Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 24

Defining a Score Scale in Relation to Measurement Error for Mixed Format Tests[*]

Jae-Chun Ban
Won-Chan Lee[†]

February 2007

[*] The authors are grateful to Robert Brennan for his valuable comments on the paper.

[†] Jae-Chun Ban is Assistant Professor, Chungnam National University, Korea (email: jban@cnu.ac.kr). Won-Chan Lee is Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

The purpose of this study was to describe the arcsine scaling procedure for equalizing measurement error for mixed format tests and to evaluate performance of the procedure in terms of constant conditional standard errors of measurement (CSEMs). The estimated observed scores using a mixture of the three-parameter logistic and generalized partial credit IRT models were transformed to establish a reported scale score that was intended to possess the property of constant CSEMs. Four real mixed format tests were used to demonstrate the applicability of the arcsine scaling procedure to achieve constant CSEMs. The results showed that the arcsine scaling procedure performed well for mixed format tests resulting in approximately equalized CSEMs for base forms as well as equated forms. When compared to some other types of scale scores, the arcsine-transformed scale scores had the most stable CSEMs.

1 Introduction

Incorporating normative and score precision information into a score scale at the time it is established can facilitate the interpretability of test scores (Kolen, 2006; Petersen, Kolen, & Hoover, 1989). In particular, when a score scale is developed, it would be useful to know the measurement error characteristics of the score scale. This paper focuses primarily on a scaling procedure that is intended to produce a score scale that has approximately equal measurement error variance along the score scale.

It is well known that the conditional standard errors of measurement (CSEMs) for number-correct scores vary along the score scale, typically showing an inverted U-shape (Lord, 1984; Brennan, 1998; Brennan & Lee, 1999). When the number-correct scores are nonlinearly transformed, the resulting scale scores will have CSEMs with a pattern that is different from that of the number-correct score CSEMs (Kolen, Hanson, & Brennan, 1992; Brennan & Lee, 1999; Lee, Brennan, & Kolen, 2000). The pattern of CSEMs for nonlinearly transformed scale scores depends on the degree of nonlinearity (Lee et al., 2000).

To enhance interpretability of scale scores, Kolen (1988) suggested a nonlinear transformation of number-correct scores to obtain nearly constant CSEMs for the scale scores. Kolen (1988) employed an arcsine transformation of number-correct scores (Freeman & Tukey, 1950). These arcsine-transformed scores or any further linear transformation of them will have the property of constant CSEMs (Kolen, 1988; Kolen et al., 1992). The CSEMs for nonlinearly transformed scale scores (e.g., arcsine-transformed scores) can be computed using strong true score models (Lord, 1965, 1969; Kolen et al., 1992) or item response theory (IRT) models (Kolen, Zeng, & Hanson, 1996; Lee et al., 2000).

The arcsine scaling procedure and previous research studies are limited to tests consisting of dichotomously-scored items. Recently, there have been an increasing number of tests that are composed of items with mixed formats. For example, a mixed format test could consist of mixtures of multiple-choice (i.e., dichotomous) and constructed response (i.e., polytomous) items. It is unknown, however, whether the arcsine scaling procedure can properly be applied to the total raw scores obtained from mixed item types. Kolen (2006, p. 167) states that "Although no published studies exist, it ... could be used to stabilize error variance for summed scores other than those that are number-correct." In an attempt to fill the gap in the literature, this study demonstrates the applicability of the arcsine scaling methodology for achieving equal measurement error variance using mixed format tests for which the total raw score is a sum of scores for all dichotomous and polytomous items.

The arcsine scaling procedure has always been associated with strong true score models. This study employs IRT as a framework for the arcsine scaling procedure. Using mixed IRT models (i.e., a dichotomous model and a polytomous model), this paper describes the entire process of scaling, linking, and equating using real data sets from mixed format tests. More specifically:

- Scaling involves the development of the scale scores for mixed format tests using the arcsine scaling procedure;

- linking (or scale transformation) is conducted to place the item parameter estimates for the two forms to be equated on the same IRT proficiency scale; and
- equating is carried out using the IRT true score equating method under the common-item nonequivalent groups design.

The arcsine-transformed scale scores developed for the mixed format tests are evaluated in terms of the equalized scale-score CSEMs for the base forms as well as the equated forms. Finally, the performance of the arcsine scaling procedure is compared to several other types of scale scores.

2 Method

The arcsine scaling procedure that equalizes the CSEMs for scale scores using mixed format data is described in this section. The IRT procedures for computing the CSEMs for raw and scale scores and conducting linking and equating for mixed format tests are also provided.

2.1 Data Source

This study used the Korean, Mathematics, Science, and English tests from the Korean National Assessment of Educational Achievement (NAEA–K). Two forms (A and B) for each test were developed, pretested, and operationally administered. The Korean and Mathematics tests were administered to about 2,700 Grade 10 students per form in 2003. The Science and English tests were administered to about 8,200 Grade 10 students per form in 2004. Schools were sampled using a two-stage stratified cluster sampling method from a nationwide high school list. Forms A and B were randomly assigned to the schools sampled. One class from each school took either Form A or Form B. Since the two forms were not spiraled within a school, the two groups of students each of which took either one of the two forms were not equivalent.

The number of items, possible raw-score points, and common items for the tests are summarized in Table 1. The total number of items across the tests ranges from 40 to 43. The number of total possible points was 46 for Science and 55 for the other three tests. Forms A and B for each test had a set of items in common. The percent of the common items in a form was a minimum of 40% (for Science) and a maximum of 57% (for Mathematics). A common-item set between the two forms for each test consisted of all polytomous items and some dichotomous items.

