Using the Circular Equating Paradigm for Comparison of Linear Equating Models

Naomi Gafni and Estela Melamed
National Institute for Testing and Evaluation

Equating error was estimated using the same test by three linear equating methods in three paradigms: (1) single-link equating of a test to itself, in which a test was administered on two different dates and the later administration was equated to the earlier administration; (2) circular equating through a chain, starting and ending at the same test; and (3) pseudo-circular equating, in which a test was equated to itself as in the first approach through equating chains containing a different number of links as in the second approach. The mean difference between the actual scores and the equated scores, as well as the root mean square of this difference, were used as the criterion measures for equating error. The results suggested a superiority of the Tucker method for the conventional circular equating chain, and the Levine and VCI methods yielded smaller errors in about half the equating chains for the pseudo-circular chain. Unexpectedly, there was not found to be a clear relationship between the number of links in the equating chain and the resulting error. *Index terms: circular equating, equating chains, equating error, equating methods, linear equating.*

The purpose of equating is to express a score on Test X (administered to population α) as an equivalent score on Test Y (administered to population β), where Tests X and Y are assumed to be similar in content and difficulty. One of the most common equating designs consists of using a common set of items (anchor items, designated U) that is included in both test forms. The anchor items are selected to represent the two forms both in their content and difficulty.

Two linear equating methods are often used with this design: the Tucker and the Levine Equally Reliable Tests methods (Angoff, 1971). These methods equate a score on Test X to a score on Test Y by a linear transformation, based on estimation of the mean and standard deviation of each of the tests for a synthetic population assumed to take both. The estimation process requires the assumption that the linear regression of total score on common item score is equal for populations α and β. The Tucker method assumes this equality for observed scores and is usually indicated for populations similar in ability, whereas the Levine method assumes this equality for true scores and is usually indicated for nonsimilar populations.

Research findings do not provide unambiguous evidence, however, for the adequacy of using the two methods in the two contexts. Kolen and Brennan (1987) showed that differences between populations α and β have a greater influence on the Levine equating method than on the Tucker equating method. One method is not necessarily preferable to the other; rather, the degree of proximity of the equated scores relative to the base group scores is reflected in these results. These results also do not provide any indication of the size of equating error.

A problem related to evaluating the results of any equating procedure concerns the choice of criterion measure. In some situations, the test is equated to itself; this is done either by single-link equating, in which a certain form of a test is administered on two different dates and treated as two different forms, or by circular-chain equating, in which a test is equated through a number of links back to itself (e.g., X→A→B→C→X). The advantage of these paradigms is that the criterion is known and the error is easily estimated (Angoff, 1987). Brennan and Kolen (1987a, 1987b) showed that the circular paradigm favors

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 3, September 1990, pp. 247-256
© Copyright 1990 Applied Psychological Measurement Inc.
0146-6216/90/030247-10$1.75

equating methods involving fewer moments, rather than more; they therefore suggested using this paradigm for comparisons of equating procedures involving the same number of moments (e.g., two different linear equating procedures, as opposed to linear equating versus equipercentile procedures).

Brennan and Kolen (1987a) claimed that "it would seem sensible that the degree of confidence in the stability of equating should be inversely related to the number of equating links necessary to progress from the new form to the initial form" (p. 281). It is not clear whether this holds true for every case. Because it is known that equating error is influenced by the degree of difference between the two populations involved in the equating process (Petersen, Marco, & Stewart, 1982), it might be possible that a longer chain involving similar populations would yield smaller error than a shorter chain involving dissimilar populations.

It is also possible that in a shorter chain the errors would not cancel one another, and an additional link might decrease the error. Error in a circular paradigm contains random error, model misspecification error (error deriving from the choice of the equating procedure), error that accumulates from multiple equating, and error related to the different dates on which the tests were taken (e.g., error involving differences in meaning the anchor items might have had at different times). It is thus possible that random error increases as the chain becomes longer, while other types of error decrease, which might yield an altogether smaller error.

An alternative design to evaluate the quality of equating would consist of a chain in which the initial form is identical to the final form, but the initial population is different from the final population. This could be done by administering the initial Test X twice on two different dates, or by dividing the population taking the initial test into two halves, and then equating the test administered to the first half of the population to the same test administered to the second half. This is referred to below as a *pseudo-circular design*.

The objective of this study was to compare empirically the equating error yielded by using the circular paradigm (in which the initial test and population were identical to the final test and population) with the error yielded by using the pseudo-circular paradigm (in which a test was equated to itself using two different populations through a chain). Error accumulated over multiple equating was compared to the error associated with single-link equating and an estimate of the error yielded by chains of different lengths.

