-PL IRT model with a uniform random number. If the *P*-value was less than the random number, the examinee received an incorrect response (i.e., 0), otherwise, a correct response (1). The four ability estimation methods (MLE, WLE, EAP, and MAP) that were discussed above were used to obtain examinees' estimated ability. The priors for EAP and MAP were standard normal distributions.

Scale Transformation and Test Termination Rule

Three measurement scales, the θ, the NC score, and the ACT scale score scales, were examined in this study. To transform a $\hat{\theta}$ to a NC score, the test characteristic curve (see Equation 10) of the base form (one of the seven test forms was designated as a base form) was used. The unrounded NC scores were transformed to the ACT scale scores using the conversion table for the base form by the linear interpolation method. ACT scores were then rounded to integer scores. The unrounded NC scores range from 0 to 60 (60 items are included in the P&P ACT Mathematics Test), and the rounded ACT scale scores range from 1 to 36. The test characteristic curve on the base form can be obtained using the following equation

$$TCC = \frac{1}{n}\sum_{i=1}^{n} P_i(\theta), \tag{10}$$

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \tag{11}$$

where i ($i = 1,2,3,\ldots,n$) represents items. The scale transformations from the θ to the NC scale, and from the NC to the ACT score scale are both nonlinear transformations.

Three types of test termination rules were studied, fixed test length, fixed SE, and target information. The test length was fixed at 30 items for the fixed length termination rule. Under the fixed SE rule, the standard error was set to 0.32 (i.e., error variance = .1) for MLE and WLE and the posterior standard deviation 0.32 (i.e., posterior variance of .1) was set for the Bayesian methods, and the maximum test length was set to 60 items for all the ability estimation methods. Target information was obtained from the P&P base form, and the simulated CATs terminated once test information exceeded the target information.

Evaluation Criteria

Three error indices, bias, SE, and RMSE, were examined to investigate the effects of scale transformation and test termination rule on the precision of different ability estimation methods. These error indices were calculated on the three measurement scales (θ, NC score, and ACT score scales) for each test termination rule, respectively. The error indices were computed based on the estimated and the true θ values. As an example, the bias, SE, and RMSE that are based on the θ scale can be expressed as

$$Bias(\hat{\theta} \mid \theta) = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta), \tag{12}$$

$$SE(\hat{\theta} \mid \theta) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \frac{\sum_{k=1}^{N}\hat{\theta}_k}{N}\right)^2}, \tag{13}$$

$$RMSE(\hat{\theta} \mid \theta) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta)^2}, \tag{14}$$

$$RMSE^2 = Bias^2 + SE^2, \tag{15}$$

where θ represents simulee's true ability, and $\hat{\theta}_j$ is the estimated ability for simulee $j = 1, 2, 3, \ldots, N$. The error indices on the NC and the ACT scales were computed similarly, except that the θ and $\hat{\theta}_j$ were transformed onto the corresponding NC or ACT scales.

Results

Figure 1 presents the bias obtained under the three measurement scales and the three test termination rules across the four ability estimation methods. In general, across the test termination rules, the bias resulting from the EAP and MAP methods were larger than the bias from MLE and WLE, especially at the two ends of the measurement scales. There was positive bias at the lower end and negative bias at the higher end of the scale. EAP had smaller bias than that of MAP. The transformations of the measurement scale and the test termination rules did affect the shapes of the bias curves. However, the general patterns of the curves were not influenced.

The use of different termination rules greatly affected the bias of MLE and WLE. In some cases, even the direction of the bias was influenced by test termination rules. Specifically, with the fixed SE termination rule, MLE had smaller bias than WLE. Both methods had positive bias at the lower end and negative bias at the higher end of the scale (with a minor exception on the θ scale). With the target information termination rule, MLE had much larger bias than WLE. MLE had large negative bias at the lower end and positive bias at the higher end of the scale. WLE, on the other hand, had minimal bias along the scale except at the two extreme ends of the scale. With the fixed length rule, both MLE and WLE had relatively small bias. Interestingly, the measurement scale transformation affected the bias direction for MLE under the fixed length rule. The change in the direction of MLE bias might be due to the nonlinear nature of the scale transformation. See Appendix for further explanation.

