Open access • Journal Article • DOI:10.1177/014662168601000402

Linking Item Parameters Onto a Common Scale — Source link

C. David Vale

Institutions: Assessment Systems Corporation

Published on: 01 Dec 1986 - Applied Psychological Measurement (SAGE Publications)

Topics: Item bank

Linking Item Parameters Onto a Common Scale

C. David Vale
Assessment Systems Corporation

An item bank typically contains items from several tests that have been calibrated by administering them to different groups of examinees. The parameters of the items must be linked onto a common scale. A linking technique consists of an anchoring design and a transformation method. Four basic anchoring designs are the unanchored, anchor-items, anchor-group, and double-anchor designs. The transformation design consists of the system of equations that is used to translate the anchor information and put the item parameters on a common scale. Several transformation methods are discussed briefly. A simulation study is presented that compared the equivalent-groups method with the anchor-items method, using varying numbers of common items, applied both to the situation in which the groups were equivalent and one in which they were not. The results confirm previous findings that the equivalent-groups method is adequate when the groups are in fact equivalent. When the groups are not equivalent, accurate linking can be obtained with as few as two common items. Linking using a more efficient interlaced anchor-items design can provide accurate linking without the expense of including explicit common items in each of the tests.

From a psychometric perspective, a test item consists of substantive content and statistical characteristics. While the substantive content of the item is most important from the examinee's viewpoint, the statistics aid in the construction of new tests with desirable psychometric properties. The statistical characteristics of an item are most usefully expressed as parameters of an item response theory (IRT; Hulin, Drasgow, & Parsons, 1983; Lord, 1980) model. The parameters of an IRT model are usually determined empirically by administering the items. For the parameters to be comparable across the items, and for arbitrary subsets of the items to be useful in combination with each other, all parameters must be expressed on a common scale. An item bank typically contains many more items than will be administered to any one examinee. Thus, parameters of the items often must be estimated from several groups of examinees and then linked together onto a common scale. Such linking is the topic of this paper.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 4, December 1986, pp. 333–344

0146-6216/86/040333-12$1.85

Basic Concepts in Linking

Two popular IRT models are the Rasch model and the three-parameter logistic model. The item characteristic curve (ICC) for the three-parameter model expresses the probability of a correct response as a function of ability (θ) and three item parameters: a, the discrimination; b, the difficulty; and c, the probability of answering the item correctly through guessing. The functional relationship is given by

$$P = P(1|\theta) = c + (1 - c)\Psi[1.7a(\theta - b)] \quad , \qquad (1)$$

where

$$\Psi(x) = 1/[1 + \exp(-x)] \quad . \tag{2}$$

The Rasch model is a restricted form of the three-parameter logistic model in which the a parameter is set to .588 and the c parameter is set to 0.

Before IRT can be used to infer test characteristics, the item parameters must be estimated. This process of estimating item parameters, called calibration, is accomplished by simultaneously estimating the θs and item parameters on a long test administered to a large sample of examinees. Although a description of calibration is beyond the scope of this paper (see Vale & Gialluca, 1985, or Wingersky, 1983, for a detailed discussion of calibration), the process involves manipulating trial item parameter estimates and trial θ estimates until the theoretical ICC fits (i.e., maximizes the likelihood of) the observed data.

Neither the Rasch model nor the three-parameter model is completely determined by the data. Since both models work from the difference between the examinee's θ and the item's difficulty, the scale of θ can be changed through a linear transformation, as long as the item parameters are also changed consistently.

Consider the case of an alternate scale, θ*, in which the item characteristic function is given by Equation 3:

$$P^* = P(1|\theta)^* = c + (1 - c)\Psi[1.7a^*(\theta^* - b^*)] \quad . \tag{3}$$

Both the θ and θ* scales are equally satisfactory representations of the same data if $P = P^*$, which implies that

$$a(\theta - b) = a^*(\theta^* - b^*) \quad . \tag{4}$$

Say that the relation between the scales is given by

$$\theta^* = k\theta + m \quad . \tag{5}$$

Then

$$a^* = a/k \tag{6}$$

and

$$b^* = kb + m \quad . \tag{7}$$

Note that the relation of b^* to b is the same as the relation of θ* to θ.

The calibration process must fix the scale, which can be done somewhat arbitrarily. Traditionally, Rasch calibrations fix the scale by setting the mean item difficulty to 0 and letting the θ scale float accordingly. Three-parameter calibration processes traditionally fix the θ scale such that the mean of the sample θ distribution is 0 and its variance is 1.