Method

A comparison between the error obtained using single-link equating and using circular equating should provide an estimate of the error accumulated through multiple equating. A substantial difference between error estimates obtained by an equating chain with initial and final forms identical, and by an equating chain with initial and final forms identical but administered to different populations, should indicate an effect related to some interaction of operation of the equating method, the paradigm, and possibly the specific dataset characteristics.

In order to examine these two effects (multiple equating and the interaction of equating method with the paradigm), equating error was estimated using the same test in three paradigms:

1. Single-link equating of a test to itself, in which a test was administered to two groups on two different dates, and the later administration was equated to the earlier one (Paradigm 1);
2. Circular equating chain, starting and ending at the same test administered at the same date (Paradigm 2);
3. Pseudo-circular equating, in which a test is equated to itself (as in single-link equating) through equating chains containing different numbers of links (as in the circular equating chain; Paradigm 3).

Instruments

The test scores in this study were based on various forms of the Psychometric Entrance Test

(PET) constructed by the National Institute for Testing and Evaluation (NITE) in Israel and written in Hebrew (Beller, in press). The test is used in undergraduate admissions by all the universities in Israel. It consists of five subtests; the length of each depends on the specific form:

General Knowledge (GK): 45 to 60 items,

Figural Reasoning (FO): 22 to 27 items,

Verbal Reasoning (RE): 40 items,

Mathematical Reasoning (MA): 30 to 35 items, and

English (EN): 50 items.

Each subtest is equated separately to two previous forms of the same subtest, based on anchor items that constitute about 20% of the subtest length. Scores on each subtest are reported on a scale with mean of 100 and standard deviation of 20. The total score on the battery (PET) is given by a linear transformation, with the subtests weighted equally, on a scale with mean of 500 and standard deviation of 100.

Form 7 of the test administered in April 1985 was used as the base form (Y) in all the paradigms; it was also used as the final form in Paradigm 2. The same form was administered again in June 1985 and was used as the final form (X) in the first and third paradigms. In addition to Form 7, the equating chains used in Paradigms 2 and 3 included the following forms: Form 10 (December 1986), Form 20 (December 1987), Form 17 (April 1987), Form 13 (June 1986), Form 4 (June 1984), Form 5 (December 1985), and Form 2 (February 1984).

Population

The examinees were all PET candidates who registered for examinations for February and June 1984, April, June, and December 1985, June and December 1986, and April and December 1987. Table 1 presents the mean and standard deviation of the raw and standard scores on the five subtests, and on the PET general score for the nine groups included in the study. The standard scores were obtained by equating each form to two previously-administered forms. The same parameters used for Form 7, as it was administered in April 85, were used for Form 7 administered in June 85.

Procedure

Figure 1 shows the five equating chains used in Paradigms 2 and 3. The arrows in each chain indicate the direction of the equating; thus, in chain 1, Form 7 (the final form) was equated to Form 17, which was equated to Form 20, which was equated to Form 10, which was equated to Form 7 (the initial form). The initial and final form were the same in Paradigm 2 (Form 7, administered in April), and in Paradigm 3 the final form was Form 7 administered in June. For example, Chain 1 for Paradigm 3 is described as illustrated in Figure 2.

Table 2 presents the number of anchor items, the mean and the standard deviation of the anchor item scores, and their correlation with the test scores for the base group and the equating group within each of the equating links in each of the chains. The anchor-test correlations for FO and RE are generally lower than for the other three tests, probably due to the smaller number of items in FO and to the relatively higher degree of heterogeneity of RE.

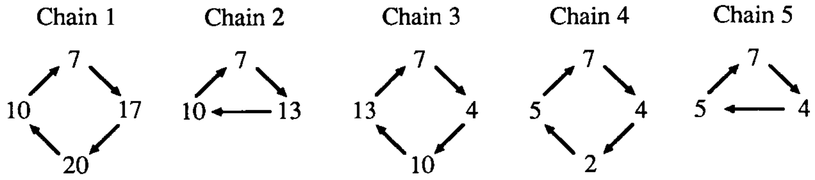
Three linear equating methods were used: the Tucker method, the Levine method, and an additional method described by Angoff (1971) as "Design V: other methods involving score data" (denoted here as VC1). In this method, a score x_1 on Form X, and a score y_1 on Form Y are found such that they predict linearly the same score on the anchor test U; the scores x_1 and y_1 are then considered equated. This method was examined because it was shown to be the most valid linear method in the prediction of grade-point averages obtained at the Hebrew University in Jerusalem from psychometric entrance scores (Melamed, 1989).