Insert Figure 1 About Here

The effects of different termination rules on bias were further investigated by examining the actual test length obtained under these rules. The means and standard deviations of the number of items administered to examinees conditional on θ for the fixed SE and the target information termination rules are presented in Tables 1 and 2. For the fixed SE termination rule, there were more items administered to simulees at the two ends of the scale than at the middle of the scale. The mean of the number of items ranged from about 9 (at $\theta = 0.4$ for EAP and MAP) to 60 (at $\theta = 4.0$ for MLE and WLE). The mean of the number of items for the target information termination rule, on the other hand, was large at the middle of the scale and small at the two ends of the scale. The mean number of items administered ranged from about 7 ($\theta = -4.0$ for MLE) to about 24 ($\theta = 1.2$ for MLE and WLE).

Insert Tables 1 and 2 About Here

Figure 2 displays the SE obtained on the three measurement scales and the three termination rules across the four ability estimation methods. Generally, EAP and MAP had the smallest SE across all conditions. EAP and MAP had a similar pattern of SE. When there was a difference, MAP had slightly smaller SE than EAP. MLE had the largest SE among the four ability estimation methods across all conditions. The transformation of the measurement scale from θ to NC score scale changed the shape of SE from concave up to concave down.

Brennan and Kolen (1989) stated that one of the goals in setting the ACT score scale for the P&P tests was to equalize SE along the entire measurement scale. In the near future, P&P tests and CATs may coexist for some period of time, thus, it is important to obtain comparable

test results from these two different test modes. Based on the findings of this study, the goal of equalizing SE along the measurement scale was maintained to some extent for the fixed length CAT, and more so for the CAT under the target information rule. In addition, the information curve that was used in this study was based on a P&P base form of ACT Mathematics Test. Thus, it is suggested that using target information as test termination criterion might help to achieve comparability between the P&P and CAT version of the test. For the fixed SE rule, SE was not equalized along the ACT scale. This is because the fixed SE rule was set to equalize SE on the θ scale. Under the target information rule, MLE had consistently larger SE than that of WLE and the Bayesian methods.

Insert Figure 2 About Here

Figure 3 presents the RMSE resulting from the conditions studied in this research. The patterns of RMSE were similar to those of SE except at the two extreme ends of the scales. As indicated in Equation 15, RMSE is composed of bias and SE. Relative to bias, SE contributed more to RMSE in the middle of the scale. The change of patterns at the extreme ends of the scale was due to the high bias of the Bayesian methods.

Insert Figure 3 About Here

Overall, under a fixed length and a target information termination rule across all three measurement scales, WLE performed better than MLE on all three error indices. When a fixed SE rule was used, WLE functioned slightly worse than MLE on bias, but slightly better on SE. The Bayesian methods always performed better than MLE and WLE on SE but worse on bias.

Discussion

The purposes of this study were to investigate the effects of the measurement scale transformations and the test termination rules on different ability estimation methods in a CAT environment. The results of the study indicated that the measurement scale transformations and the test termination rules did not severely affect the general patterns of the bias obtained under the Bayesian methods, that is, EAP and MAP resulted in negative bias at the lower end and positive bias at the higher end of the scale, and the magnitude of the bias was larger than those of MLE and WLE. Additionally, EAP and MAP had smaller SE than MLE and WLE across the conditions examined in this study.

The scale transformations and the test termination rules, on the other hand, influenced MLE and WLE. WLE resulted in less bias and SE than MLE along most of the scale points under the fixed length and the target information termination rules. The main reason Warm (1989) proposed to use WLE was to reduce bias in ability estimation. WLE achieved this goal at most of the conditions investigated in this study. This study also revealed that the transformation of the measurement scale from θ to NC or to ACT scale even changed the bias direction of MLE. This is due to the nonlinear transformation of the scales.

The findings of this study suggested that when the target information was implemented as termination rule WLE performed better than MLE on the ACT score scale. For some testing programs, the comparability between P&P and CAT versions of the test is important due to the fact that P&P tests and CATs may coexist for some time. As indicated above, the test information target used in this study was obtained from a P&P test form, and the results of this study suggested that the target information rule assisted to achieve comparability when combined with WLE. In addition, the SE was equalized along most parts of the scale (except the two extreme ends of the scale).