An often-touted virtue of IRT is its sample independence, however, and fixing the scale to characteristics of each calibration sample defeats this feature. Linking is the process whereby the item parameters are adjusted to put all of them onto a common scale. In the case of the three-parameter logistic model, only the a and b parameters are scale-related. The c parameters are not affected by linking because they are on the probability metric rather than the θ metric.

Linking is often considered a separate operation from calibration. Items may be calibrated first and then the parameters may be adjusted to a common scale. The common scale is determined by extra-calibration comparison of the item parameters or the resultant θ estimates. The distinction between calibration and linking is more apparent than real, however. If a single item were calibrated, the two scale-related parameters a and b could legitimately take on almost any values; there would be a θ scale to correspond to whatever pair of parameters was developed. If this item were to be used in conjunction with another item, however, the parameters would have to be adjusted so that the θ scales corresponding to the parameters of the two items were the same. Traditionally, if this were accomplished by estimating the parameters of both items simultaneously, it would be called calibration. If it were done by transforming the parameters to a common scale after calibration, it would be considered linking.

Anchoring Designs

There are a number of ways linking can be accomplished. All of these methods are composed of an anchoring design and a linking transformation. The anchoring design refers to the way in which tests and examinee samples are assembled. The

linking transformation refers to the equations used to put the item parameters onto a common scale.

The anchoring design ensures that there will be a basis for comparison among separate calibrations. Consider the case of two tests administered to two groups of examinees. An anchor consists of a person who answers items from both tests or an item that is taken by members of both groups. In concept, the difference in parameter estimates obtained from the two groups on the common items, or the difference in θ estimates obtained on the two tests by the common group of examinees, is used to provide the information necessary to transform the two sets of parameter estimates to a common scale.

The boxes in Figure 1 represent matrices of persons and items in which each intersection of a person and an item is a potential item administration. An item administration consists of one item administered to one person. The best calibration design would be one in which all items were administered to all persons. Economic considerations usually make this impossible, however, because the cost of a calibration and linking study is directly related to the number of item administrations. A good anchoring design is one that results in the best calibration and linkage for a given number of item administrations. In the examples shown in Figure 1, exactly half of the potential item administrations are assumed to occur. An anchoring design refers to the arrangement of item administrations in the incomplete matrices.