Table 2 shows the descriptive statistics and reliabilities for the raw scores of the four tests. The maximum mean difference between Forms A and B was 1.6 raw score points for Mathematics. The standard deviation differences were less than 1.0 across the tests. The reliabilities tended to be similar between the two

Table 1: Number of Items, Possible Points, and Common Items

			Dich Items		Poly Items		Total	
Test	Form	CI	Items	Pts	Items	Pts	Items	Pts
Korean	A, B	23	30	30	13	25	43	55
Math	A, B	24	30	30	12	25	42	55
Science	A, B	16	32	32	8	14	40	46
English	A, B	19	32	32	10	23	42	55

Note: CI = common items; Pts = points.

Table 2: Descriptive Statistics for Raw Scores

Test	Form	Mean	SD	Min.	Max.	Reliability
Korean	A	27.25	8.48	4	48	0.85
	B	28.48	9.04	4	50	0.85
Math	A	23.40	11.93	1	55	0.93
	B	25.02	12.89	2	55	0.94
Science	A	18.85	9.49	0	45	0.90
	B	19.81	9.29	0	45	0.91
English	A	23.10	12.32	2	55	0.94
	B	22.49	12.09	0	55	0.93

forms within a test, but somewhat different across the tests. The Korean test had the lowest reliability of .85. The reliability values seemed to be acceptable for trend analyses of students' achievement at the national level.

2.2 Calibration, Linking, and Equating

The item parameters and ability distribution were estimated using the computer software PARSCALE (Muraki & Bock, 2003). The dichotomous and polytomous items were calibrated simultaneously. The three-parameter logistic (3PL; Lord, 1980) model was fitted to the dichotomous items, and Muraki's generalized partial credit (GPC; Muraki, 1992) model to the polytomous items. Under the 3PL model, the probability that person i with ability θ answers item k correctly is

$$\Pr(U_k = 1|\theta) = c_k + (1 - c_k)\frac{\exp[Da_k(\theta_i - b_k)]}{1 + \exp[Da_k(\theta_i - b_k)]}, \tag{1}$$

where D is 1.7; a_k is a discrimination parameter; b_k is a difficulty parameter; and c_k is a lower asymptote parameter; and U_k is a random variable to represent item scores. With the GPC model, the probability of a particular response is given by

$$\Pr(U_k = u_k|\theta) = \frac{\exp\left[\sum\limits_{v=0}^{u_k} a_k(\theta - b_k + d_{kv})\right]}{\sum\limits_{c=0}^{m} \exp\left[\sum\limits_{v=0}^{c} a_k(\theta - b_k + d_{kv})\right]}, \tag{2}$$

where b_k is a location parameter and d_{kv}'s ($v = 0, 1, \ldots, m$) are category parameters.

Occasionally there were a few items with a very low frequency for a particular category. In such a case, the category was collapsed with an adjacent category after communicating with relevant content experts. All calibration runs converged to the default convergence criterion.

Since the two groups of examinees for Forms A and B were not equivalent and item parameters on the two forms were estimated separately, the item parameter estimates for the two forms were not on the same IRT θ scale. To place the item parameter estimates for the two forms on the same scale, the Stocking–Lord transformation method extended to mixed format tests, as proposed by Kim and Lee (2006), was employed. In so doing, Form A was considered as the base form. The extended Stocking–Lord method defines a symmetric criterion function to be minimized. Roughly speaking, the symmetric criterion suggested by Kim and Lee (2006) is defined as the sum of two (i.e., old to new and new to old) squared differences between two test characteristic curves for the two groups based on the common items. The test characteristic curves in this study were obtained by summing weighted probabilities of item and category response functions, in which the weight is a scoring function. The computer software STUIRT (Kim & Kolen, 2004) was used to carry out all the scale transformations.

Once the item parameter estimates and ability distributions for the two forms and two groups were placed on the same scale, IRT true score equatings were conducted using the computer software POLYEQUATE (Kolen, 2003). The raw-to-scale score conversion used for equating was constructed as a result of the arcsine scaling procedure, which is described in a later section.

2.3 Total Raw Scores

An extension of the Lord–Wingersky recursive algorithm (Lord & Wingersky, 1984) to polytomous items was provided by Hanson (1994) and Wang, Kolen, and Harris (2000) to estimate the conditional distribution of total raw scores. The extended recursive formula was employed in this paper to estimate the conditional distribution of total raw scores for the mixed format tests, where the conditioning is on θ.

Assume there are K items that are a mixture of dichotomous and polytomous items. Let U_k be a random variable for the score on item k, where $U_k = 0, 1, \ldots, n_k$. Let $\Pr(X = x|\theta)$ ($x = 0, 1, \ldots, T$) represent the conditional distribution of the total raw score. Let $Y_k = \sum_{j=1}^{k} U_j$ and $X = Y_K$. For item $k = 1$,

$$\Pr(Y_1 = x|\theta) = \Pr(U_1 = x|\theta), \text{ for } x = 0, 1, \ldots, n_1.$$

For item $k > 1$,

$$\Pr(Y_1 = x|\theta) = \sum_{u_k=0}^{n_k} \Pr(Y_{k-1} = x - u_k|\theta)\Pr(U_k = u_k|\theta), \text{ for } x = 0, 1, \ldots, \sum_{j=1}^{k} n_j.$$

Note that the probability of getting an invalid score for Y_{k-1} is set to be zero. To use this recursive formula, begin with $k = 1$ and repeatedly apply the formula by increasing k on each repetition. As described in Kolen, Zeng, and Hanson (1996), variance of the conditional distribution is the conditional error variance of the total raw scores at θ, $\sigma^2(X|\theta)$. The square root of this conditional error variance is the CSEM for total raw scores at θ.

