Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 31

Assessing Equating Results Based on First-order and Second-order Equity*

Eunjung Lee, Won-Chan Lee, Robert L. Brennan†

December 2010

---

*A revised version of a paper presented at the National Council on Measurement in Education, Denver, CO, April 2010.

†Eunjung Lee is a research assistant in the Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: eunjung-lee@uiowa.edu). Won-Chan Lee is Associate Director of the CASMA, College of Education, University of Iowa (email: won-chan-lee@uiowa.edu). Robert L. Brennan is the E. F. Lindquist Chair in Measurement and Testing and Director of the CASMA, College of Education, University of Iowa (email: robert-brennan@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

1 Introduction **1**
1.1 First-order and Second-order Equity 1
1.2 Dichotomous IRT procedure for FOE and SOE 1
1.3 Research Purpose 2

2 Method **3**
2.1 Equating Design and Data 3
2.2 Equating Methods 3
2.3 Evaluation Criteria 4
2.4 Computer Programs 5

3 Results **5**

4 Discussion and Limitations **7**

5 References **8**

6 Tables **11**

7 Figures **17**

1 Introduction

A variety of equating methods have been developed and widely applied in many testing areas. Accordingly, it has become increasingly important for researchers and practitioners to select the appropriate equating method for a given situation. To this end, prior studies have compared and evaluated the results of various equating methods using different equating criteria (Han, Kolen, & Pohlmann, 1997; Kim, Brennan, & Kolen, 2005; Tong & Kolen, 2005; Wang, Hanson, & Harris, 2000). The primary purpose of this study is to provide empirical evidence for the adequacy of the various equating methods, based on first-order and second-order equity (Morris, 1982).

1.1 First-order and Second-order Equity

According to *Lord's equity property of equating* (Lord, 1980), after equating, the conditional distributions of observed scores given examinees' true scores for each new and old form should be same. First- and second-order equity are a less restrictive version of Lord's equity property of equating (Lord, 1980; Morris, 1982). First-order equity (FOE) holds "to the extent that conditional expected scale scores are similar for the alternate forms" (Kolen & Brennan, 2004, p. 301). Second-order equity (SOE) holds "to the extent that the conditional standard errors of measurement, after equating, are similar for the alternate forms" (Kolen & Brennan, 2004, p. 301).

By definition, FOE and SOE are conditional on true scores. Therefore, assessing the performance of various equating methods based on FOE and SOE requires computing both expected scale scores and conditional standard errors of measurement (CSEMs), under the assumption of some psychometric model (Kolen & Brennan, 2004). Previous researchers have described procedures which can be used to compute conditional expected scores and CSEMs for different psychometric models. For example, Kolen, Hanson, and Brennan's (1992) procedure is based on a strong true score model, Kolen, Zeng, and Hanson's (1996) procedure uses a dichotomous item response theory (IRT) model, and Wang et al.'s (2000) procedure can be used with a polytomous IRT model. Brennan (2010) provides an extensive discussion of FOE and SOE, including conditions that facilitate achieving FOE and SOE, based primarily on classical test theory.

1.2 Dichotomous IRT procedure for FOE and SOE

Kolen et al. (1996) described a procedure for estimating expected (true) scores and CSEMs of scale scores using a dichotomous IRT model. To estimate the conditional distribution of scale scores, the distribution of number-correct raw scores given IRT ability (θ) and item parameters need to be obtained (Kolen et al., 1996). The conditional distribution of number-correct raw scores given ability can be obtained using a recursion formula provided by Lord and Wingersky (1984), given the item parameters for the test. Then, the number-correct

raw-score to scale-score transformation can be applied to the conditional distribution of number-correct raw scores at a given ability to produce the conditional probability distribution of scale scores given θ. The standard deviation of this scale score distribution at a given IRT ability level is taken to be the CSEM at that ability.

The mean of this conditional distribution is the true (expected) scale score at that θ, which is

$$\xi(\theta) = \boldsymbol{E}[s(X)|\theta] = \sum_{i=0} s(i)\Pr(X = i|\theta). \quad (1)$$

In this equation, $\boldsymbol{E}$ refers to the expected value, the raw to scale-score transformation is symbolized s, and $\Pr(X = i|\theta)$ represents the probability that the raw score random variable X is equal to i ($i = 0, 1, \ldots, N$) on a test of N items for ability θ.

The conditional standard error of measurement of scale scores at θ is

$$\sigma[s(X)|\theta] = \sqrt{\boldsymbol{E}[s(X) - \xi(\theta)]^2|\theta\}} = \sqrt{\sum_{i=0}[s(i) - \xi(\theta)]^2\Pr(X = i|\theta)}. \quad (2)$$

Tong and Kolen (2005) used this dichotomous IRT procedure (Kolen et al., 1996) and compared the performance of three equating methods (smoothed equipercentile equating using the log-linear presmoothing, IRT true score equating, and IRT observed score equating) in terms of FOE and SOE.

1.3 Research Purpose

This study is an extension of Tong and Kolen's (2005) study. Especially we intend to make incremental contributions to theirs. First, we compare a wider variety of equating methods including the ones which have recently been developed. Second, we compare results based on different types of scores; number-correct score, developmental standard scores and grade equivalent scores. Third, we use different evaluation criteria. We use both normal and uniform weights in examining whether a equating method satisfies FOE or SOE. Also, we compute discrepancy indices in a different way, which will be discussed in detail later. Examining a greater number of equating methods and adopting different perspectives on evaluation may lead to a better understanding of equity in equating.

In particular, the performance of seven different equating methods is evaluated and compared: (a) IRT true score equating (Lord, 1980), (b) IRT observed score equating (Kolen, 1981; Lord, 1980), (c) unsmoothed equipercentile equating (Angoff, 1971), (d) smoothed equipercentile equating using the log-linear presmoothing (Holland & Thayer, 1987), (e) smoothed equipercentile equating using the cubic-spline postsmoothing (Kolen, 1984), (f) kernel equating (von Davier, Holland, & Thayer, 2004), and (g) continuized log-linear equating (Wang, 2008). The first two methods are based on the three-parameter logistic (3PL) model.

2 Method

2.1 Equating Design and Data

The data set for this study was obtained from the Iowa Test of Basic Skills (ITBS) Forms K and L of Level 9 (Hoover, Hieronymus, Frisbie, & Dunbar, 1993). These two test forms were spiraled within classrooms for the equating study, making the two groups taking the two forms randomly equivalent. The sample size for each form was 2,000. Four tests were used in this study: (a) Vocabulary, (b) Usage and Expression (Usage/Expression), (c) Math Problem Solving and Data Interpretation (Math), and (d) Social Studies. The four tests of Form K (the new form) were equated to the corresponding tests of Form L (the old form). Descriptive statistics for the two forms of the tests, including the number of items, means, standard deviations (SD), α coefficients, and effect sizes $((\overline{X}_1 - \overline{X}_2)/\sqrt{({S_1}^2 + {S_2}^2)}/2$,Yen, 1986) are presented in Table 1. The frequency distributions of Form K and Form L for all tests are given in Figure 1. In terms of test difficulty, Form L appears to be somewhat easier than Form K for the Social Studies and the Usage/Expression tests, whereas Form K appears to be slightly easier than Form L for the Vocabulary and Math tests. The effect sizes for the two forms are bigger for Social Studies and Usage/Expression tests than for the Vocabulary and Math tests.