Variable length tests resulted from the fixed SE and the target information termination rules. However, the patterns of the mean number of items administered were different under these two conditions. There were more items given to examinees at the two ends of the scale under the fixed SE termination rule. For the target information rule, however, more items were given at the middle of the scale. The range of the number of items administered was different under these two conditions. There was as many as 60 items given to examinees under the fixed SE condition. This might due to the fact that there were not many very difficult or easy items in the item pool that was used in this study.

The results on the precision of different ability estimation methods for the fixed length and fixed SE tests on the θ scale are consistent with the findings of the previous research (Wang et al. 1999; Wang & Wang, 1999; Yi, 1998). Furthermore, this study indicated that when a target information rule was used on the ACT score scale, WLE was the best ability estimation method among the four procedures studied. Additionally, fixed SE might not result in an efficient test, especially when the item pool does not include a wide range of items. A fixed length test, on the other hand, might result in relatively small bias and SE, even after the measurement scale was transformed.

This research was conducted under unconstrained CAT conditions. For future study, a more realistic CAT administration needs to be studied. That is, item exposure rate control and content balancing need to be implemented in a CAT administration. In addition, the properties of the WLE method should be further explored.

References

ACT (1997). *ACT Assessment Technical Manual.* Iowa City, IA: Author.

Brennan, R. & Kolen, M. (1989). Scaling the ACT Assessment and P-ACT+: Rational and goals. In R. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 1-17). Iowa City, IA: Author.

Bock, R. & Mislevy, R. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Crichton, L. (1981). *Effect of error in item parameter estimates on adaptive testing.* Ann Arbor, Michigan: UMI Dissertation Information Service.

Eignor, D. & Schaeffer, G. (1995, April). *Comparability studies for the GRE CAT and the NCLEX using CAT.* Paper presented at the annual meeting of National Council of Measurement in Education. San Francisco, CA.

Eignor, D., Way, W., Stocking, M., & Steffen, M. (1993). *Case studies in computerized adaptive test design through simulation.* Research Report (RR-93-56). Princeton, NJ: Educational Testing Service.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston, MA: Kluwer-Nijhoff.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48(2),* 233-245.

Lord, F. (1986). Maximum likelihood and Bayesian parameter estimation in tem response theory. Journal of Educational Measurement, 23, 157-162.

Miller, T. & Davey, T. (1999, April). *Principles for administering adaptive tests.* Paper presented at the annual meeting of National Council of Measurement in Education, Montreal, Canada.

Mislevy, R. & Bock, R. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. [Computer program]. Chicago, IN: Scientific Software.

Sands, W., Water, B., & McBride, J. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation.* Washington DC: American Psychological Association.

Stocking, M. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21(4),* 365-389.

Stroud, A. & Sechrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice-Hall.

Wang, S. & Wang, T. (1999, April). *Precision of Warm's weighted likelihood estimation of ability for a polytomous model in CAT.* Paper presented at the annual meeting of American educational Research Association, Montreal, Canada.

Wang, T., Hanson, B., & Lau, C. (1999). Reducing bias in CAT ability estimation: A comparison of approaches. *Applied Psychological Measurement, 23,* 263-278.

Wang, T. & Vispoel, W. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35(2),* 109-135.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54(3),* 427-450.

Weiss, D. & McBride, J. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.

Yi, Q. (1998). *A comparison of three ability estimation procedures for computerized adaptive testing in the presence of nonmodel-fitting responses resulting from a compromised item pool.* Unpublished doctoral dissertation. Tampa, FL: University of South Florida.