The average error variance, $\sigma^2(E_r)$, and the marginal distribution of the total raw scores, $\Pr(X = x)$, over the θ distribution are computed as follows:

$$\sigma^2(E_r) = \int_\theta \sigma^2(X|\theta)\psi(\theta)d\theta, \tag{3}$$

and

$$\Pr(X = x) = \int_\theta \Pr(X = x|\theta)\psi(\theta)d\theta, \tag{4}$$

where $\psi(\theta)$ is the density function for θ. In this paper, Equations 3 and 4 involved using the item parameter estimates and posterior distribution of θ obtained from PARSCALE, and replacing the integrals by summations. This computational procedure, in general, was applied to all subsequent equations, as needed.

2.4 Arcsine Transformation

For a test that consists of dichotomous items, Kolen (1988) employed the arcsine transformation method to stabilize the CSEMs for scale scores. The arcsine transformation in this study was applied to the mixed format tests. The total raw scores of Form A (i.e., the base form) for each test were transformed as follows:

$$c(x) = \frac{1}{2}\left[\sin^{-1}\sqrt{\frac{x}{T+1}} + \sin^{-1}\sqrt{\frac{x+1}{T+1}}\right], \tag{5}$$

where $\sin^{-1}$ is the arcsine function and T is the maximum total score for the mixed format test. Note that the integer raw scores, X, in Equation 5 are regarded as the estimated total raw scores based on the IRT models (as defined earlier in this paper). The arcsine transformation compresses the raw-score scale in the middle and stretches it at the extremes.

The CSEMs and average standard error of measurement for the arcsine-transformed scores were then computed. Let $c(x)$ be the raw-to-arcsine transformation function. The population mean and variance for the arcsine-transformed scores are defined as:

$$\mu[c(X)] = \sum_{x=0}^{T} c(x)\Pr(X = x), \tag{6}$$

and

$$\sigma^2[c(X)] = \sum_{x=0}^{T} \{c(x) - \mu[c(X)]\}^2 \Pr(X = x), \tag{7}$$

where $\Pr(X = x)$ is the marginal distribution of the total raw scores. The conditional mean and variance for the arcsine transformed scores are given by

$$\mu[c(X)|\theta] = \sum_{x=0}^{T} c(x) \Pr(X = x|\theta), \tag{8}$$

and

$$\sigma^2[c(X)|\theta] = \sum_{x=0}^{T} \{c(x) - \mu[c(X)|\theta]\}^2 \Pr(X = x|\theta), \tag{9}$$

where $\Pr(X = x|\theta)$ is the conditional total-raw score distribution. The square root of Equation 9 is the CSEM for the arcsine transformed scores at θ. The average error variance for the arcsine transformed scores over the theta distribution, $\sigma^2(E_c)$, is

$$\sigma^2(E_c) = \int_\theta \sigma^2[(c(X)|\theta)]\psi(\theta)d\theta. \tag{10}$$

2.5 Constructing Scale Scores

The criteria used for constructing a score scale for each test were: (a) the mean should be 20; (b) possible scale scores should cover the entire score range of 8σ (-4σ to $+4\sigma$); and (c) the standard error of measurement should be close to 2 scale-score points. As a result of criterion (c), an approximate 68% confidence interval could be formed by adding ±2 points to examinees' scale scores.

The raw score reliabilities for the four tests used in this study were greater than .85 as shown in Table 2. Since the desired standard error of measurement was 2 and the 68% confidence interval was considered, the standard deviation of the score scale suggested by some rules of thumb (Kolen & Brennan, 2004) was $\sigma(S) = 2/(1\sqrt{1 - .85}) = 5.16$. The entire score range of interest was $8\sigma(S)$. Therefore, the suggested number of score points was 41.28 ($= 8 \times 5.16$) and 41 when rounded. Thus, the final score scale was set to range from 0 to 40 with an increment of 1.

Once the characteristics of the score scale were determined, the arcsine-transformed scores were, then, linearly transformed such that the desired mean and standard error of measurement for the scale scores could be achieved as follows:

$$s[c(x)] = \frac{\sigma(E_s)}{\sigma(E_c)}\{c(x) - \mu[c(X)]\} + \mu(S), \tag{11}$$

where $\mu(S)$ and $\sigma(E_s)$ are the desired mean and standard error of measurement for the scale scores, and $\mu[c(X)]$ and $\sigma(E_c)$ are the mean and standard error of measurement for the arcsine transformed scores. The linear transformation specified in Equation 11 produced unrounded scale scores with a few scores that were out of range. The unrounded scale scores were rounded and those scores out of range were truncated so that the score scale ranged from 0 to 40 integer points. Table 3 shows the descriptive statistics for the rounded scale scores for the four tests. Note that the means of the scale scores were very close to the target mean of 20 for all four tests, while the standard deviations were somewhat different across the four tests due to the differences in reliabilities. Finally, the raw-to-scale score conversion was created for each test, which was used for computing CSEMs for the scale scores.