FOE and SOE for the equated scores on alternate forms were assessed in terms of raw scores (number-correct scores), as well as two types of scale scores, developmental standard scores and grade equivalent scores. Developmental standard scores and grade equivalent scores indicate students' locations on an achievement continuum (Hoover et al., 1993). The raw score median for third-grade students was defined to be a developmental standard score of 185 and a grade equivalent score of 3.8 (Hoover et al., 1993). Operational conversion tables for Form L (Hoover et al., 1993) were used.

2.2 Equating Methods

To achieve the goals of this study, seven equating methods were used. The first method was the IRT true score equating method, based on the 3PL IRT model (Lord, 1980). IRT true score equating establishes a relationship between true scores on both forms in which the number-correct true scores for two forms are associated with the same θ through their test characteristic functions. The second method was the IRT observed score equating method, which uses the 3PL IRT model to produce an estimated distribution of the observed number-correct scores on each form. These scores were then equated using equipercentile equating.

The third method was the unsmoothed equipercentile equating method. The fourth was the log-linear pre-smoothing equating, which smooths the raw score distribution prior to obtaining the equipercentile equivalents for the new form on the scale of the old form. The log-linear method was fitted using the smoothing parameter $C = 6$. The fifth method, the post-smoothing equating, uses cubic

splines to directly smooth the equating relationship. The equating was conducted using five smoothing parameter values ($S = .05, .10, .20, .30$, and $.50$). After examining the raw-to-raw equivalents for these equating results, $S = .05$ was chosen to conduct the cubic spline post-smoothing equipercentile equating for all four tests. Using $S = .05$, the equivalents appeared to be smooth and were within the one standard error bands at most of the points for all of the tests. Extensive illustrations of these aforementioned five methods are presented in Kolen and Brennan (2004).

The sixth method, the kernel equating method, uses a Gaussian kernel (Tapia & Thompson, 1978; Silverman, 1986) to fit a continuous distribution to the discrete score distribution. The purpose of doing so is to deal with the difficulty of performing equipercentile equating for a discrete score distribution (see von Davier, Holland, & Thayer, 2004, for details). For kernel equating, two bandwidths for Form K and Form L were estimated using two penalty functions introduced by von Davier et al. (2004). Except for Social Studies, the bandwidths for the two forms were in a range from .56 to .59 for all the tests. For the Social Studies, the bandwidths for Form K was .61 and that for Form L was 1.52.

The seventh method was the continuized log-linear method, proposed by Wang (2008) that directly uses the log-linear function from the kernel smoothing step as an alternative continuization method.

In this study, equivalent scores were truncated at the upper and/or lower extremes, so that the scores don't exceed the minimum and/or maximum score point of the old forms. For all the raw and scale scores, the highest possible scores and the lowest possible scores were fixed. For example, for raw scores, the lowest possible score was zero and the highest possible score was the number of items for each test. Linear interpolation was used to compute some raw or scale score equivalents in both tails.

2.3 Evaluation Criteria

We computed the expected scale scores and CSEMs using the 3PL model (Kolen et al., 1996). Tong and Kolen (2005) computed discrepancy indices (D_1 and D_2) to empirically demonstrate the adequacy of preserving FOE and SOE. The present study computed similar types of overall discrepancy indices, motivated by Tong and Kolen (2005):

$$D_1 = \frac{\sum_i w_i |(SC_L|\theta_i) - (SC_K|\theta_i)|}{\sum_i w_i} \tag{3}$$

and

$$D_2 = \sqrt{\frac{\sum_i w_i |(EV_L|\theta_i) - (EV_K|\theta_i)|}{\sum_i w_i}}. \tag{4}$$

In Equation 3 and Equation 4, $SC_K|\theta_i$ and $SC_L|\theta_i$ refer to the expected scale scores of Form K and Form L, respectively, for quadrature point θ_i, and w_i refers to the weight of θ_i. For SOE, the D_2 index was created using $EV_K|\theta_i$ and

$EV_L|\theta_i$, referring to the conditional error variance of measurements for the two forms at quadrature point θ_i. The differences between our equations and Tong and Kolen's (2005) are as follows. In this paper, we use the absolute values of the differences between $SC_K|\theta_i$ and $SC_L|\theta_i$, instead of using the squared value for the calculation of D_1. We also use the conditional error variance of measurement (EVs), rather than the CSEMs for the calculation of D_2. Furthermore, this study uses two types of weights to compute D_1 and D_2: a weight from a normal distribution and a weight from a uniform distribution. As illustrated later, differences in weights can lead to different conclusions regarding the performance of the various equating methods.

2.4 Computer Programs

IRT calibration was conducted using BILOG-MG 3.0 (Zimowski, Muraki, Mislevy, & Bock, 1995), and all equating procedures were performed using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009). Expected scale scores and CSEMs were separately computed using POLYCSEM (Kolen, 2004). *Equating Recipes* and POLYCSEM is available on the CASMA website (http://www.education.uiowa.edu/casma).

3 Results

The differences in raw-to-raw equivalents for the four tests are depicted in Figure 2. Except for the unsmoothed equipercentile equating method, the other six equating methods resulted in relatively smooth curves. For the Vocabulary and Math tests, the raw-to-raw equivalents plots were below the identity equating line (the vertical value of zero), because the new forms for these two tests were slightly easier than the old forms. For the Social Studies and Usage/Expression tests, on the other hand, the raw-to-raw equivalents were above the identity equating line. The summary statistics for equating results of these seven equating methods are presented in Table 2. The first and second moments for the kernel method were the closest to those for Form L. Generally, traditional equating methods[1] yielded moments closer to those for Form L than the two IRT methods.

The differences in the conditional expected number-correct score distributions are presented in Figure 3. The differences in conditional expected scale score distributions for the developmental scale scores and the grade equivalent scores are presented in Figures 4 and 5, respectively. If FOE held perfectly, the curves in these graphs would have been coincident with the horizontal zero line, meaning that the expected scale scores yielded by the equivalents would be the same as the expected scale scores for the old form. The IRT true score method for all four tests seemed to preserve FOE relatively well throughout the whole score range for all three score scales. For traditional equating methods with

[1]In this paper, we categorized all five non-IRT based equating methods as traditional equating methods for descriptive purposes.

the Math test, the curves for all traditional equating methods overlapped. This indicates that these methods yield no big differences in the extent to which FOE held. For the other three tests (i.e., Vocabulary, Usage/Expression, and Social Studies), both the postsmoothing method and the continuized log-linear method were least likely to violate FOE. This pattern of findings for the number-correct scores and FOE seemed to be replicated for the developmental scale scores and the grade equivalent scores.