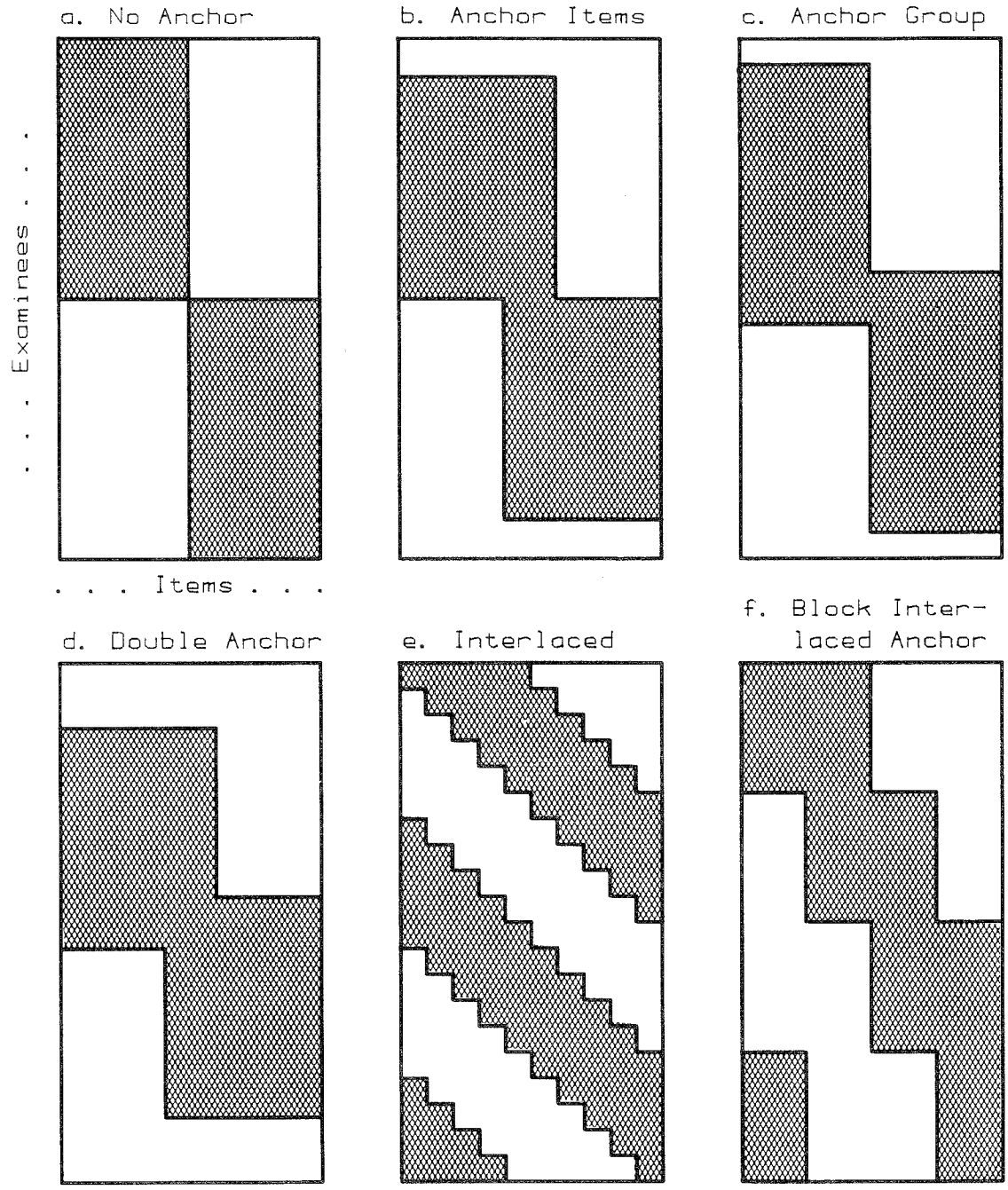
Figures 1a through 1d show, in schematic form, the basic anchoring designs. Figure 1a illustrates an unanchored design. The item pool is split into two tests with no overlap and the sample of examinees is split similarly. The first group takes the first test and the second group takes the second test. If the parameters are to be on the same metric, this must be accomplished by making either the groups or the tests equivalent. Making the groups or tests equivalent is external to the design. Groups may be made equivalent by randomly assigning, to each group, examinees sampled from one population. If the samples are large, the equivalent-groups method can be an effective method of equating (Vale, Maurelli, Gialluca, Weiss, & Ree, 1981). Similarly, tests may be made equivalent by sampling items from a common domain. Unless the tests are unusually long, however, the equivalent-tests method is unlikely to be an effective substitute for anchoring even if the items are randomly assigned to tests. The major advantage of the unanchored design is that administrations are evenly spread across the items and each item has an equal number of administrations. Although there is no guarantee that an equal number of administrations will yield equally good estimates of each item's parameters, when there is no prior knowledge of item parameter values or examinee θs an equal number of administrations provides the best opportunity for equivalent precision of estimation.

Figure 1b shows the classic anchoring design, the anchor-items design. In this design, a subset of the items is contained in both tests. The parameters of these common items are compared to determine the linear transformation necessary to put all of the items onto a common scale. The major advantage of this design is that equivalence need not be assumed for either the tests or the samples. Its major disadvantage is that the parameters of the common items tend to be well-estimated at the expense of the parameters for the unique items, which have fewer administrations.

Figure 1c is the transpose of Figure 1b. It represents an anchor-group design, in which a common group of examinees takes both tests. The item parameters of the two tests are put on a common scale by finding the transformation that yields equivalent θ distributions for the common examinees when θs are estimated on the two tests. Like the anchor-items design, this design eliminates the need to assume equivalence of tests or examinees. It also overcomes the disadvantage of uneven quality of estimation across the items. Unfortunately, it does this at the expense of uneven quality of θ estimation, which may have an effect on parameter estimation. An additional practical disadvantage is that it may be difficult to find a common group of examinees who can and will take all of the items in the bank.

Figure 1d shows a double-anchor design in which there are both common items and common ex-

Figure 1
Basic Anchoring Designs



aminees. Although there is some surface appeal to this design because of its apparently firmer anchoring, it appears to have all of the disadvantages of the two one-way anchor designs (i.e., uneven quality of estimation of both item and person parameters) with no additional benefits. Furthermore, there is no simple transformation for the double-anchor design. Not surprisingly, this design has seen little use.

These four designs are only prototypes: The tests need not be of equal length, nor must the sample sizes be equal. Furthermore, the limit on the number of tests is practically infinite, if all combinations of items are considered, as is the maximum number of possible samples.