Table 3: Descriptive Statistics for Scale Scores

Test	Mean	SD	Min.	Max.
Korean	19.99	5.36	3	34
Mathematics	20.00	7.25	3	40
Science	20.03	6.44	2	40
English	20.03	7.68	4	40

To compute the CSEMs for the scale scores, the same formula used for estimating the CSEMs for the arcsine transformed scores (i.e., Equation 9) was employed. Simply, $c(X)$ was replaced by the raw-to-scale score conversion, $s(X)$. The conditional distribution, $\Pr(X = x|\theta)$, did not change. Likewise, the conditional expected scale scores (i.e., true scale scores) were computed using Equation 8 by replacing $c(X)$ with $s(X)$.

3 Results

Figure 1 shows the raw-to-scale score transformations for the four tests. Since the arcsine transformation stretched the raw-score scale at both ends and compressed it in the middle, truncations at the upper and lower ends were inevitable in some cases. The truncated scores appeared parallel to the horizontal axis at the lower and upper ends of the score scales in Figure 1. A fitted linear line is provided for reference along with the transformation functions. Note that the raw-to-scale score transformations are clearly nonlinear showing a pattern of the lowest slope at the mid raw-score points and relatively higher slope at the score points further away from the mid point. This nonlinear pattern of the slope changes in the transformations is expected to produce a pattern of scale-score CSEMs that is different from that of the raw-score CSEMs (Kolen et al., 1992; Brennan & Lee, 1999; Lee et al., 2000). In general, it is anticipated that the transformations will reduce the size of the raw-score CSEMs at the mid raw-score points because of the very low slope of the transformations in the middle, but increase it at the score points where the slope is relatively large. Thus, if the pattern of the CSEMs for the total raw scores is an inverted U-shape, the arcsine transformation will have a flattening effect.

Figure 2 presents estimated CSEMs for the total raw scores. When computing the CSEMs for the total raw scores, a θ range of -4 to 4 was used and not every possible true raw score was attainable. Notice that the estimated total raw-score CSEMs for the four mixed tests have, in general, a non-symmetric, inverted U-shape. The magnitude of the estimated CSEMs along the true raw-score scale varied. The estimated CSEMs tended to be larger in the middle of the score scales and smaller at both ends, which resembled, in general, the shape of raw-score CSEMs for a test consisting of only dichotomous items (i.e., a smooth, symmetric inverted U-shape). Previous studies (Wang et al., 2000; Lee, in press) examined the patterns of CSEMs for tests composed of polytomous items only, and revealed that the CSEMs had an umbrella-shape or a series of M-shapes. For the mixed format tests employed in this paper, the CSEMs seemed to be more similar to the dichotomous case than to the polytomous case, except that the patterns were a bit irregular and nonsymmetric. It is also worth noting that the estimated CSEMs for a test consisting of polytomous items or mixtures of dichotomous and polytomous items tend to be vertically scattered when estimated using the classical models (e.g., Lee, Wang, Kim, & Brennan, 2006; Lee, in press). By contrast, the CSEMs estimated using IRT models tend to be smooth, because of the one-to-one relationship between θs and test characteristic curves.

Plots of the estimated CSEMs for the scale scores are shown in Figure 3. With the ± 4 range of θ, the range of minimum true scale scores across different tests was from 7 for the Korean test to 11 for the Science test. It would not be meaningful to use a wider range of θ because true scores below the sum of the lower asymptote parameter estimates (for the dichotomous items) are not definable anyway. For most of the true scale-score range, the estimated CSEMs were reasonably constant for all tests. The target standard error of measurement was 2 and the goal seemed to be achieved with the procedures used in this study. As often observed with the arcsine-transformed scale scores (Kolen et al., 1992), the estimated CSEMs were lower at high scores because of the truncations near the top to avoid scores above the maximum of 40.

Figure 4 presents plots of the estimated CSEMs for the scale scores of equated forms. The estimated CSEMs seemed to be equalized reasonably well across the true scale scores, although there were slightly more fluctuations compared to those for the base forms. However, the fluctuations seemed to be small enough to be tolerable in practice. Even if Figure 4 suggests that equalizing CSEMs using the arcsine scaling procedure works for the equated forms, it may be necessary to continue to monitor the property of equal CSEMs over subsequently developed forms for equating.

The arcsine-transformed scale scores were compared to several other types of scale scores including total raw scores, T scores, NCE scores, and Stanine scores. In order to compare those scale scores on the same plot, the estimated CSEMs for each type of the scale scores were standardized by dividing them by the standard deviation of the scale scores. Figure 5 exhibits the estimated CSEMs conditioning on θ for the five types of scale scores for the English test. Since similar patterns were observed for the other tests, only the results for the English test are presented here. The plot on the top of Figure 5 shows the results for the base form and the bottom plot displays results for the equated form. Figure 5 clearly shows that the estimated CSEMs for the arcsine-transformed scale scores were more stable than for other scales, except near the top where the CSEMs drop down due to truncations. The patterns of the estimated CSEMs for the T, NCE, and Stanine scores were irregular M-shapes, while the pattern for the raw-score CSEMs was a gradual concave-down parabola. The patterns of the estimated CSEMs for the various scale scores tended to be very similar for the base and equated forms.

