*Journal of Educational Measurement*
*Winter 1988, Vol. 25, No. 4, pp. 275–285*

# Item Pool Maintenance in the Presence of Item Parameter Drift

**R. Darrell Bock**
*University of Chicago*
**Eiji Muraki**
*NORC—National Opinion Research Center*
and
**Will Pfeiffenberger**
*Educational Testing Service*

*Differential linear drift of item location parameters over a 10-year period is demonstrated in data from the College Board Physics Achievement Test. The relative direction of drift is associated with the content of the items and reflects changing emphasis in the physics curricula of American secondary schools. No evidence of drift of discriminating power parameters was found. Statistical procedures for detecting, estimating, and accounting for item parameter drift in item pools for long-term testing programs are proposed.*

Although schemes for maintaining a test scale by classical equating methods have been in use for many years (Angoff, 1971), no similar methodology employing item response theory (IRT) has yet been proposed. The main obstacle appears to be the problem of *item parameter drift,* that is, differential change in item parameter values over time (Goldstein, 1983). It has not been clear how to take into account possible increasing or decreasing difficulty over time of some items in an IRT scale relative to others that are unchanging or changing in the opposite direction. Such effects might be expected as a result of educational, technological, or cultural change during the useful life of the scale. An example is an item from a vocational aptitude test that asks for the correct SAE number of motor oil for winter use. Knowledge of SAE numbers for the viscosity of motor oil has become superfluous since the introduction of multiple-viscosity oils.

In the present paper, a method for maintaining and updating an IRT scale over a period of time, while accounting for item parameter drift, is proposed. The method involves the fitting of what we call a "time-dependent" IRT model to data obtained from the operational use of the test. Although based on IRT and formulated at the item level rather than the test level, the model is employed in the same spirit as classical equating of successive forms of a regularly updated test. It does not depend on calibration of the items in data from a particular year, but attempts to smooth the parameter estimates over a number of years after eliminating cohort main effects.

Item parameter drift is only one of the forms of differential item functioning (DIF) that can affect objective tests. Similar effects, referred to as *item bias,* can be observed when the respondents are classified with respect to subgroup membership rather than time of testing. By placing a linear constraint on the

coefficients of the model, the procedure proposed here for item parameter drift can be applied in item bias studies. In addition, a provision for calibrating variant items by extension from operational items is included also. This enables the IRT scale to be maintained without the need for separate equating studies.

## The Item Replacement Scheme

Regular updating of items in an IRT scale is desirable both to keep the content abreast of changing education and experience in the population of respondents and to protect the test from overexposure or compromise. We assume for this purpose periodic, typically annual, retirement and replacement of a certain fraction of items in each distinct IRT scale within a test or test battery. In addition, we suppose that a number of so-called *variant* or provisional items are included with the scaled items during operational testing. These variant items are not used in the scoring of the operational test, but are provisionally calibrated by extension from the operational items in order to guide selection among them in the next update of the scale. The statistical basis for this type of provisional calibration is discussed in the section on item parameter drift in the College Board Physics Achievement Test.

A 25% item replacement scheme suitable for a time-dependent IRT model is illustrated in Figure 1. The scale is assumed to consist of 16 items, 4 of which are replaced each year, leaving none of the original items in the 5th year. In addition, 6 variant items are included at the end of the scale each year, from which 4 will be chosen as replacements in the following year.

When the system is in its steady state, all accumulated data are employed in the fitting of the time-dependent IRT model. Four years of data contribute to the