With regard to the differences in CSEMs for both forms, the results for the number-correct scores, the developmental scale scores, and the grade equivalent scores are illustrated in Figures 6, 7, and 8, respectively. The pattern of findings for SOE was less consistent than that for FOE. The IRT observed score method, relative to the other methods, seemed to yield results closest to the horizontal zero line. Among traditional equating methods, SOE was likely to hold for the kernel equating somewhat better than for the other traditional equating methods; this was the case for all tests, except for high scores in Social Studies. The postsmoothing equating was found to satisfy SOE better than the presmoothing method for Social Studies and Vocabulary tests, and vice versa, for Usage/Expression tests.

For the Math test, the SOE curves from the traditional methods did not depart much from each other, which was the same general result for FOE (see Figures 6, 7, and 8). As shown in the findings for FOE, there was little difference in the shapes of the curves, based on the three score types (number-correct, developmental, and grade equivalent scores).

The overall FOE and SOE indices (D_1 and D_2) for all tests and the three types of scores are presented in Tables 3 through 6. The values of D_1 and D_2 in Tables 3 and 5 were computed using uniform weights. Those in Tables 4 and 6 were computed using weights from a normal distribution. Note that, given the idiosyncratic characteristics of each test, it may not be meaningful to calculate the sum of the D values across tests. That is why we used standardized D values to obtain the sums and the rank order presented below the D_1 and D_2 values in Tables 3 through 6. To compute the standardized D values, D_1 and D_2 values in Tables 3 through 6 were divided by the standard deviation for the equivalents score scale of each test for Form L. In general, procedures with lower ranks better satisfied first-order (or second-order) equity.

Tables 3 and 4 report how well each equating method satisfies FOE. The IRT true score method preserved FOE better than any other method, regardless of weighting. The IRT observed score method tended to preserve FOE better than the traditional equating methods but worse than the IRT true score method. In addition, for the Math test, the overall indices from the traditional methods did not seem to differ much. This was consistent with the previously discussed graphs (Figures 3 through 5). The continuized log-linear method was found to be least likely to violate FOE among the traditional methods. Comparing Tables 3 and 4, the differences in the expected scores for two forms were smaller when the normal weights were used rather than the uniform weights regardless of the types of scores used.

Tables 5 and 6 report how well each equating method satisfies SOE. Regard-

less of weighting, the IRT true score method, which was most likely to preserve FOE, was not found to satisfy SOE well. The kernel method was found to be less likely to violate SOE than the continuized log-linear method for the Usage/Expression and the Math tests, and vice versa for Vocabulary and Social Studies. Overall, in most of the cases, the standardized D_2 values for the Vocabulary and the Math tests were larger than those for the Social Studies and the Usage/Expression tests. In other words, SOE is less likely to be preserved for the Social Studies and the Usage/Expression tests. This finding supports Tong and Kolen's (2005) results, illustrating that the greater the difficulty difference between the alternate forms, the more likely SOE will be violated.

When normal weights were used to compute D_2, the IRT observed score method was found to be the most likely method to satisfy SOE. When uniform weights were used, the postsmoothing method was found to be the most likely method to preserve SOE. In general, the differences in CSEMs were larger when normal weights were used than when uniform weights were used for raw scores. However, for the grade equivalent scores, the pattern was opposite; the differences in CSEMs were smaller when normal weights were used than when uniform weights were used.

4 Discussion and Limitations

Overall, our findings are consistent with previous research (Tong & Kolen, 2005). The results support using the IRT true score method, if preserving FOE is the main concern, and using the IRT observed score method, if preserving SOE is the main concern. Among traditional equating methods, the continuized log-linear method would be preferred if preserving FOE is the main concern. On the other hand, the postsmoothing method would be preferred if preserving SOE is the main concern. However, differences among traditional methods with respect to FOE and SOE are usually not very large. It is important to note that these general results are solely based on the random groups design and a few real data sets. In our study, the test forms to be equated do not differ much in raw score distributions. According to Tong and Kolen (2005), to preserve FOE and SOE, it is important that the forms to be equated be very similar in difficulty. Results might not be the same in a situation where big form differences exist or where a different equating design is used.

The present study uses the 3PL IRT model to estimate the true score distribution to assess FOE and SOE. Therefore, our results depend on how well the IRT model fits the data. Furthermore, IRT true score equating might have potential advantages over the traditional equating methods in terms of FOE in our study, because true score distributions were estimated here based on the IRT model. Kim, Brennan, and Kolen (2005) showed that true score equating more closely achieved estimated FOE than observed score equating when the true score distribution was estimated using the same psychometric model that was used in equating. Therefore, it would be fruitful to evaluate FOE and SOE based on an alternative model such as the four-parameter beta compound

binomial model (Lord, 1965), or beta 4, in future studies.

It is worth mentioning that FOE and SOE only consider conditional means and standard deviations of true scores for two forms. It is obvious that we do not get all the information about true score distributions from these moments. FOE and SOE should not be the only criteria that we use to assess equating results. Additionally, this study does not treat issues concerning reliability in equity. Brennan (2010) has demonstrated that for curvilinear equating FOE and SOE are more likely to be satisfied approximately as reliability gets higher. Hence, examining the role of reliability would also give us valuable information about equating relationships.

5 References

Angoff, W. H. (1971). *Scales, norms, and equivalent scores.* In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–600). Washington, DC: American Council on Education. (Reprinted as W. A. Angoff, *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service, 1984).

Brennan, R. L. (2010). *First-order and Second-order Equity in Equating* (CASMA Research Report No. 30). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating Recipes* (CASMA Monograph No. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Han, T., Kolen, M. J., & Pohlmann, J. (1997). A comparison among IRT true- and observed- score equating and traditional equipercentile equating. *Applied Measurement in Education, 10*, 105–121.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195–240.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Tech. Rep. 87–79). Princeton, NJ: Educational Testing Service.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. D. (1993). *The Iowa Tests of Basic Skills: Norms and Score Conversions: Form L: Levels79–14.* Itasca, IL: Riverside.

Kim, D. I., Brennan, R. L., & Kolen, M. J. (2005). A comparison of IRT equating and beta 4 equating. *Journal of Educational Measurement, 42(1)*, 77–99.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1–11.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25–44.

Kolen, M. J. (2004). *POLYCSEM (Windows Console Version)*. [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33(2)*, 129–140.