4 Discussion

This paper described a scaling method that incorporates score precision information for mixed format tests. In particular, the applicability of the arcsine scaling procedure was evaluated when it was applied to mixed format tests. The arcsine-transformed scale scores were nonlinear transformations of observed scores estimated using mixed IRT models (i.e., 3PL model for dichotomous items and GPC model for polytomous items). The arcsine scaling procedure was applied to real data sets from four different mixed format tests. It was observed that the arcsine transformation procedure compressed the middle of the raw-score scale and stretched the extremes, and a linear transformation of the arcsine-transformed raw scores produced scale scores having reasonably constant estimated CSEMs along the score scales, which appeared to support use of the arcsine transformation procedure for mixed format tests for equalizing CSEMs.

IRT procedures using mixed IRT models were employed for linking and equating two equivalent forms under the common-item nonequivalent groups equating design. It was found in this study that the arcsine transformation procedures successfully stabilized the estimated CSEMs along the score scale for both base and equated forms, even though the estimated CSEMs for the equated forms showed slightly more fluctuation than for the base forms. The arcsine-transformed scale scores were also compared to a few different types of scale scores, which included total raw scores, T scores, NCE scores, and Stanine scores. In general, the CSEMs estimated for those different types of scale scores resulted in more irregular patterns for the estimated CSEMs when compared to the CSEMs for the arcsine-transformed scale scores.

The method provided in this paper would be useful in practice because it incorporates score precision information into the score scale of mixed format tests to facilitate score interpretation. By equalizing CSEMs across the score scale, another advantage of the arcsine transformed score scale is that the user can fix the magnitude of the standard error at a constant value (e.g., 2 in the examples used in this paper) so that a 68% confidence interval, for example, can be constructed easily by adding and subtracting the constant from an examinee's observed scale score.

It should be emphasized, however, that the performance of the arcsine scaling procedure in equalizing scale-score CSEMs depends largely on the shape of the raw-score CSEMs. The examples used in this paper showed approximately an inverted U-shape for the estimated total raw-score CSEMs, and thus the arcsine transformation, by compressing the middle and stretching the extremes of the raw-score scale, worked fairly well. It is not always guaranteed, however, that the shape of the CSEMs for total raw scores obtained from mixtures of dichotomous and polytomous items will be an inverted U-shape, and therefore, it is important to carefully examine the pattern of the estimated total raw-score CSEMs before employing the arcsine scaling procedure for mixed format tests. Especially, when a test consists of only polytomous items, the raw-score CSEMs may not show a simple pattern (see, for example, Wang et al., 2000; Lee et al., 2006; Lee, in press). Therefore, the results reported here should be interpreted with caution because the methodology described in this paper was not investigated for tests consisting of only polytomous items, and was only partially studied for tests having various differential proportions of dichotomous and polytomous items.

It was shown in this study that IRT procedures could be used effectively in conjunction with the arcsine scaling procedure. In particular, this study demonstrated that when estimated observed scores using IRT for mixed tests were obtained, the estimated observed scores may be transformed to equalize CSEMs in the process of establishing reported score scales. A useful future study would involve using different combinations of IRT models other than the ones used in this study (i.e., the 3PL and GPC models).

5 References

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, *22*, 307–331.

Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement*, *59*, 5–24.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics*, *21*, 607–611.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* Unpublished research note.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models.* Iowa City, IA: Iowa Testing Programs, The University of Iowa.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, *43*, 53–76.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, *25*, 97–110.

Kolen, M. J. (2004). *POLYEQUATE: A computer program.* Iowa City, IA: University of Iowa. (Available from www.education.uiowa.edu/casma).

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 154–186). New York: American Council on Education and Macmillan.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). NY: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. Journal of Educational Measurement, 29, 285–307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*, 129–140.

Lee, W. (in press). A multinomial error model for tests with polytomous items. *Applied Psychological Measurement.*

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement*, *37*, 1–20.

Lee, W., Wang, T., Kim, S., & Brennan, R. L. (2006). *A strong true-score model for polytomous items* (CASMA Research Report No. 16). Iowa City, IA: University of Iowa.

Lord, F. M. (1965). A strong true-score theory with applications. *Psychometrika*, *30*, 239–270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing. (An empirical Bayes estimation problem.) *Psychometrika*, *34*, 259–299.

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

Lord, F. M. (1984). Standard errors of measurement at different score levels. *Journal of Educational Measurement*, *21*, 239–243.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Measurement in Education*, *8*, 452–461.

Muraki, E., & Bock, R.D. (2003). *PARSCALE (version 4.1): IRT item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software, Inc.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Petersen, N. S., Kolen, M. J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: American Council on Education and Macmillan.

