ACT Research Report Series 88-2

Comparison of Item Preequating and Random Groups Equating Using IRT and Equipercentile Methods

Michael J. Kolen
Deborah J. Harris

November 1988

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

COMPARISON OF ITEM PREEQUATING AND RANDOM GROUPS EQUATING USING IRT AND EQUIPERCENTILE METHODS

Michael J. Kolen
Deborah J. Harris

Table of Contents

ABSTRACT

An item preequating design and a random groups design were used to equate forms of the ACT Assessment Mathematics test. Equipercentile and three-parameter logistic model IRT procedures were used for both designs. Both pretest methods produced inadequate equating results, and the IRT item preequating resulted in more equating error than had no equating been conducted. Although neither of the item preequating methods performed well, the results from the equipercentile preequating method were more consistent with those from the random groups method than were the results from the IRT item pretest method. Item context and position effects were likely responsible, at least in part, for the inadequate results for item preequating. Such effects need to be either controlled or modeled and the design further researched before the item preequating design can be recommended for operational use.

COMPARISON OF ITEM PREEQUATING AND RANDOM GROUPS EQUATING USING IRT AND EQUIPERCENTILE METHODS

In the item preequating design for equating alternate forms of a test, items to be included in subsequent forms are pretested during the operational administrations of already equated forms. Item statistics for the pretested items are then used to equate scores on the newly constructed forms to the scale used for reporting scores. One presumed benefit of this design is that new forms can be equated prior to administering them intact. Thus, the item preequating design can be used in situations, such as those that result from test legislation, that require all items contributing to examinees' scores to be released to the examinees. In addition, because most testing programs pretest items, item statistics for pretested items are often readily available. However, for the item pretesting design to produce acceptable equating results, test items must behave similarly in pretest and operational contexts. In particular, item position and other context effects might lead to differences in the way items behave in these two settings and, therefore, result in inadequate equating.

In the present study, the feasibility of using the item preequating design to equate the ACT Assessment Mathematics test (American College Testing Program, 1988) is investigated. The equating results from the item preequating design are compared to the results from the operational ACT Assessment equating design that uses randomly equivalent groups with the operational results used as the criterion. The results from the preequating design are also compared to "no equating." In "no equating," the raw-to-scaled score conversion table is constructed under the assumption that the conversion table on a new form is identical to that for a given old form. Equipercentile and three-parameter logistic item response theory (IRT) equating methods are compared for both designs.

Eignor (1985), Eignor and Stocking (1986), and Stocking and Eignor (1986) conducted a series of studies on the use of IRT item preequating with the Scholastic Aptitude Test (SAT) (Donlon 1984). These authors concluded that the item preequating design was inadequate for the SAT Mathematical test. They also concluded that the problems resulted from a combination of test multidimensionality and differences in examinee groups having a negative effect on the IRT estimation process.

These results on the SAT suggest that item preequating might be problematic with the ACT Mathematics test. However, the extent to which the results from the SAT Mathematical test can be generalized to the ACT Mathematics test is not clear, because the tests differ markedly in their content specifications, pretest methods, and operational equating methods. In addition, the multidimensionality problems that affected the IRT method used in the SAT studies might not affect the equipercentile method that is used in the present study. Thus, the present study provides additional research evidence on the use of item preequating.

Method

The data used in this study were the item responses of examinees to the ACT Assessment Mathematics test, which is a 40-item, five-alternative multiple-choice test. The data layout for the administration of test forms is presented in Table 1.

Six alternate forms labeled A, B, C, D, E, and F were used. In the spiraled administrations (S1, S2, S3, and S4) forms were spiraled within each test center, with the result being that randomly equivalent groups of examinees were administered the alternate forms for a particular spiraled administration. For example, randomly equivalent groups of examinees were administered forms A and B in administration S1. In the pretest

administrations P1 and P2, each examinee was administered an intact test form along with a 20-item pretest unit administered at the end of the testing period. Approximately 30 pretest units were spiraled in each pretest administration, and approximately 200 examinees were administered each item. Some of the items from these pretest units were used to construct forms E and F.