Figure 1e shows an interlaced anchoring design. In this design, the number of tests is equal to the number of items. The first test begins with the first item and runs sequentially through the items until it reaches its established length. The second test begins with the second item. The final tests run to the last item and then wrap around and continue from the beginning. There are two advantages to the interlaced design. First, an anchor-items effect is achieved while keeping the numbers of administrations equal across items and test lengths equal across examinees. The second advantage is that the design also achieves an equivalent-groups effect, even when it is applied to two distinctly different groups of people; only if the design is explicitly perverted will there be a substantial difference in θ levels among examinees taking different tests. The disadvantages are that the simple linking transformations are no longer simple and a large number of tests must be printed if the calibration is done in the paper-and-pencil mode. However, a joint calibration procedure (discussed below) neatly solves the transformation problem, and the number of forms is not a problem for computer-administered tests.

Figure 1f shows a block-interlaced anchoring design. This is similar to the completely interlaced design except that the number of tests is less than the number of items. This design, like the fully interlaced design, administers all items to an equal number of examinees. It offers the practical advantage of requiring fewer tests.

Linking Transformations

The linking transformation places the parameters on a common scale. The simple procedures find a linear transformation that can be used to express one θ scale or one item difficulty scale in terms of another, and then extract the coefficients from that transformation, according to Equations 5 and 7. These methods differ in the ways in which they estimate the coefficients of the linear transformation from these equations.

A linking transformation, like an equating transformation, must be symmetric; it must yield an equivalent transformation regardless of which scale is chosen for the common scale (e.g., Test X, Test Y, or some other score scale). Thus, it cannot be a regression equation. The transformation constants can be obtained from the equations

$$k = S^*/S \tag{8}$$

and

$$m = X^* - X(S^*/S) \quad , \tag{9}$$

where X^*, X, S^*, and S are respectively the scale origins and units. The means and standard deviations of either the θs or the b parameters from the two tests are typically used. In fact, these distributions are unknown, and distributions of estimates are used instead. Vale et al. (1981), for example, used the θ estimates; Marco (1977) used the difficulty estimates.

There is a subtle biasing effect that results from using the estimates of either the θs or the bs. Specifically, although the means are unbiased estimates of the means of the true origin parameters, the variances of the distributions of estimates are inflated because the estimates of the θs and bs contain error. This problem is pronounced only when the quality of the estimates differs across the samples or tests being linked (and thus the amount of error varies). If the sample sizes or the test lengths are unequal, the quality of estimation is likely to be more uneven.

Several variations in the method of determining the transformation constants have been suggested. Bejar and Wingersky (1981) and Vale et al. (1981) used robust estimation of the moments as one alternative; this procedure gives less weight to deviant values in the estimation of the origin and unit scale constants than do standard procedures. Ironson (1982) and Reckase (1979) used principal components analysis to determine the transformation constants. More recently, Stocking and Lord (1983) suggested a substantial revision of the transformation procedure. Their method minimizes the differences between estimated true scores (i.e., test characteristic curves) on the anchor tests across samples. (An estimated true score is computed as the sum of the ICC probabilities of all common items administered to each examinee.) Divgi (1985) suggested a computationally simpler transformation that, like the Stocking and Lord procedure, uses both *a* and *b* parameters in making the transformations but also considers the errors of estimate in the parameters. To date, there has been little evidence that any of the complex procedures are superior to simple mean and standard deviation transformations.

A final transformation method was developed from a feature originally incorporated into the calibration program LOGIST (Wingersky, Barton, & Lord, 1982). The algorithm used by LOGIST estimates the θs and the item parameters in two phases. This phased process makes the simultaneous or joint estimation of θs and item parameters numerically tractable. The important feature is LOGIST's capacity to estimate θs for examinees on a subset of the items and to estimate the parameters of the items on a subsample of the examinees. When items not administered to any examinee are coded as "not reached," they are disregarded in the estimation process. Thus, the technique uses all of the available information in a simultaneous calibration.

This final transformation technique virtually eliminates the distinction between calibration and linking. Anchoring designs that traditionally would have required two calibration runs and a linking transformation can be accomplished in a single calibration run.

Linking an Item Bank

Linking a bank of items together onto a common scale should be accomplished using a design that produces a set of parameters that work well together for the estimation of θ. The choice of a design amounts to the selection of an anchoring design and a transformation method. The linking design should not be chosen without considering calibration, however. The overall goal of calibration and linking is to express each item's parameters on a common scale (i.e., to accurately estimate the parameters), not merely to adjust each subtest onto a common scale.