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

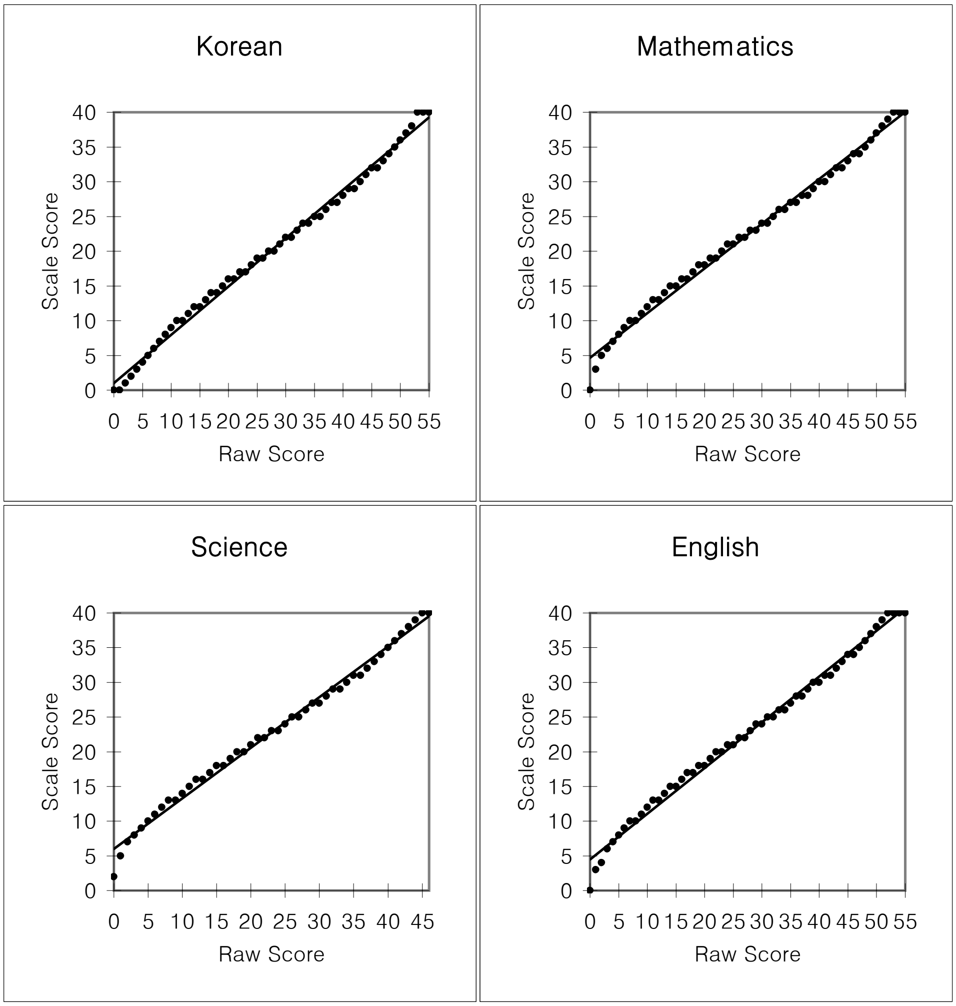

Figure 1: Raw-to-Scale Score Transformations

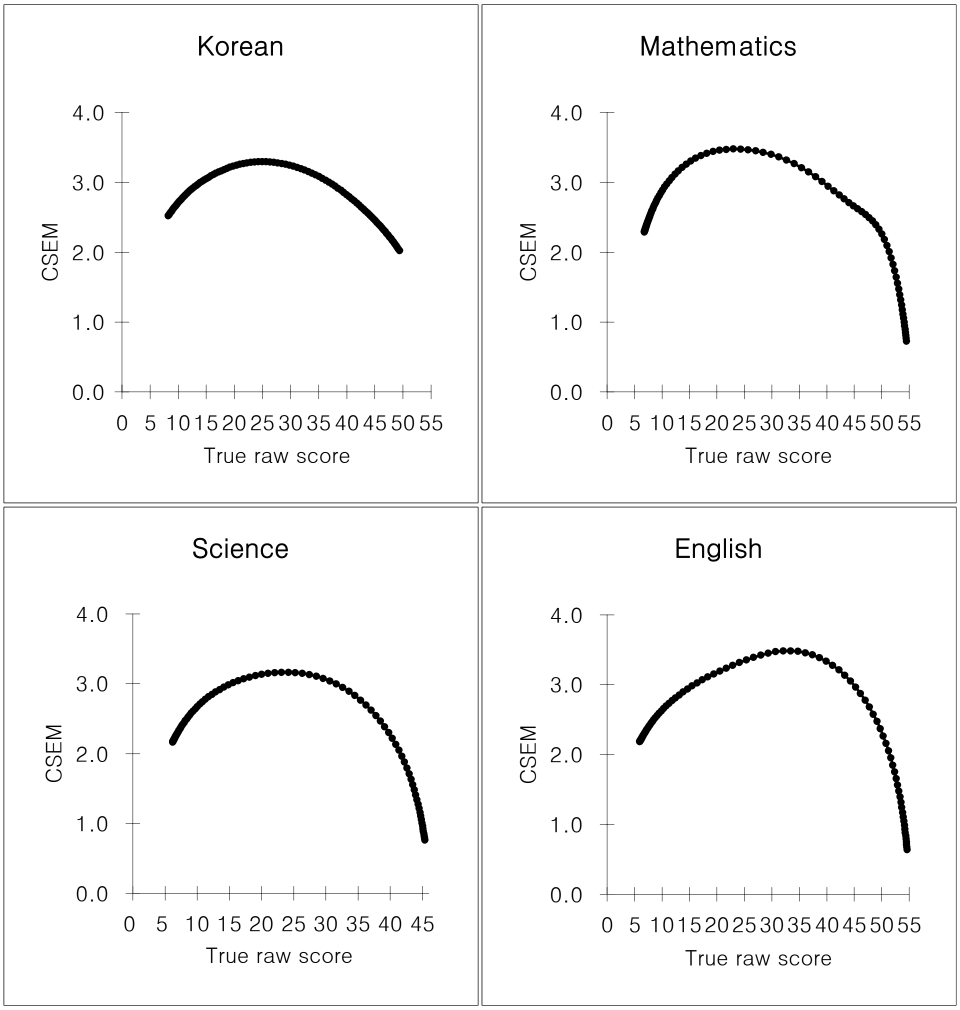