Based on the data outlined in Table 1, form E and form F were equated to form A using the random groups design and the item preequating design. An equipercentile method and an IRT method were used to conduct the equating for each design. The results of the equating process were raw-to-scaled score conversion tables for form E and form F.

Equipercentile equating using the random groups design (ER) is the current operational method for equating the ACT Assessment Mathematics test. Using this method, forms E and F were equated to form A raw scores by the equating chain of E and F to D, D to C, and C to A using data from administrations S4, S3, and S2. All equating in this chain was conducted using equipercentile equating and random groups. The method described by Kolen (1984) was used to conduct the equating and smooth the equating relationships.

Three-parameter logistic model IRT equating using the random groups design (IR) used data from administrations S1, S2, S3, and S4. LOGIST (Wingersky, Barton, & Lord, 1985) was run separately for each form/administration S1, S2, S3, and S4 combination shown in Table 1. The IRT ability (θ) scale for form A in administration S1 was considered to be the base scale, and item parameter estimates from all other forms were placed on this scale.

The LOGIST-based item parameter estimates for forms spiraled during the same administration were considered to be on the same IRT ability scale, because these forms were administered to randomly equivalent groups. To convert item parameters from one spiraled administration to another spiraled administration, the mean and standard deviations of the IRT difficulty parameter estimates were set equal for the form that was common to the two administrations (e.g., form A is common to S1 and S2). The conversion equation that resulted was used to convert the item parameter estimates from one administration to the ability scale of the other administration. (See Hambleton and Swaminathan, 1984, ch. 10 for the equations used to convert the estimates.)

These procedures were used to convert form E and form F item parameter estimates to the ability scale for form A in administration S1, based on data from administrations S1, S2, S3, and S4. These item parameter estimates then were used with IRT estimated true score equating (Lord, 1980) to estimate form A raw score equivalents of form E and form F raw scores.

Data from administrations P1 and P2 were used to equate form E and form F to form A based on equipercentile equating using the item pretest design (EP). Classical item difficulty and discrimination indices of these items for administration S1 were estimated. The resulting item statistics were then used to estimate form E and form F frequency distributions for the examinee group in administration S1. Finally, these estimated distributions were used to equate form E and form F to form A based on the examinee group in S1 using equipercentile methods.

Specifically, the item difficulties (proportion-correct) of the form E and form F items for the group of examinees in administration S1, the base administration, were estimated in the following manner. For item i pretested

in administration P1, the item difficulty observed in administration P1 is p_i. The mean item difficulty for form B items in administration P1 is $\bar{p}_{P1,B}$, and $\bar{p}_{S1,B}$ is the mean item difficulty for form B in administration S1. The difficulty of item i in administration S1 was estimated as

$$p_i^* = p_i - \bar{p}_{P1,B} + \bar{p}_{S1,B} . \quad (1)$$

Item difficulties pretested in administration P1 were converted in this manner. An analogous procedure, which involved an additional level of chaining, was followed for items pretested in administration P2.

Next, the mean and standard deviation of the observed score distributions for form E and form F were estimated for the group of examinees in S1. The mean of the form E distribution was estimated as

$$\bar{X}_E = \sum_{i:E} p_i^* , \quad (2)$$

where the summation is over form E items. The standard deviation was estimated as

$$S_E = \sum_{i:E} [p_i^*(1 - p_i^*)]^{1/2} r_i^* , \quad (3)$$

where the summation is over form E items and r_i^* is the estimated point biserial for item i (see Lord & Novick, 1968, p. 330). For Equation 3, r_i^* was taken as the point biserial correlation between the item and total score on the operational form (B or D) from administration P1 or P2. Thus, it was assumed that the point biserial for administrations P1 or P2 was an adequate estimate of the point biserial for administration S1. An analogous procedure was used for form F.