A relatively large study of linking by Vale et al. (1981), considering overall accuracy, suggested that (1) when the groups are equivalent, the equivalent-groups procedure works quite well; (2) when the groups are not equivalent, the anchor-group method or the anchor-items method should be preferred; (3) the anchor group should contain at least 30 examinees; and (4) there may be no difference between anchor tests of 5, 15, and 25 items. This last finding was probably the most bothersome, because an anchor test of five items seems rather short.

Recent work by Wingersky and Lord (1984) also suggests that short anchor tests coupled with joint calibration may be acceptable. They noted remarkably little difference in the *b* parameter error when the length of the anchor test was reduced from 50 items to 2 items.

Some other empirical work seems to suggest that five items are insufficient. One empirical finding was pointed out by Vale et al. (1981) concerning a study done by Ree and Jensen (1980). Ree and Jensen compared methods of common item linking using two 80-item tests with 20 overlapping items at various sample sizes. Because they used equivalent groups of examinees and their calibration program fixed the θ scale on the same metric each time, their items were linked using an equivalent-groups procedure before they ever applied the anchor-test equations. Comparing their results before and after linking was illuminating because their parameters were typically more accurate before their explicit anchor-test linking than after.

A Simulated Comparison of Linking Designs

Method

To further investigate the allocation of testing time to anchor items and to evaluate the utility of the interlaced design, which in theory can emulate the effects of equivalent groups and a common-items anchor without the problems inherent in either design, a small simulation study was conducted. In this study, linking effectiveness was evaluated under two conditions in two-way designs.

Conditions. The two conditions were linking with equivalent and non-equivalent examinee samples. For the equivalent-groups simulation, two examinee samples were taken from the same population which had a mean θ of 0 and a variance of 1. For the non-equivalent groups, two samples of examinees were taken from populations with means .5 units apart. One population had a mean θ of $-.25$ and the other had a mean θ of .25. The variances in both populations were .9375, the value that made their combined variances 1.0. θ was normally distributed in all populations. The non-equivalent groups in this simulation represent distributions of θ that might arise if examinees were sampled from two different sources.

Items. A pool of 60 hypothetical items was used. Difficulty parameters (b) ranged from -2.1 to 2.1 in increments of .3. At each difficulty level were four items, two with a parameters of 1.0 and two with a parameters of 1.5. All c parameters were .2. From this pool, tests of 30, 31, 32, 35, and 40 items were constructed. The items included in each of these tests are shown in Table 1.

Test length and anchoring design. The two factors in the two-way designs were test length (and thus the number of overlapping items) and anchoring design. Three anchoring designs were investigated: separate, standard, and interlaced. In the separate and standard designs, items were divided into two tests. Tests of 30 items in length

Table 1
Items Included in the Tests

b	Test 1 Unique Items a = 1	Test 1 Unique Items a = 1.5	Test 2 Unique Items a = 1	Test 2 Unique Items a = 1.5
-2.1	1	2	31	32
-1.8	3	4	33	34
-1.5	5	6	35	36
-1.2	7	8	37	38
-0.9	9	10	39	40
-0.6	11	12	41	42
-0.3	13	14	43	44
0.0	15	16	45	46
0.3	17	18	47	48
0.6	19	20	49	50
0.9	21	22	51	52
1.2	23	24	53	54
1.5	25	26	55	56
1.8	27	28	57	58
2.1	29	30	59	60
Common items:				
31-item test length:	22, 39			
32-item test length:	22, 39, 10, 51			
35-item test length:	22, 39, 10, 51, 9, 16, 21, 40, 45, 52			
40-item test length:	22, 39, 10, 51, 9, 16, 21, 40, 45, 52, 3, 4, 15, 27, 28, 33, 34, 46, 57, 58			

had no items in common, tests of 31 items in length had 2 items in common, and tests of 40 items in length had 20 items in common. In the separate design, each of the two tests in each cell was calibrated separately using ASCAL (Vale & Gialluca, 1985), a microcomputer program similar to LOGIST, and the items were linked by deriving the transformation parameters from the means and standard deviations of the *b* parameters. In the standard method, both test forms were calibrated simultaneously in a single run of ASCAL.

In the interlaced design, a form of the test began with each sequential item in the set of 60 and continued for the length of the test. Thus for each length there were 60 tests; the 30-item tests had, on average, approximately 15 items in common. All interlaced designs were calibrated using the simultaneous procedure in which all forms were calibrated in a single run of ASCAL. Note that although the constraints of the environment (i.e., two distinct groups of examinees and the capacity to use only half of the potential item administrations) were maintained, the nature of the interlaced design resulted in effectively equivalent groups and substantially more common items for a given test length, compared with the anchor-items design.

