Scale Shrinkage in Vertical Equating

Gregory Camilli, Rutgers, The State University of New Jersey

Kentaro Yamamoto and Ming-mei Wang, Educational Testing Service

As an alternative to equipercentile equating in the area of multilevel achievement test batteries, item response theory (IRT) vertical equating has produced unexpected results. When expanded standard scores were obtained to link the Comprehensive Test of Basic Skills and the California Achievement Test, the variance of test scores diminished both within particular grade levels from fall to spring, and also from lower to upper grade levels. Equipercentile equating, on the other hand, has resulted in increasing variance both within and across grade levels, although the increases are not linear across grade levels. Three potential causes of scale shrinkage are discussed, and a more comprehensive, model-based approach to establishing vertical scales is described. Test data from the National Assessment of Educational Progress were used to estimate the distribution of ability at grades 4, 8, and 12 for several math achievement subtests. For each subtest, the variance of scores increased from grade 4 to grade 8; however, beyond grade 8 the results were not uniform. *Index terms: developmental scores, equating, IRT scaling, maximum likelihood estimation, National Assessment of Educational Progress (NAEP), scale shrinkage, vertical equating.*

With the advent of item response theory (IRT), new techniques for equating tests have become possible. As an alternative to equipercentile equating of multilevel achievement test batteries, however, IRT vertical equating has produced some unexpected results. When development scale scores (Petersen, Kolen, & Hoover, 1989) were obtained to link the Comprehensive Test of Basic Skills (CTBS; CTB, 1984, 1989) and the California Achievement Test, Form E (CTB, 1987), it became evident that the variance of test scores

generally diminished from fall to spring within grade levels and from lower to upper grade levels (Yen, 1985). Expanded standard scores from the Stanford Achievement Test—scaled with the Rasch model—also show this shrinkage pattern (Gardner, Madden, Rudman, Karlsen, Merwin, Callis, & Collins, 1985). In comparison, the methods used to equate levels of the Iowa Tests of Basic Skills (Hieronymus & Hoover, 1986; Petersen, Kolen, & Hoover, 1989) and the CTBS/Form S (CTB, 1974) resulted in increasing variances both within and across grade levels.

The amount of shrinkage for some subtests of the third edition of the CTBS (Form U) is considerable. The standard deviations (SDs) for the Math Computation and Math Concepts subtests shrank by 83% and 58%, respectively, from the fall of grade 2 to the fall of grade 6. In fact, scores on all of the CTBS/U subtests showed shrinkage with the possible exception of Language Mechanics.

In order to illustrate the phenomenon of shrinkage more clearly, the scale medians and SDs for the Math Computations and Math Concepts subtests are given in Figure 1. The general trend includes both rapidly decelerating growth (Figure 1a) and shrinking variability (Figure 1b). Note that variability generally decreases until grade 11, but then shows a moderate increase.

Some researchers who have noted the discrepancy between the CTBS IRT scale and other scales have taken a neutral stance. For example, Yen (1986) argued that it is not possible to determine a scale uniquely; consequently, it is not possible to know which scale reflects educational progress more accurately. The CTBS/U IRT scale and other developmental score scales are simply

APPLIED PSYCHOLOGICAL MEASUREMENT

Vol. 17, No. 4, December 1993, pp. 379-388

© Copyright 1993 Applied Psychological Measurement Inc.
0146-6216/93/040379-10\$1.75

Figure 1
Score Medians and SDs by Grade for CTBS/U
Math Concepts and Math Computations

“different,” although Burket (1984) made the stronger claim that the shrinkage phenomenon does reflect something important about growth in basic skills. In an opposing point of view, Hoover (1984) contended that both shrinkage and decelerating growth of achievement were unsupported empirically and intuitively, and therefore the CTBS/U scale should be viewed with a degree of skepticism. Rather than being a challenge to commonly held beliefs about academic growth, from this perspective the scale shrinkage property is simply an artifact of the equating methodology. Camilli (1987) provided partial support for this hypothesis by showing that the decrease in variance within grade levels is probably the result of measurement error.

The most useful sources of empirical information concerning scale shrinkage are the CTBS/U, CAT/E, and CTBS/4 technical manuals (CTB, 1984, 1987, 1989). Although the scales of other tests also show shrinkage, their technical manuals provide fewer details of item calibration and scaling.