Based on the means and standard deviations from Equations 2 and 3, the E and F frequency distributions for the group of examinees in administration S1 were estimated using the negative hypergeometric distribution (Lord & Novick, 1968, pp. 515-520). The negative hypergeometric distribution was also used to estimate a raw score frequency distribution for form A in administration S1

based on the mean and standard deviation of the form A raw scores in administration S1. Equipercentile methods were used to equate the form E and form F distributions to form A.

Data from administrations P1 and P2 also were used to equate form E and form F to form A using IRT equating with the item pretesting design (IP). The first step in this equating process was to convert item parameter estimates from administrations PI and P2 to the form A scale for administration S1. Three parameter logistic model item parameters and abilities for form B were estimated using LOGIST for administration P1. Because of small sample sizes for the pretested items, the examinee abilities based on form B were fixed, and these fixed abilities were used to estimate the item parameters for the pretest items. The pretest item parameter estimates were converted to the form A ability (θ) scale in administration S1, by setting equal the mean and standard deviation of the form B item difficulty parameter estimates for administration P1 and S1. A similar procedure that involved an additional chaining step was followed to convert the item parameter estimates for items pretested in administration P2 to the form A ability scale in administration S1. Based on the resulting estimated item parameters, estimated IRT true score equating methods (Lord, 1980) were used to estimate form A raw score equivalents of form E and form F raw scores.

There were practical problems that complicated the item preequating methods. These problems included the fact that not all of the form E and form F items were pretested in administration P1 and P2. Also, one item had poorly estimated IRT pretest item parameters (very low estimated discrimination and very high estimated difficulty). As a result of these problems, only 37 of the 40 items in form E and 30 of the 40 items in form F had valid pretest-based item statistics for both pretest methods.

To conduct item preequating, the statistics for items without pretest-based item statistics (3 items on form E and 10 items on form F) were estimated from administration S4. For method EP, the S4 item difficulties for these items were converted to administration S1 values using the procedure described in conjunction with Equations 1-3, chained from S4 to S3 to S2 to S1. The administration S4 point biserials were used as the administration S1 estimates. For method IP, the item parameter estimates from method IR for the items without pretest statistics were used as the estimates for these items.

The procedures just described would not be adequate if item preequating were used operationally. However, these procedures can be used to investigate the feasibility of the item preequating design with the ACT Assessment Mathematics test using already available data.

Results

Summary statistics are shown in Table 2 for all intact forms used in the study. One notable result in this table is that the examinee groups appear to be similar from one spiraled administration to another. For example, the largest difference between means for a given form administered in two spiral administrations is .38 for form D in administrations S3 and S4. The differences between the groups in the pretest administrations and the spiral administrations are more marked. For example, the mean score for form D in administration P2 is 3.08 points lower than the mean for the same form in administration S3. These differences probably result because the pretest administrations are at times of the year different from the spiral administration.

An item statistics summary is presented in Table 3. The statistics in this table are based only on the 37 form E items and 30 form F items that were estimated in both the pretest and operational modes. The item difficulties in

this table for the spiral administration were calculated by converting the item difficulties from administration S4 to estimated values for administration S1 by a chaining process that used procedures similar to those used in Equations 1 through 3. The discrimination indices for the spiraling design summarized in Table 3 were the point biserials from administration S4.

The mean classical item difficulties shown in Table 3 are greater for the spiral design than for the item pretest design. Because the random groups design is used operationally and is based on realistic assumptions, the spiral administration methods are used as the criterion. It appears that items tend to be more difficult in pretesting than when administered operationally. A similar conclusion can be reached by inspecting the IRT difficulties, noting that IRT difficulties are inversely related to classical difficulties.