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

Morris, G. M. (1982). *On the foundations of test equating*. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 169–191). New York: Academic Press.

Peterson, 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: Macmillan.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Tapia, R. A., & Thompson, J. R. (1978). *Non–parametric probability density estimation*. Baltimore: Johns Hopkins University Press.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29(6)*, 418–432.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer–Verlag.

Wang, T. (2008). The continuized log-linear method: An alternative to the kernel method of continuization in test equating. *Applied Psychological Measurement, 32*, 527–542.

Wang, T., Hanson, B. A., & Harris, D. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement, 24*, 195–210.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23(4)*, 299–325.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1995) *BILOG–MG: Multiple-group item analysis and test scoring.* Chicago: Scientific Software Int'l.

6 Tables

Table 1: Summary Statistics for Both Forms

	# of	Form L			Form K			Effect
Subject	items	Mean	SD	α	Mean	SD	α	Size
Vocabulary	26	13.58	5.74	0.858	14.09	5.79	0.865	-0.088
Usage/Expression	31	16.60	6.54	0.859	15.26	6.44	0.852	0.206
Math	24	12.29	4.87	0.817	12.92	4.90	0.819	-0.128
Social Studies	30	15.67	5.78	0.825	14.18	5.43	0.788	0.265

Table 2: Number-Correct Score Moments for the Four ITBS Tests

	Eq%ile	Pre-Eq	Post-Eq	Kernel	CLL	IRT-True	IRT-Obs	Form L
Vocabulary								
$\widehat{\mu}$	13.576	13.574	13.579	13.578	13.601	13.585	13.566	13.579
$\widehat{\sigma}$	5.728	5.727	5.740	5.736	5.758	5.682	5.704	5.739
$\widehat{sk}$	0.041	0.040	0.047	0.043	0.052	0.087	0.072	0.046
$\widehat{ku}$	2.056	2.057	2.066	2.063	2.073	2.120	2.121	-0.936
Usage/Expression								
$\widehat{\mu}$	16.601	16.601	16.595	16.600	16.596	16.515	16.581	16.600
$\widehat{\sigma}$	6.532	6.534	6.540	6.542	6.539	6.631	6.549	6.543
$\widehat{sk}$	-0.058	-0.057	-0.062	-0.055	-0.060	-0.091	-0.028	-0.059
$\widehat{ku}$	2.002	1.997	2.011	2.005	2.003	2.065	2.032	-0.991
Math								
$\widehat{\mu}$	12.289	12.290	12.289	12.291	12.290	12.333	12.293	12.291
$\widehat{\sigma}$	4.858	4.861	4.870	4.865	4.866	4.869	4.828	4.868
$\widehat{sk}$	0.093	0.095	0.081	0.091	0.097	0.114	0.074	0.095
$\widehat{ku}$	2.175	2.182	2.185	2.179	2.182	2.284	2.222	-0.820
Social Studies								
$\widehat{\mu}$	15.662	15.666	15.658	15.666	15.658	15.622	15.674	15.666
$\widehat{\sigma}$	5.771	5.775	5.786	5.781	5.787	5.971	5.844	5.781
$\widehat{sk}$	0.014	0.015	0.004	0.016	0.007	-0.107	-0.066	0.016
$\widehat{ku}$	2.097	2.100	2.145	2.199	2.116	2.194	2.220	-0.895

Note. Eq%ile = Unsmoothed equipercentile equating; Pre-Eq = Log-linear presmooting with C = 6; Post-Eq = Cubic-spline postsmoothing with S = .05; Kernel = Kernel equating; Cll = Continuized log-linear equating; IRT-True = IRT true score equating; IRT-Obs = IRT observed score equating.

Table 3: Overall First-Order Equity Index (D_1) for the Seven Equating Methods (Uniform Weights)

	Eq%ile	Pre-Eq	Post-Eq	Kernel	CLL	IRT-True	IRT-Obs
Raw							
Vocabulary	0.199	0.201	0.138	0.178	0.110	*0.043*	0.099
Usage/Expression	0.302	0.283	0.247	0.291	0.273	*0.095*	0.271
Math	0.176	0.177	0.186	0.177	0.175	*0.080*	0.147
Social Studies	0.374	0.392	0.351	0.491	0.360	*0.098*	0.230
SUM	0.182	0.182	0.161	0.197	0.159	*0.055*	0.129
RANK	5	6	4	7	3	1	2
Developmental Standard Scores							
Vocabulary	1.483	1.511	0.817	1.252	0.540	*0.270*	0.609
Usage/Expression	1.724	1.442	1.186	1.579	1.402	*0.504*	1.347
Math	1.491	1.555	1.575	1.563	1.551	*0.591*	1.134
Social Studies	1.433	1.544	1.183	2.076	1.245	*0.361*	0.765
SUM	0.293	0.292	0.229	0.315	0.227	*0.082*	0.181
RANK	6	5	4	7	3	1	2
Grade Equivalent Scores							
Vocabulary	0.107	0.109	0.053	0.088	0.032	*0.019*	0.041
Usage/Expression	0.149	0.109	0.083	0.125	0.105	*0.036*	0.099
Math	0.120	0.125	0.125	0.125	0.124	*0.045*	0.088
Social Studies	0.081	0.101	0.071	0.165	0.075	*0.024*	0.047
SUM	0.341	0.340	0.254	0.391	0.254	*0.093*	0.204
RANK	6	5	4	7	3	1	2

Note. SUM = sum of standardized D_1s, where the summation was taken over tests; RANK = rank order of the SUM, where 1 indicates the smallest discrepancy and 7 indicates the largest discrepancy; Figures in italics indicates the corresponding equating method yields the smallest discrepancy for a test.

Table 4: Overall First-Order Equity Index (D_1) for the Seven Equating Methods (Normal Weights)

	Eq%ile	Pre-Eq	Post-Eq	Kernel	CLL	IRT-True	IRT-Obs
Raw							
Vocabulary	0.127	0.122	0.123	0.121	0.122	*0.046*	0.072
Usage/Expression	0.151	0.148	0.141	0.145	0.146	*0.084*	0.134
Math	0.119	0.114	0.123	0.119	0.115	*0.084*	0.113
Social Studies	0.222	0.219	0.224	0.241	0.218	*0.088*	0.162
SUM	0.108	0.105	0.107	0.109	0.105	*0.054*	0.084
RANK	6	4	5	7	3	1	2
Developmental Standard Scores							
Vocabulary	0.509	0.502	0.462	0.481	0.456	*0.189*	0.273
Usage/Expression	0.686	0.671	0.643	0.660	0.674	*0.431*	0.540
Math	0.610	0.624	0.641	0.643	0.632	*0.487*	0.546
Social Studies	0.621	0.636	0.628	0.739	0.622	*0.227*	0.440
SUM	0.116	0.117	0.114	0.122	0.114	*0.063*	0.086
RANK	5	6	3	7	4	1	2
Grade Equivalent Scores							
Vocabulary	0.032	0.032	0.028	0.030	0.028	*0.013*	0.017
Usage/Expression	0.045	0.043	0.042	0.043	0.044	*0.028*	0.032
Math	0.041	0.043	0.043	0.044	0.043	*0.034*	0.036
Social Studies	0.034	0.035	0.034	0.044	0.034	*0.012*	0.024
SUM	0.115	0.117	0.113	0.123	0.113	*0.065*	0.083
RANK	5	6	3	7	4	1	2

Note. SUM = sum of standardized D_1s, where the summation was taken over tests; RANK = rank order of the SUM, where 1 indicates the smallest discrepancy and 7 indicates the largest discrepancy; Figures in italics indicates the corresponding equating method yields the smallest discrepancy for a test.