The present paper is primarily concerned with

the apparent shrinkage of developmental score scales, and not with particular tests. Other facets of standardized test batteries are discussed only as they pertain to the shrinkage phenomenon. It is argued that scale shrinkage does not result from the use of IRT per se, but rather particular applications of IRT in which test content and specification, IRT estimation methods, and examinee populations interact.

Scale Shrinkage

Definition

Although scale shrinkage generally is characterized by shrinking variances—within and between grade levels—and rapidly decelerating growth across grade levels (Camilli, 1987), the pattern shows much variation depending on particular grade level, test form, and subtest. As mentioned above, developmental score scales from both the CTBS/U and CAT/E show moderate expansion in later grades. Becker & Forsyth (1992) developed one IRT equating design using standard IRT methodology that showed scale

expansion for the Iowa Test of Educational Development across grades 9–12. Shrinkage is most evident in grades 1–4. However, the amount of shrinkage also varies to some degree by subtest. For example, the Language Mechanics subtest of the CTBS/U shows an irregular pattern of SDs that shrink from grades 2.1 to 3.7, expand from 4.1 to 9.1, shrink again from 9.7 to 10.7, and remain constant from 11.1 to 12.7 (where .1 refers to the fall of a grade and .7 refers to the spring of the grade).

Yen (1985) examined shrinkage in terms of the SDs of scaled item difficulties for the CTBS/U Levels D (grades 2.1–2.7) and J (grades 9.1–12.7). All subtests showed shrinkage except the Reading Vocabulary and Language Mechanics subtests. In terms of scale scores, however, Reading Vocabulary showed a clear pattern of shrinkage, and Language Mechanics showed the irregular pattern noted above. It is preferable to conceptualize shrinkage in terms of the ability scale

rather than the item difficulty scale, because items can be written to have more or less homogeneous difficulties independently of the ability distribution. The most interesting question concerns whether the dispersion of ability level shrinks, not the dispersion of item difficulties. Furthermore, because different segments of the IRT score scale behave differently, statements regarding shrinkage must be qualified to some extent with regard to subject matter and grade.

Although shrinkage first became evident on examination of the CTBS/U subtests, less pronounced shrinkage also occurred for some CTBS/4 subtests. For example, the SDs for the Math Concepts subtest contracted approximately 11% (CTBS/U) and 13% (CTBS/4) from grade 2.7 to 3.1. Over the same period, the SDs for Math Computation contracted 23% (CTBS/U) and 8% (CTBS/4). Table 1 shows the actual patterns of shrinkage for the Math Concepts subtest in SD units for the CTBS/U, CTBS/4, and CAT/E. The

Table 1
Score Medians (in z Units Relative to Fall Grade 2) and SDs
(in q Units Relative to Fall Grade 2 SDs) for the CTBS/U,
CTBS/4, and CAT/E Math Concepts and Applications

Test, Level, and Grade	Fall		Spring	
	Median	SD	Median	SD
CTBS/U				
Levels B & C, Grade 1	–1.57	1.32	–.46	1.12
Level D, Grade 2	0.00	1.00	.59	.79
Level E, Grade 3	.79	.70	1.06	.54
Level F, Grade 4	1.19	.54	1.37	.48
Level G, Grade 5	1.47	.46	1.63	.44
Level H, Grade 6	1.70	.42	1.80	.39
CTBS/4				
Levels 10 & 11, Grade 1	–1.62	1.19	–.18	1.16
Level 12, Grade 2	0.00	1.00	.59	1.13
Level 13, Grade 3	1.28	.98	1.82	.94
Level 14, Grade 4	1.82	.87	2.18	.82
Level 15, Grade 5	2.21	.80	2.52	.76
Level 16, Grade 6	2.56	.72	2.82	.72
CAT/E				
Level 11, Grade 1	–1.81	1.22	–.30	1.19
Level 12, Grade 2	0.00	1.00	1.24	1.14
Level 13, Grade 3	1.32	.95	1.99	.99
Level 14, Grade 4	1.97	.89	2.37	.99
Level 15, Grade 5	2.33	.77	2.68	.82
Level 16, Grade 6	2.69	.76	2.98	.84