Examinees. The number of examinees ranged from 750 to 1,000 for each calibration. For each design, the number of examinees was adjusted so that the number of item administrations was approximately 60,000. Thus 1,000 examinees took the 30-item tests, 968 took the 31-item tests, 938 took the 32-item tests, 857 took the 35-item tests, and 750 took the 40-item tests. Such adjustment of the number of examinees simulated a constant amount of examinee time for each test length.

Each cell of each experimental design contained four replications. Each replication contained independently simulated examinees. For each replication, responses were independently generated and the items were calibrated.

Criteria. Two criteria of linkage quality were evaluated in each cell. The first was an efficiency criterion developed by Vale et al. (1981; Vale & Gialluca, 1985). This criterion is the ratio of the psychometric information (Birnbaum, 1968) that can be extracted using the estimated parameters to the information that can be extracted using the true parameters. The second criterion was the root mean squared error (RMSE) of the *b* parameters. This was computed as the square root of the average of the squared differences between the true and the estimated *b* parameters across the 60 items. The four replications in each cell were then averaged.

RMSE is an index often used in evaluations of calibration and linking. It is useful, however, only if the scale onto which the parameters are linked is the same as the true scale. In simultaneous calibrations, the scale is defined to have a mean of 0 and a variance of 1, the parameters of the true θ distributions used in the simulations. In separate calibrations, the scale of one administration is typically expressed on the scale of the other. This makes RMSE comparisons with true parameters meaningless. RMSE was thus not computed for the separate calibration cells.

The significance of the differences among test lengths and anchoring designs was investigated separately for the equivalent- and non-equivalent-groups conditions using two-way fixed-effects analyses of variance (ANOVAS). Efficiencies and RMSEs were evaluated separately as dependent variables, without transformation. The ANOVAS were run excluding the 30-item tests. In a sense, the 30-item tests represented an unfair comparison of the separate and standard designs with the interlaced design because, at that length, only the interlaced design had any items in common among forms.

Results

Mean efficiencies of calibration in the equivalent-groups condition are summarized in the upper left portion of Table 2 and graphically portrayed in the upper left graph of Figure 2. Little difference was observed between the three anchoring designs. The conclusion of equivalent efficiency with equivalent groups was supported by the lack of significant differences between designs in the analysis of variance (Table 3). There were significant differences among test lengths, however, with longer test lengths producing higher efficiencies. The effects of test