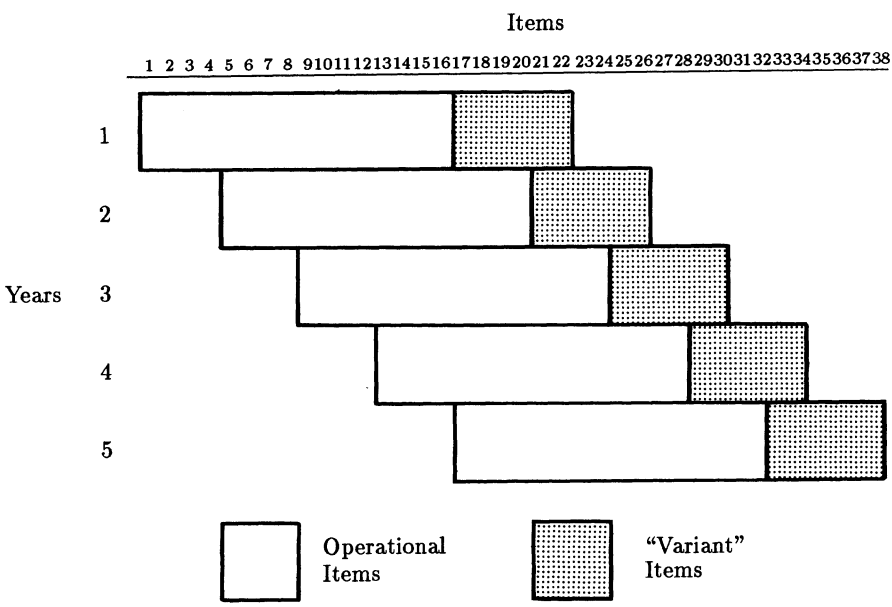


FIGURE 1. *An item replacement scheme*

estimation of smoothed item parameters in the earliest entered 25% of the items, 3 years contribute to the second earliest, 2 in the third, and 1 in the new items for the current year. Because Bayes modal estimation of the trend coefficients is employed, all items can be fitted, but the new items receive default trends with zero slopes.

## Assumptions of a Time-Dependent IRT Model

Provisionally, we base the model on the assumptions detailed in this section. In the next section, we present empirical evidence of the reasonableness of some of these assumptions, which are as follows:

1. The response functions of all items can be described by a logistic model with lower asymptote (guessing), slope (discrimination), and location (threshold) parameters. The procedure could be extended to the 4-parameter model of Thissen and Bock (1986) for nonmonotonic response functions, but we have not done so on the grounds that such items are relatively uncommon.

2. The proficiency in the achievement areas is distributed with arbitrary mean and variance in successive samples from the population of respondents. In the present context, we will refer to these samples as *year-groups* or just *groups,* but they can also be the demographic groups of item-bias studies.

3. Drift is confined to the location parameter. Neither the asymptote nor slope parameter is assumed to drift during the lifetime of the item. Undoubtedly it is true that the validity of an item is also impaired as it becomes obsolete and unfamiliar to the respondents, but long before that its increasing difficulty will be apparent and make it a candidate for retirement from the scale. The guessing parameter could also change if test-taking strategies change, but we will assume such changes are small as long as the test instructions remain the same.

4. Only the item $\times$ group interaction is considered drift; average drift for all items is considered a change in the population of potential examinees and is absorbed in the estimate of the year-group mean. There is obviously no way, within the test data themselves, to separate overall change in item difficulty from overall change of ability in the population. This definition of item drift is consistent with that of differential item bias, in which an overall difference in the difficulty of the items is considered "adverse impact" of the test as a whole and thus has no implication for scalability.

5. For purposes of the drift model, the item $\times$ group interactions can be described by a low degree polynomial in time. In item bias studies with crossed or nested structures on the demographic groups, a general linear model could be assumed.

According to these assumptions, the model for the probability of a correct response from a person of ability, $\theta_k$, randomly selected from the population corresponding to group $k$, and responding to item $j$, is

$$P(x_{jk} = 1 \mid \theta) = g_{jk} + (1 - g_{jk})\Psi[z_{jk}(\theta)],$$

where $g_{jk}$ is the lower asymptote or "guessing" parameter, $\Psi$ is the logistic response function, and $z_{jk}(\theta)$ is a logistic deviate. This model may be specialized by choice of $z_{jk}(\theta)$ as follows.

1. Location, slopes, and asymptotes constant over groups:

$$\begin{aligned} z_{jk}(\theta) &= a_j\,(\theta - b_j) \\ &= c_j + a_j\theta, \end{aligned} \tag{1}$$

where $a_j$ is the item slope, $b_j$ is the item location or "difficulty," and $c_j = -a_j b_j$ is the item intercept.