median scale score and SDs (τ_{ij} and σ_{ij}) for each grade i and month j were standardized as follows:

$$z_{ij} = \frac{\tau_{ij} - \tau_{21}}{\sigma_{21}} \quad (1)$$

and

$$q_{ij} = \frac{\sigma_{ij}}{\sigma_{21}}. \quad (2)$$

Note that the CAT/E, which used nearly the same equating methodology as the CTBS/4, showed approximately the same SDs (within and between grades) in q units. This can be seen by comparing the q indexes in Table 1 for the three different tests. More growth was observed for the CTBS/4 and CAT/E than for the CTBS/U. Table 1 shows that the pattern of shrinkage within grades was quite different for the CTBS/U and CTBS/4. For example, for the CTBS/U Math Concepts, the grade 2 SD (in q units) decreased by approximately 21% (1.00 to .79) from fall to spring; for the CTBS/4, the SD increased by approximately 13%. In general, the CTBS/4 showed some shrinkage within grades, but less than the CTBS/U. It appears, then, that references to "scale shrinkage" also should be qualified with respect to test and test edition. Variations in patterns of shrinkage apparently result from changes in test development, equating, and scaling.

Potential Causes

Scale shrinkage may originate from many sources. However, only two have been proposed and investigated to some degree in the research literature: statistical artifact and multidimensionality. Below, measurement error and restriction of range are considered artifactual causes of shrinkage. These potential problems are associated with the IRT method of joint maximum likelihood (JML) analysis, which was the method used to scale the CTBS/U, CTBS/4, and the CAT/E. Multidimensionality then is examined as a cause of scale shrinkage.

Measurement error. Camilli (1987) showed in a simulation study that the pattern of shrinkage

within grades could be explained by measurement error in estimated trait levels ($\hat{\theta}$ s). More specifically, an IRT $\hat{\theta}$ from a three-parameter IRT model near a floor (the lowest possible score on a test) has a large standard error of measurement (s_e), and a $\hat{\theta}$ near a ceiling (the highest possible score) has a small s_e (Camilli, 1987, Figure 4). Note that s_e is calculated directly from the $\hat{\theta}$ s, and does not refer to the expected value of s_e , which has a U-shaped curve (Lord, 1980).

Therefore, with a three-parameter IRT model, groups that are lower on the θ continuum will have observed (i.e., estimated) score distributions with larger amounts of measurement error. Lower scoring individuals also may attempt fewer items, resulting in a higher amount of measurement error at lower θ levels (Mislevy, personal communication). Therefore, fall norming groups might be expected to have lower test reliabilities than spring norming groups.

The pattern of the reliabilities of some CTBS subtests from lower to higher grades confirms this expectation. For the CTBS/4, Table 2 gives the

Table 2
KR20 Reliabilities for Math Concepts
Based on Number-Correct Scoring
for Levels 11-16 on the CTBS/4

Level	Grade	KR20	
		Fall	Spring
11	1	.77	.83
12	2	.77	.85
13	3	.85	.86
14	4	.83	.85
15	5	.91	.93
16	6	.91	.94

estimated Kuder-Richardson formula 20 (KR20) values for fall and spring (CTB, 1989). These coefficients range from .77 in fall first grade to .94 in spring sixth grade. Reliabilities for the ITBS (Hieronymus & Hoover, 1986) also show this pattern in grades 2-6 for Math Concepts (.77-.89) and Math Problem Solving (.79-.90). This indicates that measurement error contributes more to observed score variances at lower grade levels in the fall, and suggests that observed score SDs

decrease to some degree because the error component diminishes. Score distributions based on the ITBS scaling test were adjusted so that they would reflect true score variances (see Petersen, Kolen, & Hoover, 1989, p. 232).

Restriction of range. The $\hat{\theta}$ for some item response patterns, including all correct and all incorrect, cannot be obtained by the JML method—the numerical algorithm does not converge. Therefore, arbitrary lowest and highest scores must be selected. In the presence of strong ceiling and floor effects, this procedure will produce spikes at the lower and upper end of the frequency distribution of $\hat{\theta}$. Obviously, the usefulness of the variance estimate of such a distribution is severely limited. However, it appears that calculations for scale means, medians, and variances for the CTBS/U included such scores. Given a high proportion of “substituted” values, distributions of observed scores that are distorted are expected.