Table 5: Overall Second-Order Equity Index (D_2) for the Seven Equating Methods (Uniform Weights)

	Eq%ile	Pre-Eq	Post-Eq	Kernel	CLL	IRT-True	IRT-Obs
Raw							
Vocabulary	0.598	0.593	0.569	0.578	0.558	0.543	*0.524*
Usage/Expression	0.867	0.856	0.917	0.866	0.884	1.081	*0.855*
Math	0.582	0.570	*0.544*	0.562	0.572	0.596	0.575
Social Studies	0.793	0.835	*0.757*	0.819	0.819	1.083	0.986
SUM	0.493	0.496	*0.482*	0.490	0.492	0.570	0.511
RANK	4	5	1	2	3	7	6
Developmental Standard Scores							
Vocabulary	3.550	3.412	2.824	3.184	2.888	2.616	*2.574*
Usage/Expression	3.158	3.120	*3.014*	3.099	3.137	3.654	3.427
Math	4.357	4.297	4.646	*4.221*	4.384	5.205	4.579
Social Studies	3.258	3.352	*3.021*	4.003	3.203	3.452	3.263
SUM	0.680	0.675	*0.637*	0.695	0.646	0.705	0.656
RANK	5	4	1	6	2	7	3
Grade Equivalent Scores							
Vocabulary	0.252	0.239	0.174	0.218	0.190	0.169	*0.165*
Usage/Expression	0.354	0.331	0.362	*0.330*	0.340	0.393	0.366
Math	0.229	0.236	*0.224*	0.235	0.238	0.267	0.246
Social Studies	*0.202*	0.232	0.217	0.312	0.220	0.223	0.222
SUM	0.775	0.783	*0.730*	0.837	0.742	0.785	0.748
RANK	4	5	1	7	2	6	3

Note. SUM = sum of standardized D_2s, where the summation was taken over tests; RANK = rank order of the SUM, where 1 indicates the smallest discrepancy and 7 indicates the largest discrepancy; Figures in italics indicates the corresponding equating method yields the smallest discrepancy for a test.

Table 6: Overall Second-Order Equity Index (D_2) for the Seven Equating Methods (Normal Weights)

	Eq%ile	Pre-Eq	Post-Eq	Kernel	CLL	IRT-True	IRT-Obs
Raw							
Vocabulary	0.601	0.603	0.590	0.592	*0.565*	0.651	0.618
Usage/Expression	0.987	0.983	0.986	0.981	0.994	1.090	*0.868*
Math	0.699	0.708	0.674	0.696	0.713	0.749	*0.649*
Social Studies	1.155	1.164	*1.150*	1.160	1.167	1.389	1.245
SUM	0.599	0.602	0.591	0.597	0.599	0.674	*0.589*
RANK	5	6	2	3	4	7	1
Developmental Standard Scores							
Vocabulary	2.699	2.663	2.432	2.541	*2.246*	2.488	2.327
Usage/Expression	4.279	4.196	4.324	4.088	4.221	4.689	*3.712*
Math	3.221	3.296	*3.137*	3.244	3.331	4.083	3.299
Social Studies	*3.294*	3.406	3.300	3.662	3.368	3.964	3.552
SUM	0.641	0.646	0.626	0.647	0.627	0.727	*0.618*
RANK	4	5	2	6	3	7	1
Grade Equivalent Scores							
Vocabulary	0.178	0.176	0.153	0.166	0.141	0.148	*0.139*
Usage/Expression	0.289	0.281	0.295	0.270	0.282	0.315	*0.248*
Math	0.212	0.220	*0.209*	0.217	0.223	0.289	0.225
Social Studies	0.195	0.205	*0.195*	0.239	0.199	0.229	0.206
SUM	0.658	0.667	0.641	0.680	0.638	0.742	*0.623*
RANK	4	5	3	6	2	7	1

Note. SUM = sum of standardized D_2s, where the summation was taken over tests; RANK = rank order of the SUM, where 1 indicates the smallest discrepancy and 7 indicates the largest discrepancy; Figures in italics indicates the corresponding equating method yields the smallest discrepancy for a test.