2. Linear location drift over time (or quantitatively structured groups):

$$\begin{aligned} z_{jk}(\theta) &= a_j(\theta - b_j - \delta_j t_k) \\ &= c_j + a_j\theta + d_j t_k, \end{aligned} \tag{2}$$

where $t_k$ is time at occasion $k$ measured from an arbitrary origin, and

$$\sum_{j=1}^{n} \delta_j = 0. \tag{3}$$

3. Quadratic location drift over time (or quantitatively structured groups):

$$\begin{aligned} z_{jk}(\theta) &= a_j\,(\theta - b_j - \delta_{1j} t_k - \delta_{2j} t_k^2) \\ &= c_j + a_j\theta + d_{1j} t_k + d_{2j} t_k^2, \end{aligned} \tag{4}$$

where

$$\sum_{j=1}^{n} \delta_{1j} = 0,$$

and $\delta_{2j}$ is unconstrained.

4. Bias model: slopes and asymptotes constant over groups; locations unrestricted except for a linear constraint. This case could be regarded as a polynomial trend model in which the degree is one less than the number of time points, but it is more natural to consider it a bias model:

$$\begin{aligned} z_{jk}(\theta) &= a_j(\theta - b_{jk}) \\ &= c_{jk} + a_j\theta, \end{aligned} \tag{5}$$

where the threshold estimates for group $k$ are adjusted so that their weighted mean, weighted inversely as their squared standard errors of estimates, is equal to the weighted mean for a reference group; that is

$$\frac{\sum_{j}^{n} w_{jk} b_{jk}}{\sum_{j}^{n} w_{jk}} = \frac{\sum_{j}^{n} w_{j0} b_{j0}}{\sum_{j}^{n} w_{jo}}, \tag{6}$$

and

$$w_{jk} = [\mathrm{SE}(b_{jk})]^{-2}, \tag{7}$$

where $\mathrm{SE}(b_{jk})$ is a standard error obtained from the Fisher-scoring solution described in the following section. Note that the item bias model is one in which the item slopes and asymptotes are assumed constant over groups, but the locations are unrestricted apart from Equation 6.

5. Constant asymptotes only:

$$P(x_{jk} = 1 \mid \theta) = g_j + (1 - g_j)\Psi[z_{jk}(\theta)],$$

where $z_{jk}(\theta) = a_{jk}(\theta - b_{jk})$.

## Estimation

The parameters of these models, jointly with the means and variances of the groups, are estimated by the marginal maximum likelihood (MML) method, implemented by the EM algorithm and Fisher-scoring iterations (see Bock, in press; Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). The MML equations are evaluated numerically using Gauss-Hermite quadrature on 10 quadrature points. Bayes (stochastic) constraints are placed on all parameters: Normal priors are assumed for the intercept and trend parameters, lognormal for the slopes, and a beta prior for the guessing parameter (see Mislevy, 1986). Empirical prior distributions of ability, represented as a discrete distribution on the quadrature points, are employed within groups. The estimated means and standard deviations for the groups are computed from these discrete distributions.

Priors on the intercepts are very mild and serve only to keep the item locations approximately in the plus-or-minus four sigma interval. The priors on the trend parameters are also mild and are introduced primarily to permit the procedure to work for an item that appears with as few as one time point. For the item slopes, priors are necessary to prevent Heywood cases when the sample size is only moderate relative to the number of items. Typically, only a few items are affected by the constraints on the item slopes. The priors on the asymptote parameters are stronger and constrain these parameters to lie between zero and approximately one-half. These constraints are required to assure admissible values when the item is so easy that the lower asymptote is not otherwise estimable.

## Evidence for Item Parameter Drift in the College Board Physics Achievement Test

For an empirical test of the assumptions listed in the section on a time-dependent IRT model, we are fortunate in having available, through the kind cooperation of Linda Cook, Educational Testing Service, the item-response data from the administration of the College Board English Achievement Test and Physics Achievement Test on five occasions over a period of 10 years. The same form of the respective test was administered with exactly the same instructions on each of these occasions. These forms have since been retired.