Conversion tables relating scores on form E and form F to form A are given in Tables 4 and 5. In addition, the number of examinees at each form E or form F raw score point in administration S4 is also given. To maintain comparability among methods, scores below the sum of the IRT lower asymptote parameter estimates (7 and below) and a score of all correct (40) are not shown. For both forms, the random groups conversion tables (ER and IR) appear to be similar to one another, the pretest methods conversion tables (EP and IP) appear to be similar to one another, but the results from the random groups methods differ from those for the pretest methods.

Two indices that summarize the differences between a pair of conversion tables over score points were calculated. These indices are

$$\text{BIAS} = \sum f_i (X_{1i} - X_{2i}) / \sum f_i \qquad \text{and} \tag{4}$$

$$\text{RMSD} = \{\sum f_i (X_{1i} - X_{2i})^2 / \sum f_i\}^{\frac{1}{2}} . \tag{5}$$

In these equations, X_{1i} is a form A equivalent of form E or form F integer score i using equating method 1, X_{2i} is a similar quantity for equating method 2, f_i is the frequency of form E or form F score i, and all summations are over the integer scores on form E or form F shown in Tables 4 and 5. The BIAS index is a measure of the mean difference between converted scores on two forms. The root mean-squared difference index (RMSD) is a measure of the squared difference between converted scores, and provides an index of similarity of conversion tables that is weighted by score frequency. To implement Equations 4 and 5 for comparisons with no equating (NEQ), X_{1i} is the conversion for the method to be compared and $X_{2i} = i$.

The BIAS and RMSD indices are presented in Tables 6 and 7. For implementing Equations 4 and 5 in these tables, the row method is method 1 in Equations 4 and 5 and the column method is method 2. The BIAS indices are shown above the diagonal and the RMSD are shown below the diagonal.

Recall that the random groups design methods are used as the criterion. Refer to the RMSD indices shown below the diagonal in Tables 6 and 7, where smaller index values reflect more similar conversion tables. The random groups methods ER and IR agree fairly closely in their RMSD indices. The results for the pretest methods (EP and IP) were fairly similar to one another, but they differed markedly from those for the random groups method. For both forms, even no equating (NEQ) produced results more consistent with the ER and IR results than did the IP method. Similar statements hold for the BIAS indices.

Discussion

The pretest methods (EP and IP) did not compare favorably with the random groups methods (ER and IR). The IRT pretest method (IP) introduced more estimated error into the equating process than had no equating been done.

Based on the summary statistics shown in Table 3, the operational item difficulties estimated from the pretest process (see Equation 1) suggested that the items would be more difficult operationally than they actually were and also less discriminating. The overestimate of item difficulty might have been caused by a lack of motivation or fatigue resulting from the items being pretested in a separate section at the end of the testing period.

The results of the present study in combination with the results from the SAT studies mentioned earlier suggest that the item pretest design may produce inadequate equating in many situations. The results of this study clearly indicate that item context and position effects should not be ignored; if they are ignored item preequating may perform poorly and add error into the testing process.

The IRT pretest method (IP) performed particularly poorly. The factors discussed in the SAT studies mentioned earlier, group differences and multidimensionality, might have contributed to the problems encountered with this method in the present study. In addition, small pretest sample sizes might have contributed to the problems for this method.

Although neither of the pretest-based methods performed well in this study, the equipercentile method (EP) performed more adequately than the IRT method (IP). The equipercentile pretest method (EP) might be expected to perform better than the IRT method (IP) in certain situations, because estimation of item characteristics in the EP method is more straightforward and may be more stable with small pretest sample sizes, and the assumption of unidimensionality at the item level is not required for the EP method.