Appendix

Nonlinear Scale Transformation Changes the Direction of Bias

It is presumed that a monotonic scale transformation would not change the direction of bias of a CAT ability estimate (i.e., changing from positive bias to negative bias, or vice versa). However, the simulation results from this study indicated that the bias direction was changed when the measurement scale is nonlinearly transformed. A possible explanation for this finding is provided in the following section. The bias on the θ scale and on the transformed scale S for a given sample of size N is

$$Bias(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta) \tag{16}$$

and

$$Bias[S(\hat{\theta})] = \frac{1}{N}\sum_{i=1}^{N}[S(\hat{\theta}_i) - S(\theta)] \tag{17}$$

where θ represents an examinee's true ability. The bias on the θ scale would be positive if the absolute value of the sum of positive terms in the summation is greater than the absolute value of the sum of the negative terms. Because the transformations are monotonic, the positive terms in Equation 16 would correspond to the positive terms in Equation 17, and vice versa. However, the nonlinear transformation function may make the magnitude of the positive terms smaller relative to the negative terms and make the absolute value of their sum smaller than that of the negative terms, and furthermore may result in negative bias. Figure 4 illustrates an extreme case of the two terms. The solid lines represent the true ability and the dashed lines are the two estimates. This graph presents how a larger positive deviation on one scale is transformed into a smaller positive term on the other scale.

Insert Figure 4 About Here

TABLE 1

Mean and Standard Deviation of Number of Items Administered under Fixed SE Termination Rule Across Four Ability Estimation Methods

Theta	MLE Mean(SD)	WLE Mean(SD)	EAP Mean(SD)	MAP Mean(SD)
		Fixed SE		
-4.0	59.6 (4.4)	59.7 (3.9)	59.6 (3.6)	59.4 (4.5)
-3.6	59.8 (3.3)	59.6 (4.4)	59.3 (4.6)	59.0 (5.7)
-3.2	59.7 (3.4)	59.5 (4.5)	58.6 (5.8)	58.6 (6.0)
-2.8	58.8 (6.0)	58.6 (6.4)	55.6 (9.7)	55.5 (9.8)
-2.4	56.8 (7.9)	56.2 (9.0)	48.4 (12.8)	48.4 (12.8)
-2.0	47.3 (12.4)	46.2 (12.9)	35.6 (12.5)	36.1 (12.7)
-1.6	32.4 (11.0)	31.4 (10.5)	24.9 (8.2)	25.2 (8.7)
-1.2	22.3 (6.3)	21.8 (6.2)	17.8 (5.5)	18.1 (5.5)
-0.8	16.0 (4.0)	15.3 (4.1)	13.7 (3.3)	13.9 (3.4)
-0.4	12.6 (2.7)	11.8 (2.5)	11.0 (2.2)	11.1 (2.3)
0.0	10.9 (1.7)	10.3 (1.5)	9.6 (1.5)	9.5 (1.5)
0.4	11.0 (1.5)	10.6 (1.3)	9.2 (1.1)	9.2 (1.2)
0.8	11.5 (1.5)	11.5 (1.2)	9.7 (1.1)	9.7 (1.1)
1.2	12.4 (1.5)	12.5 (1.2)	10.6 (1.3)	10.6 (1.3)
1.6	13.7 (1.7)	13.5 (1.6)	12.3 (1.8)	12.2 (1.8)
2.0	16.5 (4.0)	15.5 (3.2)	14.6 (2.9)	14.7 (2.9)
2.4	25.3 (10.5)	22.3 (8.5)	19.7 (6.5)	19.8 (7.0)
2.8	42.0 (14.9)	38.6 (15.2)	29.3 (12.8)	29.4 (12.9)
3.2	56.2 (8.9)	53.5 (11.5)	43.9 (15.5)	44.5 (15.3)
3.6	59.3 (3.9)	58.5 (5.9)	54.0 (11.2)	53.3 (11.8)
4.0	60.0 (0.6)	59.9 (1.6)	58.6 (5.5)	58.2 (6.7)

TABLE 2

Mean and Standard Deviation of Number of Items Administered under Target Information Termination Rule Across Four Ability Estimation Methods