A preliminary two-way analysis of variance (items × year groups) of the arcsine transformed sample proportion correct for these data showed clear indications of item parameter drift in the Physics Achievement data but little evidence of such effect in English Achievement. We attributed this observation to the greater likelihood of change in physics curricula compared to that of English during the 10-year interval. We therefore narrowed the study to the physics items, and, among them, concentrated on the Mechanics items in the test because there were a sufficient number (29) to justify IRT scaling.

## Tests of Fit

To test the fit of the drift model to the Mechanics achievement data, we began with a model in which all $\delta$ were assumed null, then estimated parameters for the linear and quadratic polynomial model, then relaxed the assumption of equal slopes over groups. Constraints on the asymptotes could not be relaxed because of the rather tight beta priors used to condition their estimation. The other priors are sufficiently mild that their effect can be ignored when interpreting the likelihood ratio tests. The change in marginal log likelihood between models thus provides a test of the statistical significance of the parameters added to the model. The results of these analyses expressed as likelihood ratio chi-squares are shown in Table 1.

The very large change of chi-square between model 1 and model 2 shows that a significant item $\times$ year-group interaction is clearly present in these data. The change of chi-square upon entering the quadratic drift parameters, or leaving the item locations free, is in both cases less than twice the change in degrees of freedom and not definitely established. (The large sample standard deviation of the change in chi-square is equal to the change in degrees of freedom.) Similarly, the change in chi-square upon relaxing the assumption of common slopes gives no reason to reject that assumption.

The Mechanics subscale of the Physics Achievement Test thus appears to be adequately described by a linear item-location drift model in these data. The linear trend lines calculated from the estimated trend coefficients, $\delta$, are shown for each of the items in Figure 2. Separate estimates of the item locations from the item bias model, in which the asymptote and slope parameters are constrained to have the same values as in the drift model, are shown also. The ordinates of these graphs are in standard deviation units in scale-score metric. Negative values correspond to easy items, and positive values correspond to difficult items. Because of the large sample sizes, the standard errors of the

Table 1

29 Physics Achievement (Mechanics) Items
Administered 5 Times in 10 Years

| Model | $\chi^2$ | Change $\chi^2$ | *df* |
|---|---|---|---|
| Constant locations, slopes and asymptotes | 430,081.4 | 678.1 | 28 |
| Linear location drift | 429,383.3 | 50.4 | 29 |
| Quadratic location drift | 429,332.9 | 105.2 | 55 |
| Bias: Constant slopes and asymptotes | 429,227.7 | 104.0 | 120 |
| Constant asymptotes | 429,123.7 | | |

estimated locations are at most .03 on this scale; differences larger than .06 between the estimated locations and the fitted line can be considered statistically significant.

Twenty-one of the 29 plots in Figure 2 show a reasonably regular progression of locations, and thus of difficulty, over the 10 years. For 11 of these 21, plus or minus two standard error intervals around the estimated locations would include the line fitted by the trend model. In the remaining 10, one or two of the points are somewhat discrepant, but not so much as to obscure the trend.

## Interpretation of the Trends

Among these 21, 11 items (10, 25, 33, 39, 40, 62, 63, 69, 70, 73, 75) become relatively easier, and the other 10 items (items 9, 16, 24, 26, 34, 41, 42, 49, 50, 51) become relatively harder. These results have a fairly clear interpretation in terms of the item content of the Mechanics questions of the Physics Achievement Test (see the College Board item specifications for the Physics Achievement Test). Whereas all items in the first set are classified as basic mechanics items, such as kinematics, dynamics, or energy and momentum, only 4 items in the second set deal with those topics. All 6 of the items classified as other topics (rotation, oscillations, gravitation, etc.) are included in the second set that are becoming relatively harder, indicating that their difficulty has increased over the years.

From these results it would appear that teachers are now choosing to concentrate more on the basic topics in mechanics that are covered earlier in the course. The physics curriculum survey of Pfeiffenberger and Wheeler (1984) shows that the topics most frequently covered in physics courses are the basic mechanics. Other topics in mechanics, and indeed topics in other areas of physics, appear to be receiving relatively less emphasis. These trends may reflect textbook changes influenced by Physical Science Study Committee (PSSC) physics in the late fifties and Project Physics in the early sixties.