Further research needs to be conducted on the item preequating design using both of the methods. One area of research could study different pretest procedures that attempt to minimize or eliminate context effects, position

effects, the effects of multidimensionality, and the effects of group differences on the item preequating results. For example, a research study could be conducted to ascertain whether or not position and context effects are minimized when pretest items are interspersed in operational sections and then approximately maintain their position when used operationally. Because the existence of position and context effects is indicative of a violation of the IRT local independence assumption, this type of study would attempt to lessen the effects on IRT equating of violating the local independence assumption. Another study could focus on finding ways to minimize the effects of group differences on item preequating results. Still another area of research could attempt to model context and position effects and incorporate this modeling into the item preequating process. In short, the present study and previous similar studies with the SAT strongly suggest that there are serious problems with item preequating, and additional research needs to be conducted and solutions found for these problems before item preequating could be suggested for use in practice.

References

American College Testing Program. (1988). ACT Assessment Program technical manual. Iowa City, IA: Author.

Donlon, T. F. (Ed.). (1984). The College Board technical handbook for the Scholastic Aptitude Test and achievement tests. New York: College Entrance Examination Board.

Eignor, D. R. (1985). An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections. (Research Report 85-10). Princeton, NJ: Educational Testing Service.

Eignor, D. R., & Stocking, M. L. (1986). An investigation of the possible causes for the inadequacy of IRT preequating. (Research Report 86-14). Princeton, NJ: Educational Testing Service.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory, principles, and applications. Boston: Kluwer-Nijhoff.

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

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Stocking, M. L., & Eignor, D. R. (1986). The impact of different ability distributions on IRT preequating. (Research Report 86-49). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST V users guide. Princeton, NJ: Educational Testing Service.

TABLE 1

Data Layout

Administration	Alternate forms
S1 (spiraled)	A B
S2 (spiraled)	A C
S3 (spiraled)	C D
P1 (pretest)	B (+ pretest units)
P2 (pretest)	D (+ pretest units)
S4 (spiraled)	D E F

TABLE 2

Test Score Moments for 40-Item
ACT Assessment Mathematics Forms

Administration	Form	N	Mean	Standard Deviation	Skewness	Kurtosis
S1	A	3074	20.26	7.88	.29	2.32
	B	3025	20.00	8.16	.22	2.17
S2	A	3753	20.07	7.89	.27	2.36
	C	3708	21.28	8.67	.21	2.14
S3	C	2438	21.45	8.45	.15	2.20
	D	2442	21.39	8.73	.27	2.17
P1	B	5155	19.22	8.68	.34	2.23
P2	D	5708	18.31	8.66	.42	2.22
S4	D	2664	21.01	8.51	.29	2.17
	E	2738	21.43	7.95	.27	2.36
	F	2712	20.87	7.95	.24	2.25

TABLE 3

Item Statistics Summary for Spiral and Pretest Designs for Form E (37 Items) and Form F (30 Items)

Item Statistic	Form	Spiral Mean	Spiral SD	Pretest Mean	Pretest SD	Spiral/Pretest Correlation
Classical						
Difficulty	E	.56	.16	.52	.13	.95
	F	.57	.15	.55	.12	.83
Discrimination	E	.42	.06	.38	.09	.57
	F	.42	.08	.39	.07	.41
IRT						
Difficulty	E	.04	.77	.43	.76	.85
	F	-.04	.88	.35	.60	.77
Discrimination	E	.86	.22	1.00	.41	.57
	F	.84	.27	.99	.36	.45
Pseudo-chance	E	.16	.07	.18	.08	.23
	F	.15	.09	.21	.10	.59