Figure 2

Efficiency and *b*-Parameter RMSE

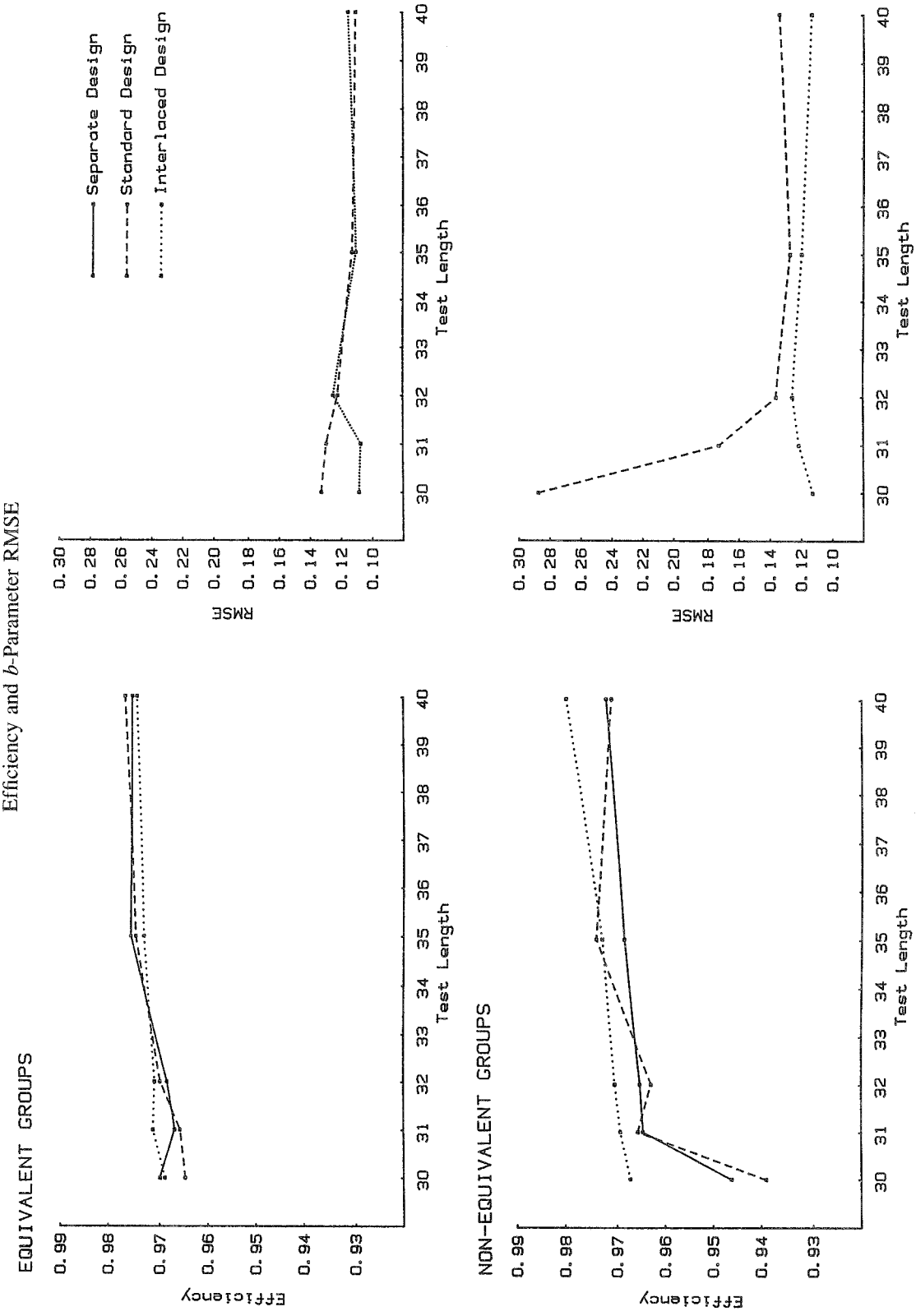


Table 2
Mean Efficiency and b Parameter RMSE

	Efficiency				RMSE		
Length	Separate	Standard	Inter-laced	Mean	Standard	Inter-laced	Mean
Equivalent Groups							
30	.9695	.9644	.9685	.9675	.1322	.1082	.1202
31	.9665	.9654	.9709	.9676	.1288	.1070	.1179
32	.9680	.9694	.9705	.9693	.1216	.1244	.1230
35	.9754	.9744	.9727	.9742	.1124	.1100	.1112
40	.9749	.9764	.9740	.9751	.1100	.1148	.1124
Mean	.9712	.9714	.9720		.1182	.1141	
Non-Equivalent Groups							
30	.9463	.9393	.9670	.9509	.2876	.1124	.2000
31	.9644	.9654	.9690	.9663	.1718	.1210	.1464
32	.9650	.9627	.9701	.9659	.1353	.1250	.1302
35	.9683	.9740	.9728	.9717	.1268	.1194	.1231
40	.9719	.9708	.9800	.9742	.1334	.1130	.1232
Mean	.9674	.9682	.9730		.1418	.1196	

Note: Column means exclude the 30-item tests.

length accounted for 27% of the variance in efficiency (i.e., $\eta = .52$).

The upper right graph in Figure 2 shows plots of the *b*-parameter RMSE for the equivalent groups. No differences are apparent in the figure, and no significant differences were found in mean RMSE (Tables 2 and 3) for either test length or design. Thus, when the parameters are considered individually, if the groups are equivalent it appears to make no difference how the available administration time is distributed among tests or how the tests are anchored.

The lower left graph in Figure 2 shows plots of efficiency for the non-equivalent groups. These are markedly different from the comparable plots for the equivalent groups. The interlaced design produced the highest efficiencies in four of the five comparisons; the separate and standard designs produced approximately equivalent efficiencies. The separate and standard designs were distinctly inferior for the 30-item (no-overlap) tests. The effect of test length on efficiency (disregarding the 30-item tests) was significant. The η (.55) was remarkably similar to that of the equivalent-groups condition (.52).

The lower right graph in Figure 2 shows the *b*-parameter RMSE for the non-equivalent-groups condition. The interlaced design provided smaller errors at all levels of test length than did standard anchoring. The RMSE in the standard design dropped until tests were 32 items long, after which it leveled out. As shown in Table 3, test length, anchoring design, and the interaction between them were all significant at the 5% level. This interaction apparently occurred because the RMSE of the standard anchoring design decreased with test length while that of the interlaced design did not.