7 Figures

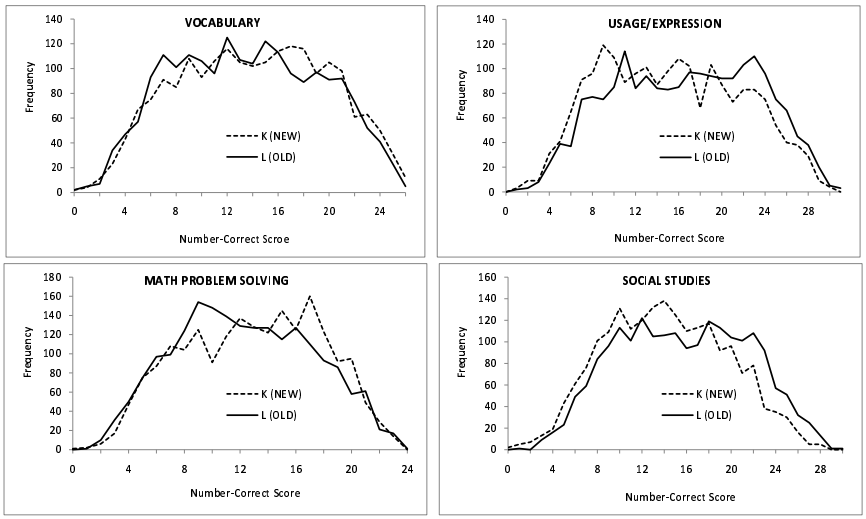

Figure 1: Raw Score Distributions for Both Forms of the Four ITBS Tests

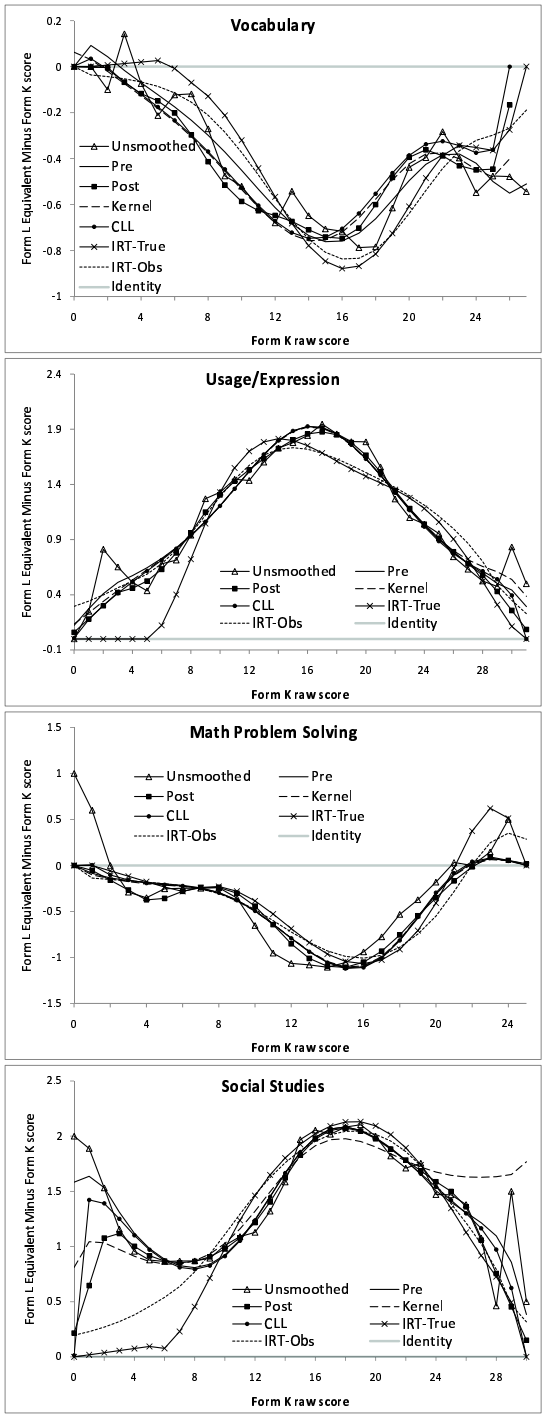

Figure 2: Raw-to-Raw Score Equivalents for the Four ITBS Tests

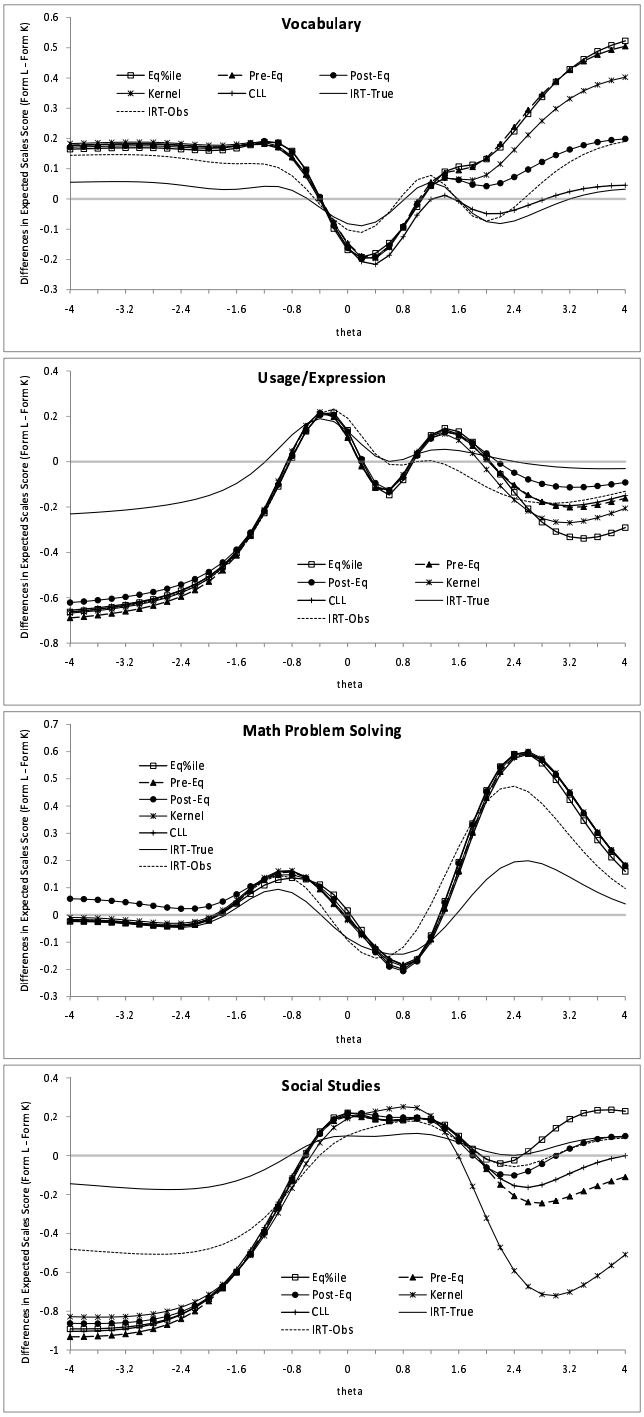