Another noteworthy trend is the phasing out of English units of measurement in the physics curriculum. This effect is evident, for example, in the contrast between items 33 and 34 from a set that includes the distinction between mass and weight. Item 33 uses metric units (kilograms for mass), and item 34 uses English units (pounds for weight). Earlier, both systems of units were used significantly in physics texts and in forms of the Physics Achievement Test. But about 1977, English units were excluded from all new forms of the test, and metric units are now used almost exclusively in most textbooks. This difference could account for the opposite slope of changes in items 33 and 34.

Eight of the plots have a clearly outlying point, disrupting what would have otherwise been a systematic trend (items 7, 8, 10, 12, 18, 33, 49, 51, 70). The points suggest the existence of cohort effects specific to particular items. It is difficult to imagine how such effects could occur in a nationwide sample, unless some information relevant to these items appeared in the national media in these years, or perhaps in publications read by the physics teachers. In the present data, these effects are apparently random and infrequent enough to justify the

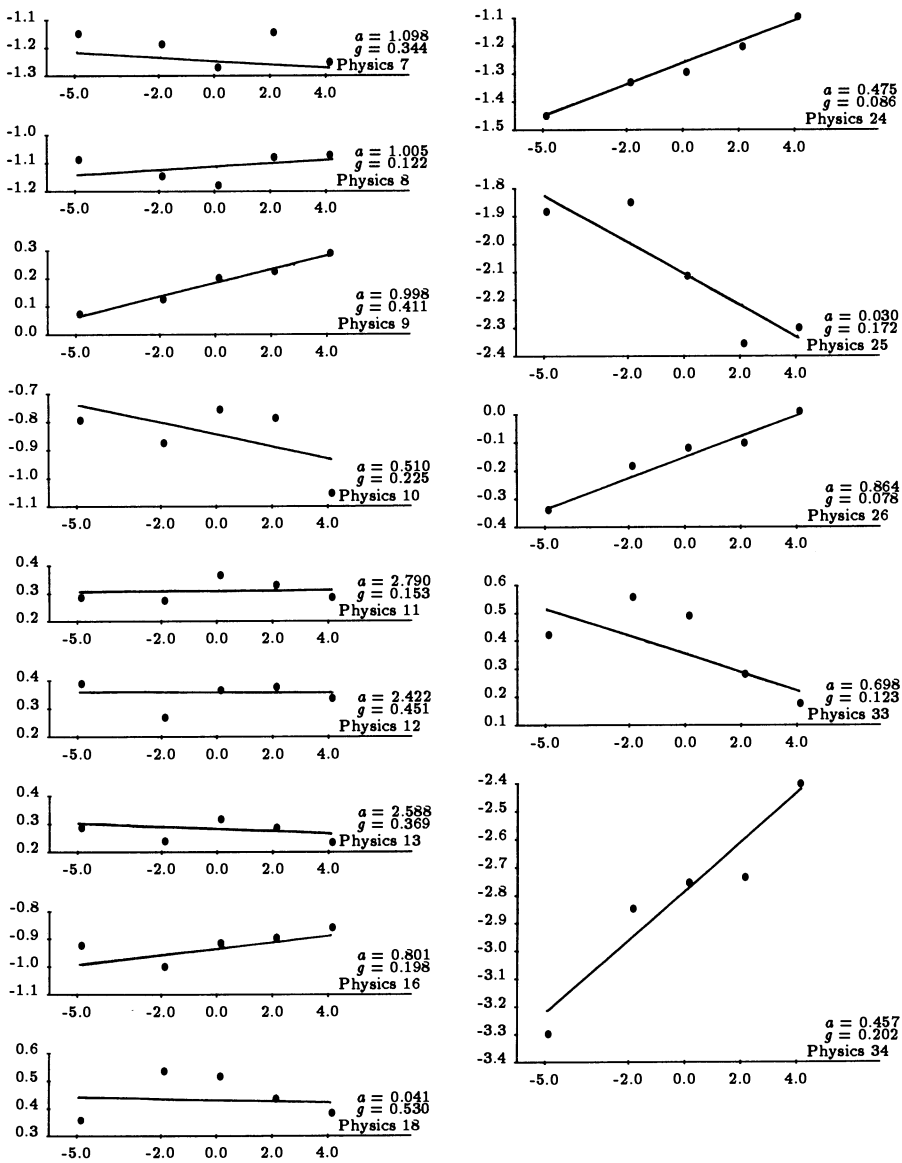


FIGURE 2. *Item trend lines*

*Note.* Abscissa is time in years, ordinate is location of item in standard deviation units.

averaging of the data that the trend lines provide. They suggest, however, that long-term testing programs should not rely on item calibrations in a single cohort.