Theta	MLE Mean(SD)	WLE Mean(SD)	EAP Mean(SD)	MAP Mean(SD)
		Target Information		
-4.0	6.6 (4.7)	9.5 (5.3)	7.6 (2.7)	9.3 (3.6)
-3.6	7.0 (4.5)	9.8 (5.3)	7.6 (2.8)	9.2 (3.7)
-3.2	7.7 (4.8)	10.3 (5.3)	7.9 (3.1)	9.4 (3.8)
-2.8	8.3 (5.0)	11.5 (5.4)	7.9 (2.9)	9.7 (3.9)
-2.4	9.4 (4.9)	12.1 (5.3)	8.6 (3.1)	10.0 (3.8)
-2.0	10.7 (5.1)	13.2 (4.8)	9.1 (3.3)	10.8 (4.0)
-1.6	12.8 (4.8)	14.7 (4.7)	10.6 (3.6)	12.5 (4.1)
-1.2	14.8 (4.8)	16.5 (4.0)	12.7 (3.9)	14.8 (4.1)
-0.8	17.7 (4.1)	18.9 (3.2)	16.0 (3.4)	17.4 (3.4)
-0.4	20.3 (3.5)	21.1 (2.3)	19.0 (2.5)	19.6 (2.4)
0.0	21.9 (3.4)	22.6 (2.2)	20.6 (2.3)	20.7 (2.1)
0.4	23.0 (2.3)	23.3 (1.5)	21.7 (2.4)	21.3 (2.2)
0.8	23.6 (1.8)	23.5 (1.9)	22.4 (2.2)	22.3 (2.2)
1.2	23.7 (2.0)	23.7 (2.2)	23.0 (1.9)	23.1 (2.0)
1.6	22.4 (2.6)	22.1 (3.1)	22.5 (1.8)	22.9 (1.8)
2.0	18.9 (3.6)	19.1 (3.1)	20.3 (2.1)	21.1 (1.8)
2.4	16.1 (3.8)	17.4 (3.1)	18.1 (1.9)	19.5 (1.6)
2.8	14.0 (4.2)	15.9 (2.8)	16.7 (1.5)	18.8 (1.3)
3.2	12.3 (4.1)	15.0 (2.8)	15.9 (1.3)	18.3 (1.0)
3.6	10.5 (3.8)	14.3 (2.6)	15.5 (1.0)	18.0 (0.8)
4.0	9.5 (3.4)	13.6 (2.4)	15.2 (0.7)	17.8 (0.7)