Criterion measure. Form 7 served as the initial and the final form in the process of equat-

Table 1
Number-Correct and Standardized Score Mean (M) and Standard Deviation (SD) of the Nine Samples for the Five Subtests and PET

	Form																	
	7 (April) (*N* = 6484)		10 (*N* = 2060)		20 (*N* = 1947)		17 (*N* = 6659)		13 (*N* = 3225)		5 (*N* = 1543)		2 (*N* = 3648)		4 (*N* = 4596)		7 (June) (*N* = 1215)	
ubtest	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
umber-Correct Score																		
K	25.91	8.54	27.11	8.37	27.16	8.06	25.92	8.07	25.98	8.24	26.69	8.34	37.68	10.40	31.98	11.01	24.87	9.10
O	14.98	4.52	15.88	5.30	16.99	4.83	15.54	4.87	13.74	4.71	17.40	4.72	12.83	4.00	10.64	4.02	13.67	4.43
E	24.18	6.92	23.16	7.22	24.01	6.70	23.09	6.41	20.06	6.11	22.07	6.82	21.73	5.83	18.38	5.88	21.63	6.93
A	19.42	6.07	21.65	7.44	22.20	6.94	21.91	6.94	17.31	6.29	18.48	6.56	17.42	6.25	13.47	6.06	17.73	5.99
N	27.02	9.88	26.04	9.88	31.33	9.72	27.98	9.96	21.72	9.10	28.45	10.49	28.39	10.08	21.94	9.56	24.27	10.63
andardized Score																		
K	102.20	18.82	101.70	18.84	102.04	18.06	101.40	18.37	98.35	19.34	104.20	18.72	100.80	17.53	94.87	19.27	99.92	20.02
O	104.60	18.09	107.30	19.38	109.96	17.74	106.30	17.41	99.21	17.45	105.00	19.27	103.93	20.00	92.00	18.53	99.69	17.70
E	103.50	18.92	103.30	18.80	107.51	18.16	105.30	18.23	96.00	18.06	103.20	20.46	100.87	19.47	91.87	18.67	96.54	18.97
A	104.10	19.49	107.10	20.32	108.49	19.32	105.90	19.41	95.59	18.42	104.10	18.49	102.07	17.33	92.40	16.67	98.67	19.26
N	100.30	19.38	103.40	20.18	108.49	18.87	103.40	18.81	94.62	18.48	104.30	19.54	100.07	17.00	90.87	16.13	94.99	20.80
ET	518.40	92.90	528.00	98.07	545.81	90.99	528.20	91.69	479.47	89.84	526.10	95.40	508.56	85.25	453.96	82.61	487.10	95.75

Figure 1
Five Equating Chains Used in the Study



ing in all three paradigms; the difference between the actual score and the equated score (BIAS) could serve as an estimate of equating error. Because mean BIAS can be small while its variance can be fairly large, BIAS by itself is not sufficient. An additional measure that takes into consideration both mean BIAS and its variance is the root mean square error (RMSE), which is the square root of the sum of the squared mean BIAS and its variance.

Results and Discussion

Tables 3 and 4 present means of BIAS and RMSE, respectively, for the five subtests and PET scores using single-link (SL) equating, the five chains of Paradigm 2, and the five chains of Paradigm 3. A comparison of the two tables shows that the absolute size of the two measures is fairly similar, which indicates small BIAS variances; they can therefore be used interchangeably. The sign of the BIAS has not been considered because it is affected by the direction of the equating chain.

Paradigm 2 Versus Paradigm 3

Findings based on the five subtests are expected to be consistent with each other because they actually serve as five replications of the same design. Table 4 reveals that although the Tucker method yielded smaller RMSEs for almost every case using Paradigm 2, this was not true using Paradigm 3—in 13 out of 25 cases (five subtests × five chains), the Levine method yielded smaller RMSEs than the Tucker method. Moreover, in the three cases in which the Levine method yielded smaller RMSEs using Paradigm 2, the corresponding RMSEs for the Levine method using Paradigm 3 were also smaller than for the Tucker method.

This result throws some doubt on the use of the conventional circular paradigm for comparison of equating methods with the same number of moments. It indicates an interaction of equating method with the paradigm used. As discussed above, the Levine method, as well as the VCI method, emphasizes the difference between groups in single-link equating more than the Tucker method does. This property does not necessarily hold mathematically for a chain involving more than two groups. It might be, however, that this attribute, along with group characteristics and the differences among them involved in real equating situations, results in smaller error estimates overall for the Tucker method when it is used in the circular paradigm (Paradigm 2) than when it is used in the pseudo-circular paradigm (Paradigm 3).