Also shown in Figure 2 are the estimated slopes and asymptotes for each item. There is no apparent relationship between the item parameters and the goodness-of-fit of the trend lines.

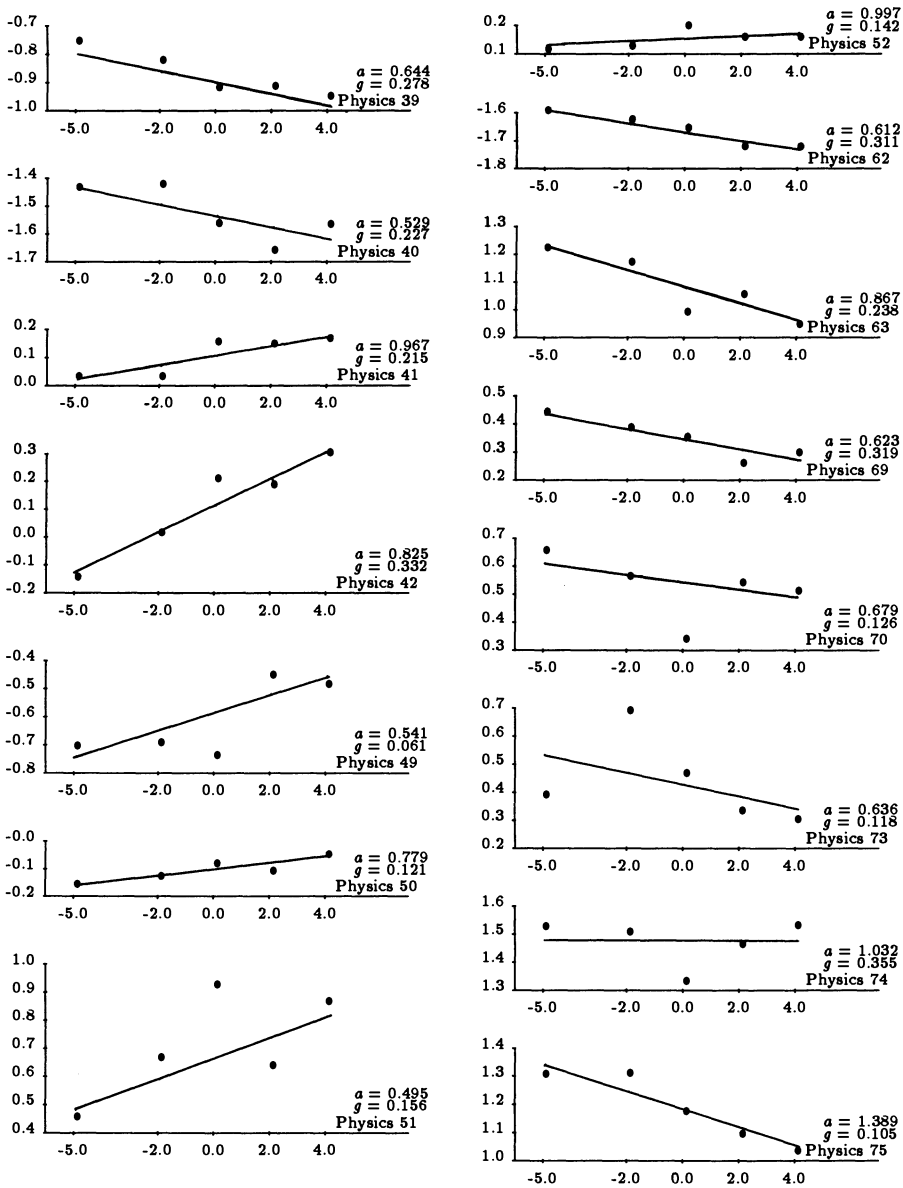


FIGURE 2 (cont'd): *Item trend lines*

The year-group sample sizes, means, and standard deviations estimated in these data, shown in Table 2, are also of interest. Overall location and scale of the estimates have been set so that the mean is zero and standard deviation is unity in group 3 (1978). Apart from the 1976 sample, where achievement appears to be somewhat depressed, the estimated population means are remarkably constant