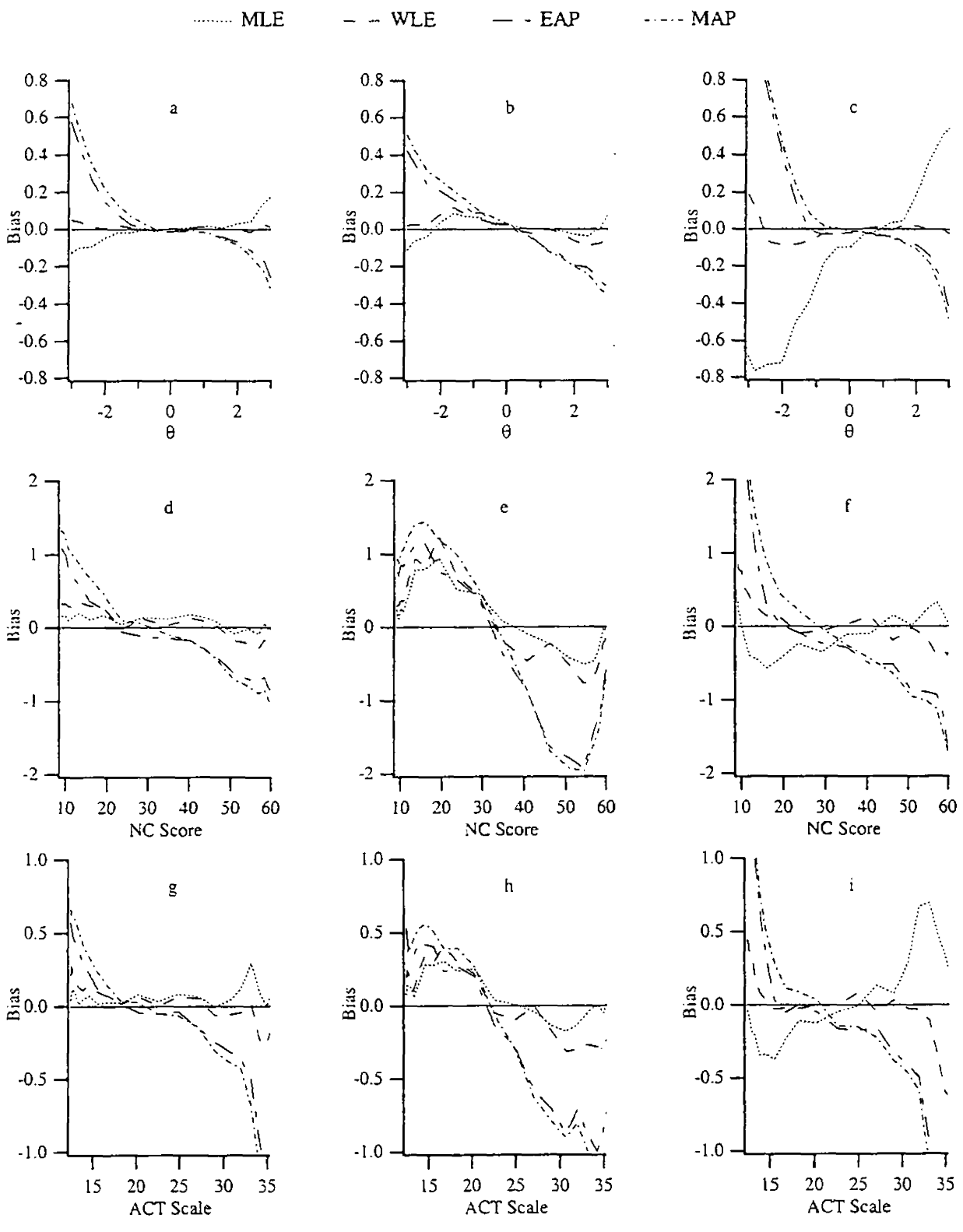

FIGURE 1. The effects of measurement scale transformations and test termination rules on bias estimation across four ability estimation methods

Note: a. θ scale, fixed length; b. θ scale, fixed SE; c. θ scale, target information
d. NC score scale, fixed length; e. NC score scale, fixed SE; f. NC score scale, target information
g. ACT scale, fixed length; h. ACT scale, fixed SE; i. ACT scale, target information