Because the chains in Paradigm 3 contain an additional population (June, Form 7) to those contained in Paradigm 2, larger RMSEs were expected using Paradigm 3. Such an expectation is reasonable, even if the additional population had a similar ability distribution, due to an additional source of noise (i.e., sampling error). In particular, such an increase in error is expected when the additional population has a quite different ability distribution, as in the present case. However, the current findings do not support this expec-

Figure 2
Chain 1 as Used in Paradigm 3

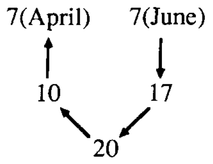


Table 2
Number of Items (*n*), Mean and Standard Deviation of the Anchor Items Within the Base and Equating Groups, and the Correlation (*r*) of the Anchor Items With the Test for Both Groups (Pairs of Groups are Presented by the Order of Their Appearance in the Chains; Pairs That are Identical Except for Their Direction Appear Only Once)

		Base Group			Equating Group				Base Group			Equating Group		
Subtest	*n*	Mean	SD	*r*	Mean	SD	*r*	*n*	Mean	SD	*r*	Mean	SD	*r*
10 → 7 (April)								4 → 10						
GK	10	5.70	2.33	.87	5.68	2.29	.85	9	5.22	2.06	.81	5.43	1.96	.82
FO	6	3.73	1.47	.80	3.59	1.38	.76	6	2.91	1.43	.78	3.77	1.52	.81
RE	8	5.04	2.00	.79	4.93	1.93	.79	8	5.12	1.69	.67	5.61	1.67	.74
MA	7	4.57	1.77	.84	4.29	1.72	.82	7	3.40	1.74	.82	4.43	1.91	.86
EN	12	6.64	3.17	.91	6.29	2.98	.89	12	5.32	2.89	.88	6.70	3.15	.90
20 → 10								7 (April) → 4						
GK	9	5.20	1.99	.82	5.35	2.10	.84	9	5.52	2.09	.83	5.05	2.12	.83
FO	6	4.02	1.43	.79	3.69	1.55	.82	6	3.39	1.53	.77	2.98	1.55	.78
RE	8	5.62	1.79	.78	5.08	1.93	.80	8	4.82	1.92	.74	3.97	1.89	.74
MA	7	4.77	1.83	.84	4.59	1.91	.84	7	4.47	1.85	.85	3.59	1.81	.84
EN	11	7.21	2.52	.88	6.34	2.71	.88	13	6.84	3.35	.90	5.80	3.20	.90
17 → 20								7 (June) → 4						
GK	9	5.27	2.11	.83	5.54	2.04	.82	9	5.16	2.25	.85	5.05	2.12	.83
FO	6	3.51	1.46	.76	3.87	1.43	.76	6	3.10	1.53	.75	2.98	1.55	.78
RE	8	5.13	1.79	.74	5.39	1.83	.76	8	4.33	1.98	.75	3.97	1.89	.74
MA	7	4.59	1.84	.86	4.85	1.80	.85	7	3.98	1.82	.85	3.59	1.81	.84
EN	11	6.45	2.71	.89	7.18	2.68	.90	13	6.05	3.37	.91	5.80	3.20	.90
(April) → 17								5 → 7 (April)						
GK	9	5.32	2.11	.83	5.07	2.13	.83	9	5.60	2.07	.85	5.35	2.06	.83
FO	6	3.60	1.52	.79	3.75	1.55	.81	6	3.87	1.46	.79	3.80	1.45	.79
RE	9	5.72	2.09	.77	5.89	1.99	.75	8	4.83	2.02	.76	4.74	2.01	.77
MA	7	4.37	1.79	.85	4.65	1.77	.84	7	4.27	1.78	.84	4.36	1.74	.83
EN	11	5.96	2.87	.88	6.31	2.88	.89	18	10.31	4.71	.94	9.59	4.55	.93
7 (June) → 17								2 → 5						
GK	9	5.11	2.19	.84	5.07	2.13	.83	9	5.24	2.02	.81	5.19	2.13	.83
FO	6	3.29	1.52	.79	3.75	1.55	.81	6	4.29	1.44	.83	4.09	1.47	.80
RE	9	5.13	2.22	.78	5.89	1.99	.75	8	4.00	1.76	.71	3.42	2.01	.75
MA	7	3.92	1.79	.85	4.65	1.77	.84	6	3.66	1.58	.81	3.55	1.60	.80
EN	11	5.26	2.92	.89	6.31	2.88	.89	10	6.22	2.29	.84	6.07	2.35	.87
13 → 10								4 → 2						
GK	9	5.32	2.01	.79	5.43	1.95	.79	15	8.13	3.17	.89	8.97	2.96	.88
FO	10	5.26	2.17	.87	6.18	2.38	.88	7	3.41	1.64	.80	4.32	1.72	.85
RE	8	4.23	1.73	.70	4.65	1.85	.76	9	3.06	1.93	.75	4.08	2.14	.77
MA	8	3.80	2.00	.83	4.92	2.19	.88	8	3.53	1.93	.84	4.51	2.02	.87
EN	13	6.00	2.89	.87	7.06	3.21	.90	13	5.20	2.86	.88	6.57	3.01	.88
7 (April) → 13								4 → 5						
GK	10	5.80	2.35	.85	5.21	2.36	.85	9	4.89	2.20	.84	5.48	2.13	.83
FO	9	5.73	1.97	.84	5.24	1.97	.84	6	3.04	1.42	.77	3.76	1.43	.79
RE	8	5.16	1.96	.78	4.25	1.97	.76	8	4.08	1.71	.70	4.82	1.82	.77
MA	8	4.79	1.97	.86	4.07	1.91	.84	6	2.97	1.66	.83	3.75	1.68	.82
EN	12	6.68	3.01	.89	5.68	2.97	.89	10	4.69	2.55	.86	5.81	2.75	.89
7 (June) → 13														
GK	10	5.44	2.47	.87	5.21	2.36	.85							
FO	9	5.29	2.03	.85	5.24	1.97	.84							
RE	8	4.55	2.01	.79	4.25	1.97	.76							
MA	8	4.27	2.01	.86	4.07	1.91	.84							
EN	12	5.90	3.13	.91	5.68	2.97	.89							