Figure 3: First-Order Equity for Number-Correct Scores

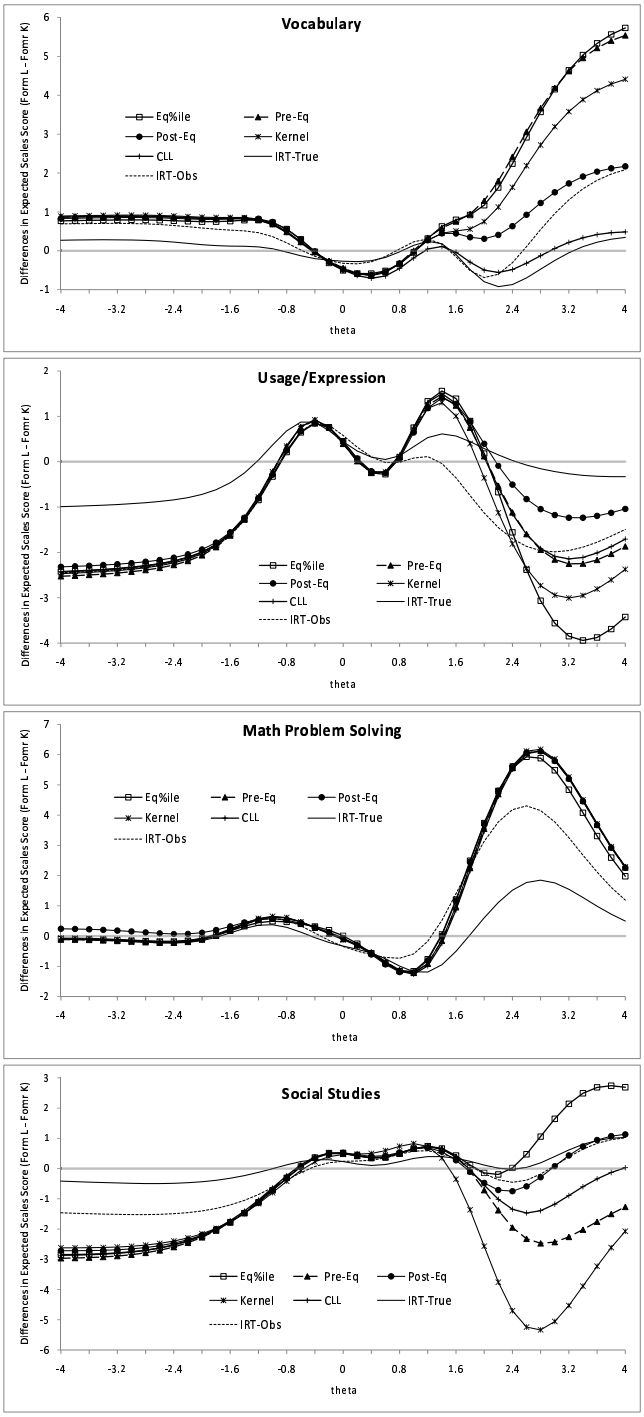

Figure 4: First-Order Equity for Developmental Standard Scores

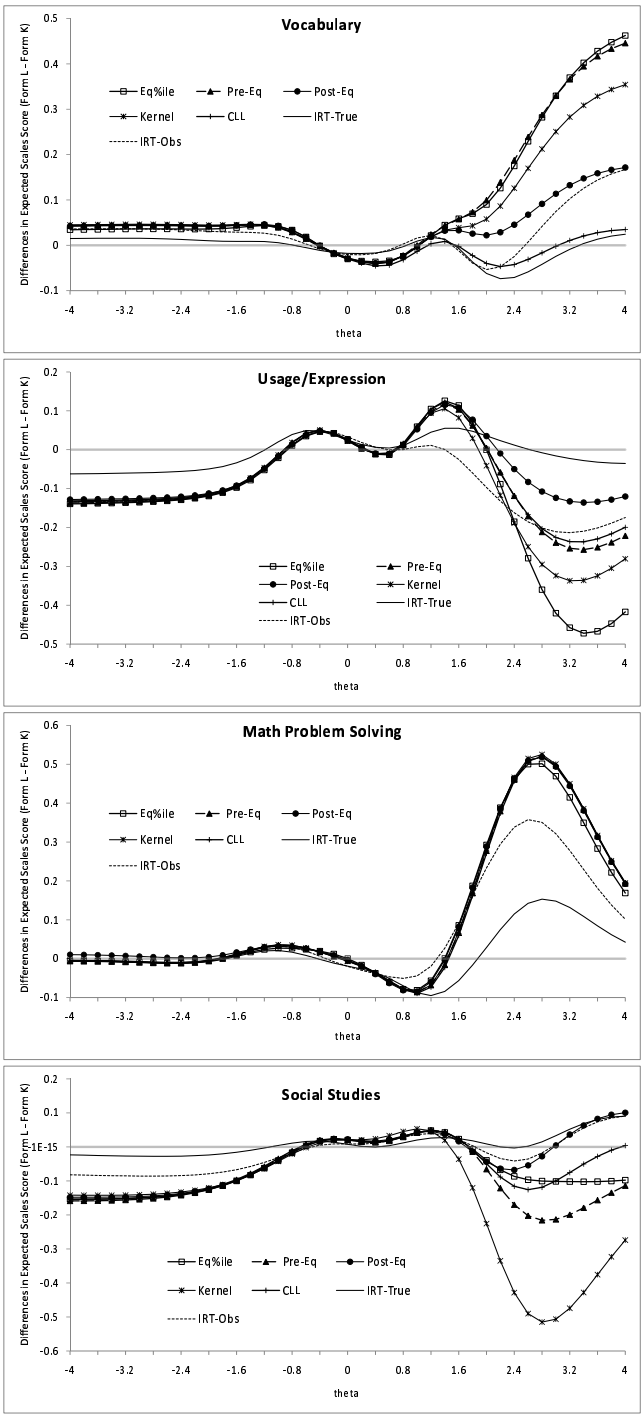

Figure 5: First-Order Equity for Grade Equivalent Scores

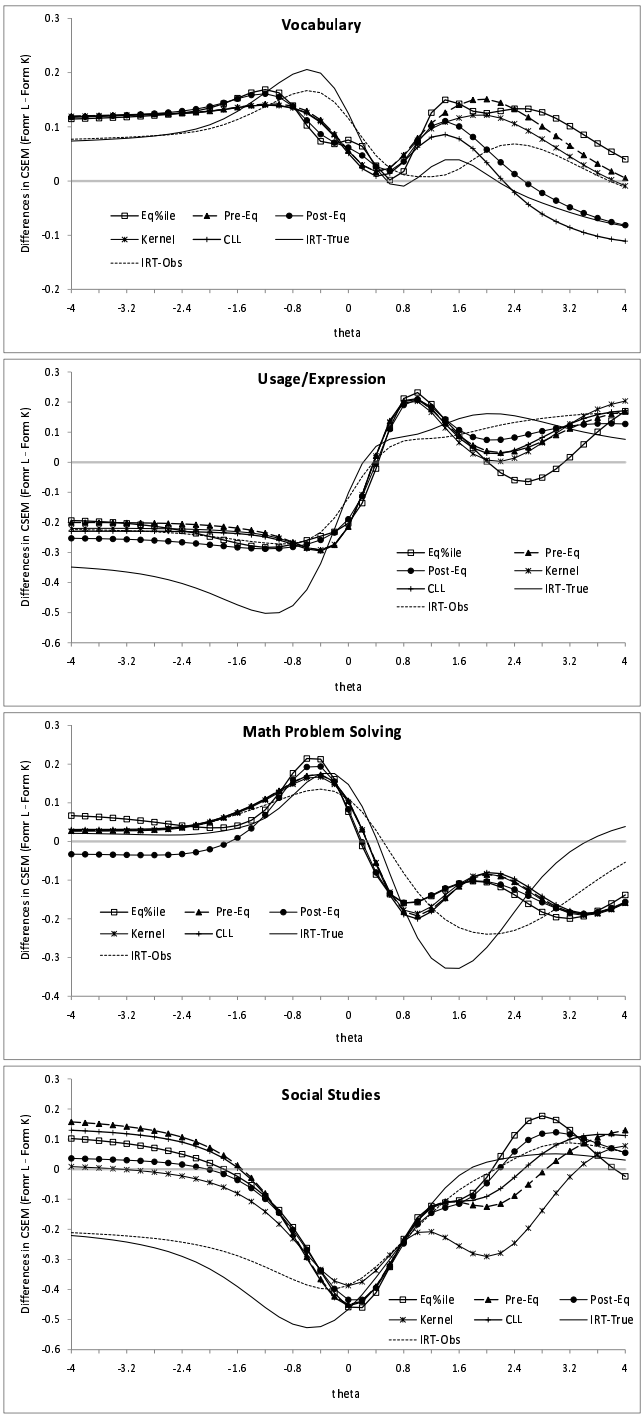