TABLE 4

Tables for Converting Form E Raw Scores to Raw Score Scale of Form A

Score	Frequency	ER	IR	EP	IP
39	25	37.25	38.70	39.34	39.11
38	27	35.76	37.44	38.43	38.16
37	36	34.48	36.19	37.50	37.20
36	46	33.27	34.96	36.57	36.23
35	45	32.07	33.76	35.61	35.27
34	44	31.11	32.58	34.62	34.32
33	59	30.08	31.45	33.62	33.38
32	56	29.23	30.35	32.60	32.45
31	63	28.49	29.29	31.57	31.53
30	54	27.75	28.26	30.52	30.63
29	82	26.98	27.25	29.48	29.74
28	81	26.16	26.25	28.43	28.85
27	74	25.33	25.27	27.37	27.97
26	97	24.54	24.29	26.31	27.08
25	114	23.66	23.32	25.24	26.19
24	122	22.67	22.34	24.17	25.29
23	125	21.57	21.36	23.09	24.37
22	132	20.54	20.38	22.01	23.43
21	114	19.52	19.39	20.94	22.48
20	129	18.54	18.40	19.86	21.50
19	136	17.60	17.40	18.78	20.49
18	133	16.68	16.41	17.71	19.44
17	94	15.74	15.43	16.64	18.36
16	109	14.73	14.46	15.57	17.24
15	126	13.78	13.52	14.50	16.08
14	125	12.85	12.59	13.44	14.88
13	91	11.86	11.69	12.38	13.65
12	94	10.84	10.81	11.33	12.39
11	69	9.95	9.96	10.28	11.10
10	63	9.03	9.13	9.25	9.79
9	59	7.88	8.32	8.22	8.48
8	49	6.72	7.50	7.20	7.16

TABLE 5

Tables for Converting Form F Raw Scores to Raw Score Scale of Form A

Score	Frequency	ER	IR	EP	IP
39	10	38.33	39.31	39.22	39.24
38	36	36.99	38.31	38.28	38.31
37	26	35.58	37.17	37.31	37.28
36	25	34.31	35.96	36.33	36.21
35	39	33.11	34.72	35.34	35.11
34	50	31.92	33.46	34.33	34.01
33	61	30.90	32.21	33.31	32.92
32	63	29.83	30.98	32.28	31.86
31	60	28.97	29.79	31.25	30.83
30	89	28.16	28.64	30.21	29.83
29	64	27.40	27.53	29.16	28.87
28	88	26.52	26.47	28.11	27.93
27	83	25.63	25.43	27.05	27.02
26	83	24.81	24.43	26.00	26.12
25	110	24.00	23.45	24.94	25.23
24	107	23.04	22.49	23.87	24.35
23	104	21.96	21.54	22.81	23.46
22	110	21.00	20.60	21.75	22.56
21	103	20.00	19.67	20.69	21.64
20	124	19.10	18.74	19.63	20.71
19	107	18.17	17.81	18.57	19.75
18	129	17.27	16.88	17.51	18.77
17	129	16.40	15.96	16.46	17.75
16	123	15.45	15.03	15.40	16.70
15	113	14.45	14.10	14.35	15.61
14	107	13.47	13.17	13.31	14.49
13	121	12.48	12.23	12.27	13.33
12	110	11.42	11.29	11.24	12.13
11	88	10.41	10.35	10.21	10.88
10	77	9.45	9.40	9.19	9.59
9	72	8.35	8.45	8.18	8.21
8	34	7.22	7.49	7.18	6.61

TABLE 6

Bias and MSE Indices for Form E

	ER	IR	EP	IP	NEQ
ER	--	-0.11	-1.51	-2.47	-1.57
IR	0.60	--	-1.40	-2.36	-1.46
EP	1.78	1.54	--	-0.96	-0.06
IP	2.56	2.51	1.18	--	0.90
NEQ	1.66	1.50	0.45	1.07	--

Note: BIAS index is above diagonal and RMSD index is below diagonal.

TABLE 7

Bias and RMSD Indices for Form F

	ER	IR	EP	IP	NEQ
ER	--	0.04	-0.72	-1.28	-1.01
IR	0.60	--	-0.76	-1.32	-1.06
EP	1.12	0.99	--	-0.56	-0.30
IP	1.38	1.48	0.83	--	0.27
NEQ	1.13	1.14	0.48	0.50	--

Note: BIAS index is above diagonal and RMSD index is below diagonal.