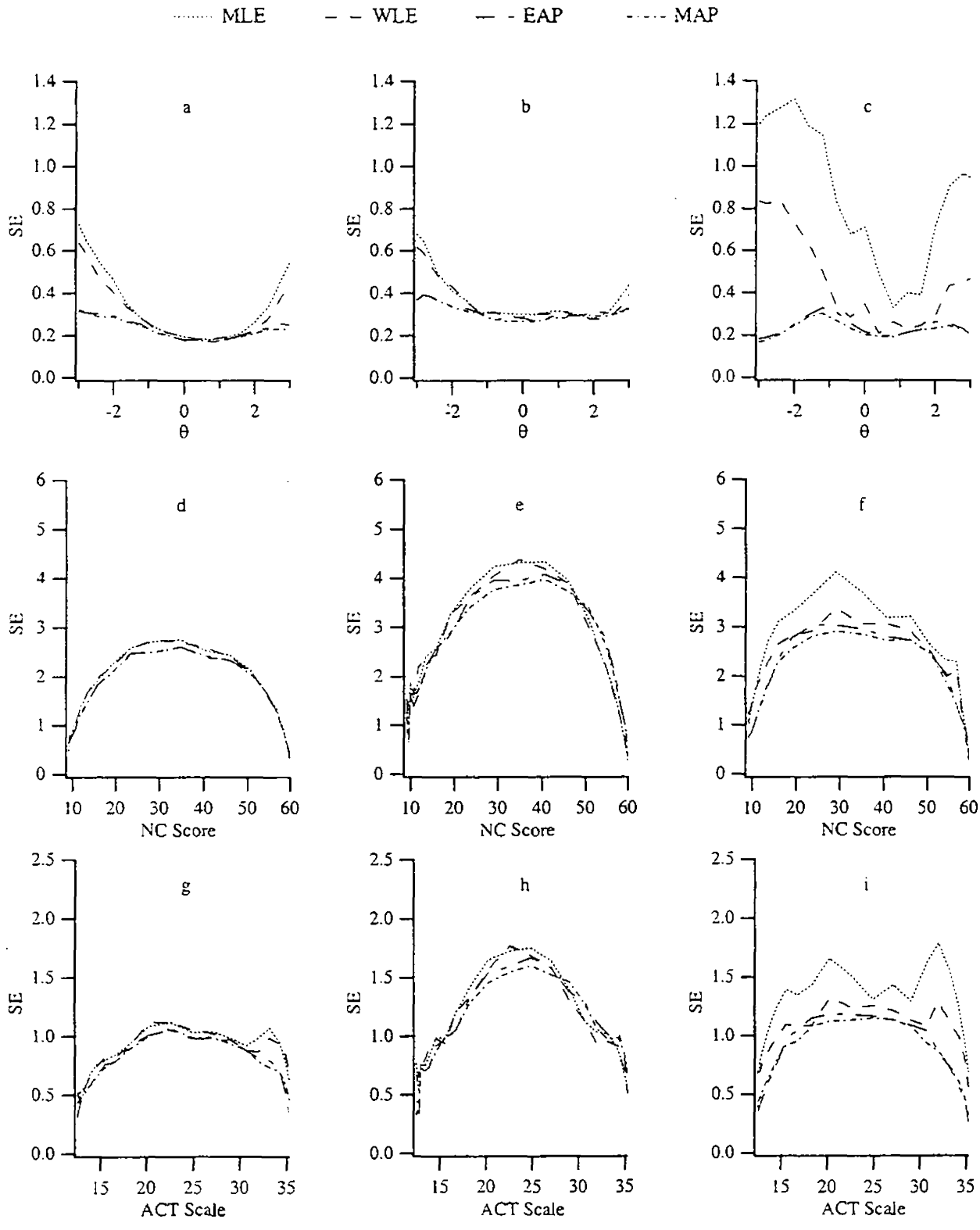

FIGURE 2. The effects of measurement scale transformations and test termination rules on standard error of estimation across four ability estimation methods

Note: a. θ scale, fixed length
b. θ scale, fixed SE
c. θ scale, target information
d. NC score scale, fixed length
e. NC score scale, fixed SE
f. NC score scale, target information
g. ACT scale, fixed length
h. ACT scale, fixed SE
i. ACT scale, target information