There is evidence of floor and ceiling effects on the CTBS/U. Camilli (1987) noted that the average item proportion correct for the CTBS/U Math Concepts and Math Computations subtests were approximately .8 for second graders in the spring (Level D). Furthermore, the first estimable scale scores below the maximum score for these subtests were at the 94th (Math Concepts) and 82nd (Math Computations) percentiles. Approximately 18% of the examinees were close to the ceiling on Math Computations in the spring. By the same logic, approximately 16% of second graders were fairly close to the floor in the fall. [This information is available in the CTBS Norms Book, Grades K-3 (CTB, 1983).] To some degree, floor and ceiling effects may present problems for all multilevel achievement batteries, especially at lower grade levels. For example, for the ITBS approximately 18% of fourth graders score at or below chance on Math Problem Solving in the fall (Hieronymus & Hoover, 1986).

The statistics for judging the severity of ceiling and floor effects are given in IRT scale scores (and IRT-derived number-correct scores) for the CTBS/U and in the number-correct metric for the

ITBS. Thus, these statistics are not strictly comparable. In fact, the IRT scale scores may be more informative because chance scores are more precisely defined, and estimable scale scores implicitly give greater weight to more discriminating items.

Multidimensionality. This explanation was induced from the observation that the IRT item discrimination, a , and item difficulty, b , parameters tend to be positively correlated for some achievement tests (Yen, 1985). This suggests a secondary construct that becomes more influential to performance on later items on the test. Accordingly, scale shrinkage is a function of item dimensionality, and

The variance of the estimated item difficulties and traits will decrease (i.e., the scale will “shrink”) as the number of traits influencing the items increases if the correlation between the items’ difficulties on the underlying dimension is less than the correlation between the underlying traits (Yen, 1986, p. 321).

By means of a simulation study, Yen (1985) showed that shrinkage is predictable when a unidimensional scaling model is applied to a multidimensional test. To show this, a formula was derived (Yen, Equation 22, p. 403) for the variance of unidimensional estimates of b for a test that has J actual dimensions. For the case of $J = 1$, the variance of the b s (across items) with unit normal θ is defined as σ_b^2 .

For $J = 2$, the result can be simplified with three assumptions. First, assume that b_1 and b_2 are uncorrelated across items, and that θ_1 and θ_2 are unit normal. Second, assume that both b_1 and b_2 have variances equal to σ_b^2 across items. Third, assume that a_1 and a_2 are constant across items. The variance of the b s estimated with a unidimensional model for $J = 2$ then can be expressed as

$$V_b = \frac{\sigma_b^2(a_1^2 + a_2^2)}{a_1^2 + 2a_1a_2\rho_{12} + a_2^2}. \quad (3)$$

These predictions were tested by Yen in a simulation study in which data were simulated for a

unidimensional test and a multidimensional test according to the simplifying assumptions made above (the correlation between b_1 and b_2 was .03). Then, the estimated ratios of variances of item difficulties (obtained with LOGIST) were compared to the expected ratio. The results were very close to expectation: The variance of the multidimensional test ($J = 2$) shrank relative to the variance of the unidimensional test ($J = 1$). (Yen's simulated test was very closely approximated by the unidimensional model.) However, note that the ratio of variances for the two tests ($J = 1, 2$) can be expressed very simply by

$$\frac{V_b}{\sigma_b^2} = \frac{a_1^2 + a_2^2}{a_1^2 + 2a_1a_2\rho_{12} + a_2^2}. \quad (4)$$

This ratio does not depend on the variance of the b s because σ_b^2 has canceled out of the equation.

The cause of this apparent shrinkage is that the unidimensional composite, created from two correlated traits, has greater true variance than the single trait level for $J = 1$. (Also note that the ratio in Equation 4 would be 1.0 if the correlation between traits were 0.) Hence, the trait level estimated from the two-dimensional test is more reliable. In fact, Yen (1985) noted that the items of the two-dimensional test had higher a s and therefore yielded estimated θ s with smaller variances because they "have smaller standard errors of measurement" (p. 405). Multidimensionality produced shrinkage only to the degree that the multidimensional test was more reliable, that is, contained less error. (In practice, however, multidimensionality need not increase the reliability of a test.)