Table 2

Estimated Population Means and Standard Deviations Under the Linear Drift Model

| Year | Month | N | Mean | S.D. |
|---|---|---|---|---|
| 1973 | January | 2380 | -0.0508 | 1.0010 |
| 1976 | April | 2110 | -0.3085 | 0.9511 |
| 1978 | May | 1268 | 0.0000 | 1.0000 |
| 1980 | January | 3625 | -0.1522 | 0.9487 |
| 1982 | January | 3634 | -0.1460 | 1.0303 |

over the 10 years. There is no evidence among these self-selected respondents of the decline that has been noted in Scholastic Aptitude Test (SAT) scores.

## Variant Items

When variant items are included in a test (in order to anticipate how they will function if used operationally), it is desirable to estimate their item parameters without allowing them to influence the underlying dimension of the scale defined by the existing items. In MML estimation of item parameters, this is easily accomplished by excluding the variant items from the calculations of the likelihoods used in estimating the expected numbers of correct responses and numbers of respondents at the quadrature points (the E step of the EM algorithm). The estimation of item parameters from these expected values can then be carried out for all items, including the variant items (the M step of the EM algorithm). The procedure is the IRT analogue of computing factor loadings for a new test by extension from the factor solution of an existing test battery.

Because of possible position effects when a variant item is introduced into the operational test, parameter estimates obtained by extension should not be used when scoring the operational test. A new calibration based on current operational data, if available, is preferable. For the same reason, the variant items should be appended at the end of the operational scales and not inserted among the functioning items.

## Conclusions

The present study supports empirically a number of conclusions about item parameter drift that seem reasonable on prior grounds.

1. Differential drift can occur over a period of years in a nationally administered educational test if curricular emphasis on the relevant subject matter is changing.

2. Drift will affect item locations (difficulties) much more strongly than item slopes (validities) during the tenure of the items in the test.

3. Differential drift of item locations are relatively steady in large populations and are describable as a linear function of time.

4. A linearly time-dependent item response model can describe educational test data accurately enough to support an IRT-based system for maintaining

consistent scales of measurement over an extended period as items are retired and replaced in the item pool. A system of this type, based on an extension of the BILOG program to multiple groups, has been prepared by Muraki and Bock (1987).

## References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Bock, R. D. (in press). Measurement of human variation: A two-stage model. In R. D. Bock (Ed.), *Multilevel analysis of educational data.* New York: Academic Press.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369–377.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Muraki, E., & Bock, R. D. (1987). *BIMAIN: A program for item pool maintenance in the presence of item parameter drift and item bias.* Mooresville, IN: Scientific Software.

Pfeiffenberger, W., & Wheeler, G. F. (1984). A curriculum survey of high school physics courses. *The Physics Teacher, 22,* 569–575.

Thissen, D., & Bock, R. D. (1986). *A 4-parameter logistic item response model.* Unpublished manuscript.

## Authors

R. DARRELL BOCK, Professor, University of Chicago, Dept. of Behavioral Sciences, 5858 S. University Ave., Chicago, IL 60637. *Degrees:* BS, Carnegie Institute of Technology; MA, PhD, University of Chicago. *Specializations:* psychometrics, educational assessment, measurement of growth.

EIJI MURAKI, Research Analyst, National Opinion Research Center, 1155 E. 60th St., Chicago, IL 60637. *Degrees:* BA, Waseda University; MA, University of Minnesota; PhD, University of Chicago. *Specializations:* item response theory, multivariate analysis.

WILL PFEIFFENBERGER, Senior Examiner, Educational Testing Service, Princeton, NJ 08534. *Degrees:* BS, University of Notre Dame; MS, University of Iowa; MS, Rutgers University. *Specialization:* test development.