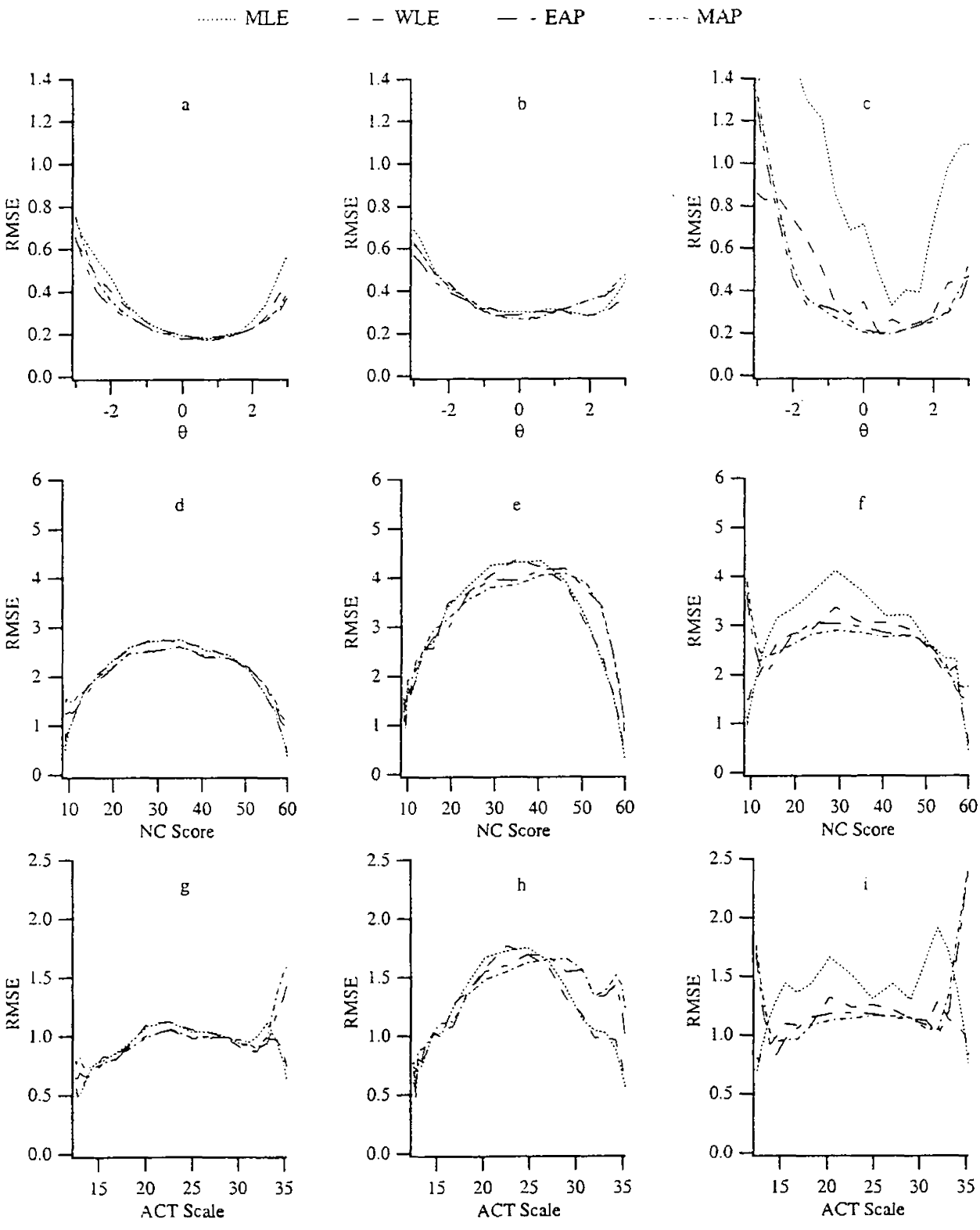

FIGURE 3. The effects of measurement scale transformations and test termination rules on root mean square error across four ability estimation methods

Note: a. θ scale, fixed length
b. θ scale, fixed SE
c. θ scale, target information
d. NC score scale, fixed length
e. NC score scale, fixed SE
f. NC score scale, target information
g. ACT scale, fixed length
h. ACT scale, fixed SE
i. ACT scale, target information

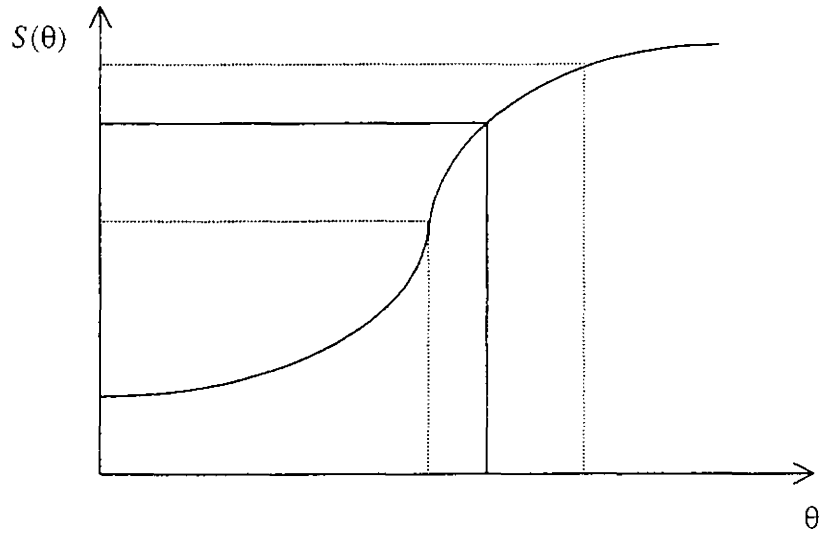

FIGURE 4. The effect of a nonlinear transformation of scale on the direction of bias