The NAEP Scaling Model

An alternative strategy for scaling multilevel achievement tests is to use IRT methods to estimate simultaneously item parameters and the characteristics of the θ distributions. Unlike methods used for the CTBS/U, the use of estimated θ s in number-correct formulas for mean and variance is bypassed. Thus, scales established with this method may be less sensitive to measurement error and restriction of range. Camilli

(1987) showed that more accurate estimates of true θ means and variances were obtained with this approach than those obtained by using estimated θ s in standard formulas. Software exists for this purpose (Mislevy & Bock, 1982) and has been applied in the NAEP. A summary description of the IRT model is provided below. More detailed information is given in Johnson & Allen (1992).

Description of the Model

Different θ distributions (i.e., population distributions) are hypothesized for each group, and an IRT function that relates the probability of a correct response on each item to a set of item parameters is described. Once this formulation—called the "model"—is achieved, the θ and item parameters must be estimated.

The vertical equating or scaling problem is thus translated to one of defining and estimating population θ distributions. When examinees are from known subpopulations such as 4th, 8th, and 12th graders for the NAEP test data, it is appropriate to define separate θ distributions for each grade. Accordingly, the mathematical model is motivated by the fundamental Bayesian idea that the best estimate of the probability of a correct response to an item is derived from a three-parameter IRT model combined with prior information concerning the probability of encountering an examinee from a particular grade at a particular θ level.

In the scaling process, both the IRT item parameters and the prior information are estimated in stages. The prior information for a particular grade is defined as the grade's frequency (population) distribution of ability, which can either be assumed to be normally or irregularly shaped. This frequency distribution is known as the "population distribution." At the end of the estimation stages, the θ distribution is combined with information in the groups' item response patterns to obtain the posterior distribution. If the θ distribution was assumed to be normal, but was actually skewed, information in the item response patterns would update the θ

distribution and the posterior distribution would correctly reflect the degree of skewness. When many examinees respond to a large number of items, the prior θ distribution plays a minor role; in this case, the estimated θ s are sufficient for determining the moments of score distributions. However, criteria for determining what constitutes a sufficient number of examinees and items have not been developed.

Scale Linking in NAEP

One of the main objectives of the NAEP is to assess subject-matter proficiency in three grade/age cohorts (grade 4/age 9, grade 8/age 13, and grade 12/age 17) within an assessment year. Analyses pertaining to this objective are referred to here as cross-sectional analyses, and they are intended to provide comparisons of the performances of subpopulations that differ in background variables such as age, grade, sex, ethnicity, and region of the country.

The cross-sectional analyses of the NAEP are based on data that were collected according to a focused, balanced incomplete block item sampling design. Each examinee receives a test booklet containing items from a single subject area. For example, the 1990 mathematics cross-sectional assessment used seven blocks of items for each grade/age cohort. There were five content-area subscales, and each of the seven blocks contained items from these content areas. For each cohort, the blocks were assembled into seven different test booklets and each booklet was administered to a random subsample of the relevant cohort by "spiraling" test booklets within each administration site (i.e., different test booklets are combined in an alternating order so that equivalent samples of examinees receive each test booklet).

In the NAEP cross-sectional assessments, the combined item pools for three cohorts contain two types of items. The first, referred to here as *grade-age only items*, are administered to a single grade/age cohort. The second type, referred to here as *linking items* are administered to at least two of the three cohorts. For example, the Num-

bers and Operations scale of the 1990 mathematics cross-sectional item pool contained 73 grade-age only items (33 at grade 4/age 9, 21 at grade 8/age 13, and 20 at grade 12/age 17), and 25 linking items (8 items common to grade 4/age 9 and grade 8/age 13, 6 items common to grade 8/age 13 and grade 12/age 17, and 11 items common to all three cohorts). The 1990 NAEP technical report (Johnson & Allen, 1992) contains more detailed information on the instrument.

The purposes of the cross-sectional analyses require that results (within and across cohorts) be reported in terms of a common scale. To achieve this, scales for a given subject area are derived by using a concurrent calibration scheme in the item parameter estimation phase of the analysis. The concurrent method capitalizes on both the common-item structure (due to the inclusion of the linking items) and common-population structure (due to the spiraling of test booklets within grade/age cohort) inherent in the overall design. Parameters are estimated simultaneously for the items from all three grade/age cohorts. Separate θ distributions are concurrently estimated for each grade/age cohort. At the initial stage of parameter estimation, identical standard normal distributions were used for all three cohorts. Subsequent EM steps (Mislevy & Bock, 1982) estimated three distinct θ distributions concurrently with item parameters.