Figure 2: Conditional Standard Errors of Measurement for Total Raw Scores

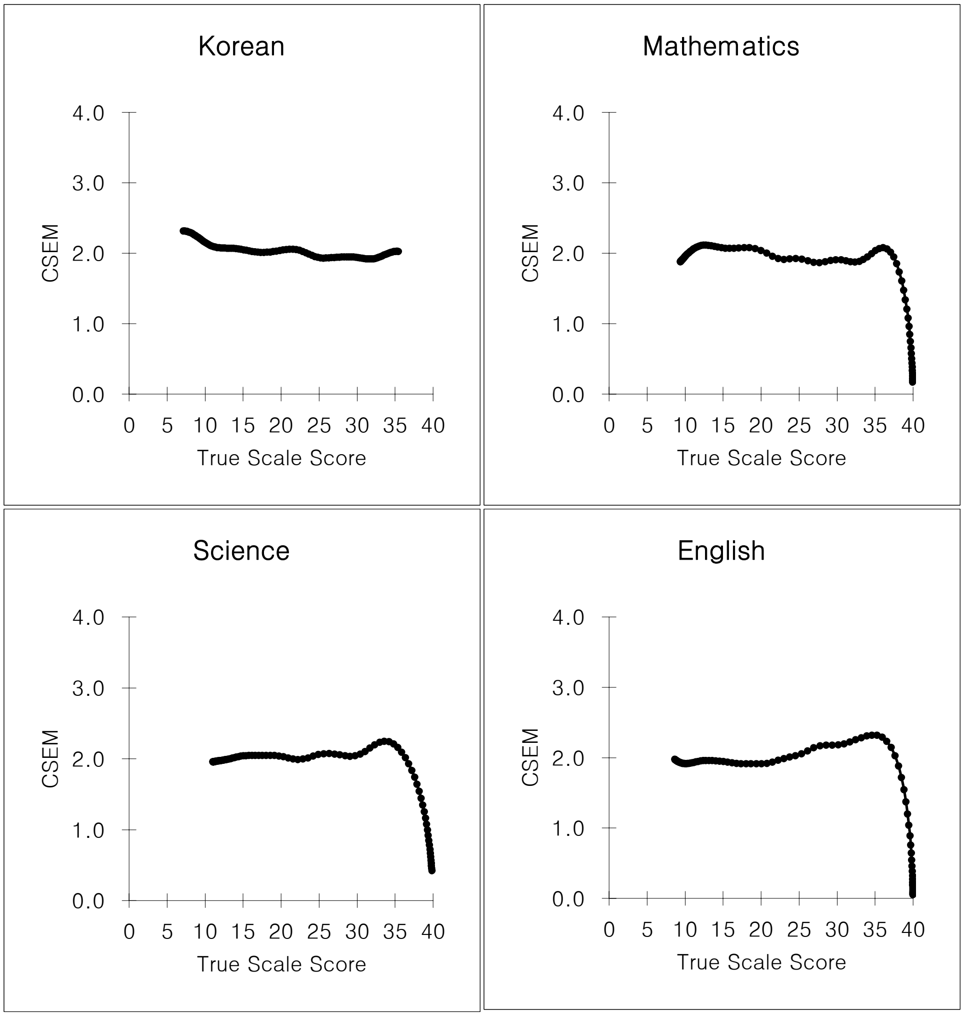

Figure 3: Conditional Standard Errors of Measurement for Scale Scores: Base Form

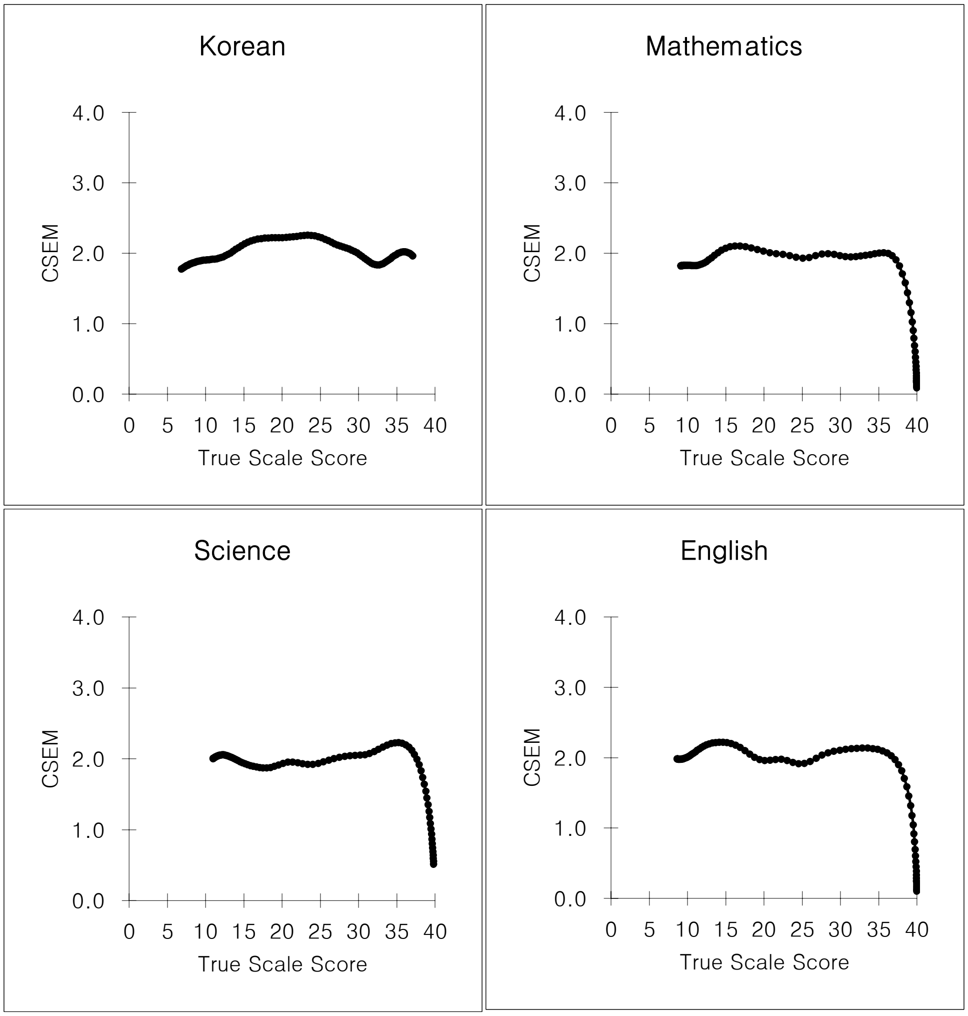