Discussion

This study attempted to elucidate the best way to design a calibration and linking study with a fixed amount of examinee resources. The results suggest that the tests should be as long as possible and that the anchoring should be done using the interlaced design, if feasible. The study did not carry the design to its logical extreme, where the number of examinees was halved and all examinees took all items. Extrapolation from the trends found in the data suggests that this would be the optimal design.

The data also suggested that anchoring is unnecessary but not harmful when the groups are equivalent. In this condition, the anchoring design

Table 3
Analysis of Variance Excluding the 30-Item Tests

Criterion and Source	df	Equivalent Groups F	p	η	Non-Equivalent Groups F	p	η
Efficiency							
Length	3,36	4.746	.007	.52	7.567	.001	.55
Design	2,36	0.093	.912	.06	5.525	.008	.38
Interaction	6,36	0.498	.806		.907	.501	
RMSE (b)							
Length	3,24	0.733	.542	.28	4.230	.016	.44
Design	1,24	0.183	.673	.08	17.396	.001	.52
Interaction	3,24	0.560	.647		3.473	.032	

appears to make little difference in terms of efficiency and RMSE of *b* parameters. The test length still is important, however.

Significant improvements in efficiency were observed as the test lengths, and the numbers of common items, increased (at the expense of sample size) in both the equivalent- and non-equivalent-groups conditions. Since test length and overlap were confounded in this study, this could be due either to test length effects (noted to be substantial by Vale et al., 1981) or to the concomitant increases in the number of common items.

Results suggest that the number of common items does not improve efficiency in the equivalent-groups condition: The effective number of items in common in the interlaced design was substantially greater than for the other two methods for all test lengths. If anchor items improve linking in the equivalent-groups condition, a significant difference among anchoring designs should have been observed; it was not. This suggests that common items are not useful when groups are equivalent.

Similarly, some evidence indicates that the number of common items is not important when the groups are not equivalent. The variance in efficiency accounted for by test length (i.e., η^2) in the non-equivalent-groups condition was essentially equal to that of the equivalent-groups conditions. This suggests that the differences in efficiencies were due to improvements in calibration accuracy (from increased test length) rather than from linking accuracy (from more common items). Although there were differences among the designs when the groups were not equivalent, this was due primarily to the greater efficiency of the interlaced design; this may have been due to its ability to make the groups effectively equivalent, rather than its having more common items.

Thus it appears that interlaced, simultaneous calibration is the best way to link items. The data also suggest that as few as two common items may be sufficient for linking, using non-interlaced designs. This conclusion is not definitive, however, because test lengths were confounded with numbers of common items in this study. Furthermore, other factors that may complicate practical linking studies (e.g., multidimensionality) were not considered in this study.

Conclusions

Large item banks must often be developed and administered in parts. For IRT to be a viable means of calibrating the items in the bank, the parameters must be expressed on a common scale by means of a linking technique.

In general, calibration and linking should be done with tests as long and having as many common items as possible. For a fixed number of available item administrations, it is better to give longer tests at the expense of sample size. For the general case, the best design appears to be the jointly calibrated interlaced design. If, however, the examinee groups are large random samples from a common popu-

lation, the linking method is largely irrelevant; even so, longer tests are still preferable to short tests because of improved calibration accuracy.

References

Bejar, I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81–8/ETS No. 81–35). Princeton NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood IL: Dow Jones-Irwin.

Ironson, G. (1982). Chi-square and item response theory techniques. In R. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Reckase, M. D. (1979, April). *Item pool construction for use with latent trait models*. Paper presented at the 1979 Convention of the American Educational Research Association, San Francisco.

Ree, M. J., & Jensen, H. E. (1980). *Item characteristic curve parameters: Effects of sample size on linear equating* (AFHRL-TR-79-70; AD-A082 341). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Vale, C. D., & Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters* (Research Report ONR-85-4). St. Paul MN: Assessment Systems Corporation.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). *Methods for linking item parameters* (AFHRL-TR-81-10). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *ERIBC monograph on applications of item response theory*. Vancouver BC: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Acknowledgments

The author thanks Marcia M. Andberg, Kathleen A. Gialluca, and J. Stephen Prestwood for their assistance in preparing this paper, and an anonymous reviewer for helping to clarify some characteristics of the interlaced design.

Author's Address

Send requests for reprints or further information to C. David Vale, Assessment Systems Corporation, 2233 University Avenue, Suite 310, St. Paul MN 55114, U.S.A.