Item parameters were estimated with a calibration sample of nearly equal size from each grade/age cohort. The calibration sample for the 1990 NAEP analysis included a subsample of 12,905 of the 25,572 examinees. The number of examinees responding to a particular item ranged from 963 to 5,740 with all but one item having greater than 1,000 responses. The calibration sample sizes were balanced to eliminate the possibility of a particular sample dominating the item parameter estimation—the fit of the model parameters was maximized with respect to the composite posterior distribution of equally weighted multiple grade/age samples. Using this method, only one set of item parameters for the linking items was derived. Compared to the

parallel common-item linking approach, this variant of the concurrent estimation method directly incorporates information about both the shape and the location of the θ distribution for each grade/age cohort. [For more information on the NAEP procedures see the summer 1992 special issues of the *Journal of Educational Measurement* (Volume 29) and *Journal of Educational Statistics* (Volume 17), which cover numerous topics including the data collection design, estimating item and population characteristics, scale linking, differential item functioning, dimensionality, and content area frameworks.]

Results

Table 3 presents the means and SDs of the NAEP math subscale scores by grade. There is no evidence of decreasing variance as grade level increases from grade 4 to grade 8 for any of the five subscales; for the four applicable scales, SDs increased by 4%–29%. However, for three of the five subscales the SDs of the grade 12 score distributions were less than that of grade 8. The other two subscales had SDs that increased by 17% (Geometry) and 1% (Algebra and Function). This is the opposite pattern of shrinkage exhibited by the CTBS tests.

Discussion

The IRT scale for the NAEP math subtests generally showed scale expansion from grade 4 to 8, and some shrinkage from grade 8 to 12. However, these results must be carefully qualified. It cannot be said that this scale is "correct" or "equal interval," but rather that this obtained IRT scale is consistent with a measurement model

that recognizes growth and diversity. Scale shrinkage is not an inevitable property associated with IRT scaling. The NAEP scalings used here did not show shrinkage from grade 4 to grade 8. Thus, the CTBS results should not be used to support substantive arguments about growth. The NAEP results are somewhat more defensible regarding scale characteristics. Substantial item context effects, multidimensionality, differential reliability, and other sources of potential non-equivalence between grade/age cohort scales would be expected to alter the item response functions of the linking items. However, Beaton & Johnson (1992) reported that 96% of the item response functions for the NAEP linking items showed evidence of invariance across the grade/age cohorts. But the NAEP procedure has not resolved completely the issue of scaling (Linn & Dunbar, 1992).

In the vertical equating situation, the tests to be equated are intended to vary in difficulty, and the equating process is complicated by the fact that different skills are brought to bear on problem solving in different grades. The *Standards for Educational and Psychological Testing* recommend use of the term "comparable" scores rather than "equated" scores in this situation (American Psychological Association, 1985). It should not be concluded, however, from the present study that multidimensionality has no effect on the vertical equating process. On logical grounds, there is good reason for the concern that test dimensionality may affect multilevel test equating.

It cannot definitively be determined whether developmental scales should show expansion or

Table 3
NAEP Subscale Score Means and SDs for Grades 4, 8, and 12

Subscale	Grade 4		Grade 8		Grade 12	
	Mean	SD	Mean	SD	Mean	SD
1. Number and Operations	212.5	33.5	268.7	34.8	294.3	33.4
2. Measurement	221.7	33.5	261.3	43.3	294.2	37.2
3. Geometry	217.0	30.9	261.7	35.5	296.2	41.5
4. Data Analysis, Statistics and Probability	—	—	266.1	41.2	295.2	36.3
5. Algebra and Function	215.5	31.5	264.2	36.8	296.7	37.1

contraction. The criteria for determining useful vertical scales constitute a controversial topic for debate and research. However, some measurement problems were addressed holistically with the NAEP IRT approach: A plausible measurement model was selected to describe test responses from multilevel test batteries, and a developmental score scale was defined in terms of the estimated posterior distribution of θ . Rival scales, from this perspective, are proposed as rival models that can be evaluated logically for utility and empirically for fit. Even within this framework, however, it is generally acknowledged that IRT methods do not yield interval scales. Any number of monotonic scale transformations fit the observed data equally well, and impressions of group variability can change with the transformation selected. Therefore, methods of enhancing scale interpretation are being developed that do not depend on interval-level properties (Beaton & Johnson, 1992).