Figure 4: Conditional Standard Errors of Measurement for Scale Scores: Equated Form

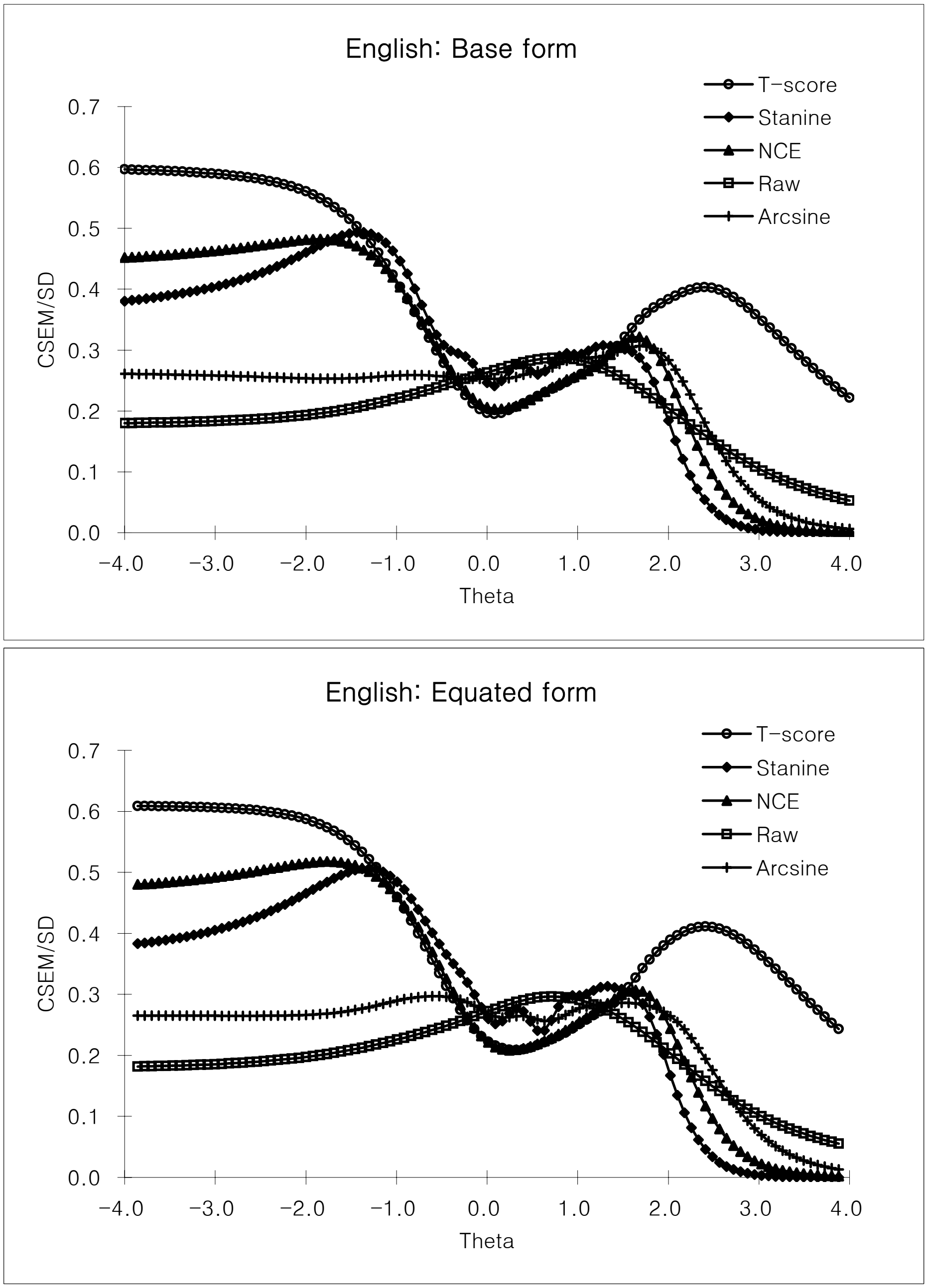

Figure 5: Conditional Standard Errors of Measurement for Various Scale Scores