Figure 6: Second-Order Equity for Number-Correct Scores

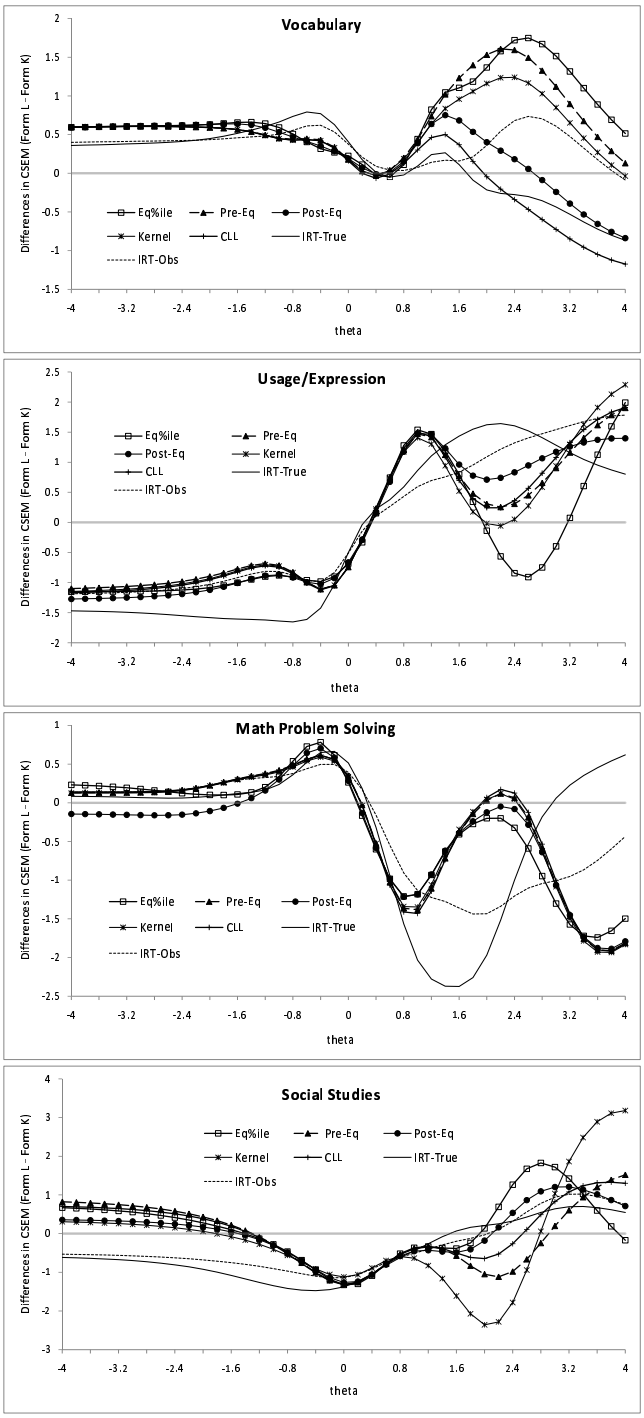

Figure 7: Second-Order Equity for Developmental Standard Scores

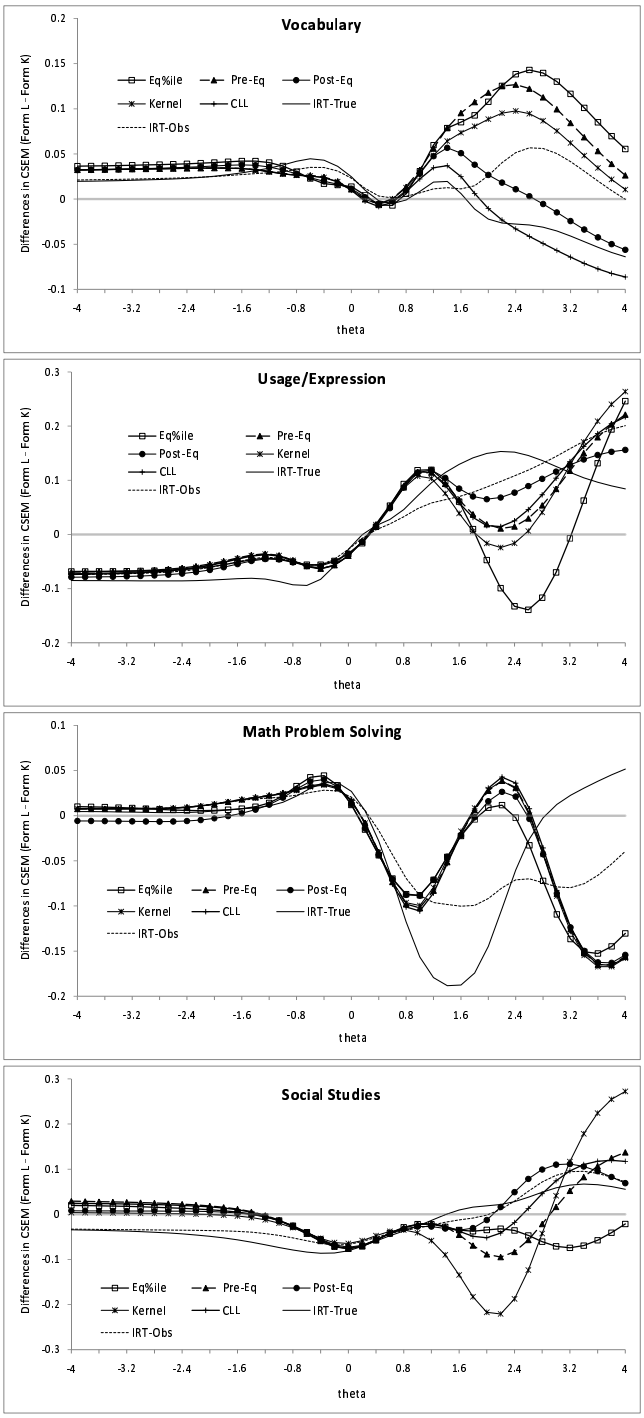

Figure 8: Second-Order Equity for Grade Equivalent Scores