It is noted with some irony that in the NAEP program item parameters are now estimated separately for each age/grade cohort (i.e., within age), and then linked to create a cross-group scale. Early results indicate that this new procedure gives virtually the same results as those from the former cross-age calibration (Haertel, 1991). This suggests that linking procedures are less likely the source of notable cases of scale shrinkage than JML estimation of ability; however, only a comparative study on a set of real multilevel achievement data has the potential to validate this claim definitively. Such a study might not be considered desirable by critics of cross-age scaling. In fact, the NAEP Technical Review Panel found that the interpretations given to NAEP cross-age quantitative comparisons are often problematic (Haertel, 1991).

The present study had several limitations. First, arguments for particular scales or transformations of scales are difficult to support empirically with real achievement test data, and simulation studies are probably not very useful for this purpose. Second, the gap between grades 4 and 8 (and 8 and 12) is quite large, which might

render scaling procedures sensitive to undesirable or incidental properties of the linking items. Test developers are in a much better position to link tests, having to link only adjacent grades. It would be quite helpful to repeat the analyses presented here on such test data. Finally, it has been suggested, but not proven, that the cross-age NAEP scaling method is more robust to measurement problems than standard IRT methods. Studies have not yet been performed to test these models on the same set of data. Given the relatively few items that any one examinee takes in the NAEP assessment, individual scores are highly unreliable. It is therefore not practical to use standard JML methods in which estimated θ s are used to calculate scale transformation coefficients and moments of θ distributions.

References

- American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington DC: Author.
- Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement*, 29, 163-175.
- Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, 29, 341-354.
- Burket, G. R. (1984). Reply to Hoover. *Educational Measurement: Issues and Practices*, 3, 15-16.
- Camilli, G. (1987). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics*, 13, 227-241.
- CTB. (1974). *Comprehensive Test of Basic Skills, Form S* (Technical Bulletin 1). Monterey CA: CTB/McGraw-Hill.
- CTB. (1983). *Forms U and V Norms Book, Grades K-3*. Monterey CA: CTB/McGraw-Hill.
- CTB. (1984). *Comprehensive Tests of Basic Skills, Forms U and V* (Technical Report). Monterey CA: CTB/McGraw-Hill.
- CTB. (1987). *California Achievement Tests, Forms E and F* (Technical Bulletin 2). Monterey CA: CTB/McGraw-Hill.
- CTB. (1989). *Comprehensive Tests of Basic Skills, Fourth Edition* (Technical Bulletin 1). Monterey CA: CTB/Macmillan/McGraw-Hill.
- Gardner, E. F., Madden, R., Rudman, H. C., Karl-

- sen, B., Merwin, J. C., Callis, R., & Collins, C. S. (1985). *Stanford Achievement Test Series* (Technical Data Report). Harcourt Brace Jovanovich.
- Haertel, E. H. (1991). *Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress*. Unpublished manuscript, Stanford University.
- Hieronimus, A. N., & Hoover, H. D. (1986). *Iowa Tests of Basic Skills: Manual for school administrators*. Chicago IL: Riverside.
- Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GEs. *Educational Measurement: Issues and Practices*, 3, 8-14.
- Johnson, E. G., & Allen, N. L. (1992). *The 1990 NAEP technical report* (Rep. No. 21-TR-20). Washington D.C.: National Center for Education Statistics.
- Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29, 177-194.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
- Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Maximum likelihood item analysis and test scoring with logistic models*. Mooresville IN: Scientific Software.
- 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.
- Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399-410.
- Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23, 299-325.

Acknowledgments

The author acknowledges two anonymous reviewers and Elizabeth Celentano for their helpful comments.

Author's Address

Send requests for reprints or further information to Gregory Camilli, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick NJ 08903, U.S.A. Internet: camilli@zodiac.rutgers.edu.