Table 3
Mean BIAS Using Paradigms 2 and 3, the Three Equating Methods for the Five Chains, and Single-Link Equating (SL)

	Paradigm 2			Paradigm 3		
Chain	Tucker	Levine	VC1	Tucker	Levine	VC1
GK						
SL				-.48	-.25	-.23
1	1.06	1.38	1.39	.27	1.33	1.32
2	-3.01	-3.91	-3.93	-3.49	-3.61	-3.62
3	-.17	.27	-.24	.13	3.23	2.25
4	-3.90	-4.43	-4.58	-3.55	-1.72	-2.16
5	-1.65	-2.24	-1.89	-1.31	.64	.67
FO						
SL				-.40	.13	.17
1	.15	.68	.43	-2.35	-.08	-.27
2	1.25	.97	1.09	-.89	.39	.66
3	.63	4.10	3.70	-1.77	3.11	.01
4	3.03	4.11	4.30	.29	2.55	3.10
5	3.05	5.43	5.02	.53	4.35	4.08
RE						
SL				-1.25	-.51	-.31
1	-2.15	-2.63	-2.56	-5.24	-4.22	-3.49
2	-4.16	-4.86	-5.17	-6.58	-5.60	-5.40
3	.23	-.28	2.37	-2.41	-.39	2.37
4	-1.55	-1.67	-2.52	-5.42	-4.86	-5.14
5	-1.17	-.36	.01	-4.37	-2.15	-1.53
MA						
SL				-.61	-.11	-.11
1	1.39	1.39	1.55	.02	1.89	2.17
2	.00	-.26	-.02	-1.30	.26	.42
3	-3.35	-2.02	-2.56	-4.51	-1.60	-1.94
4	-.16	.37	-6.10	-1.32	.84	-4.78
5	-.64	.17	-6.41	-1.61	1.03	-4.66
EN						
SL				-.47	-.20	-.14
1	-1.20	-1.67	-1.64	-2.33	-1.80	-1.66
2	-1.35	-1.31	-1.26	-2.09	-1.31	-1.20
3	1.97	2.26	2.19	1.00	2.14	2.28
4	-1.61	-1.33	-1.63	-2.51	-1.57	-1.72
5	-1.34	-.77	-.68	-2.24	-1.02	-.83
PET						
SL				-10.78	-2.21	-1.14
1	-.47	-.60	-.57	-11.64	-3.13	-1.94
2	-8.67	-11.33	-11.21	-17.55	-11.94	-11.03
3	-.39	5.93	7.33	-9.01	8.65	6.75
4	-4.80	-3.25	-12.78	-15.25	-5.52	-12.97
5	-1.73	3.28	-4.49	-10.86	4.07	-2.38

tation. An attempt was made to explain the difference in the results using Paradigms 2 and 3, in terms of the difference between the final group (June, Form 7, as opposed to April, Form 7) in each paradigm and the group immediately preceding it. This attempt failed to support the hypothesis that the larger this difference was, the larger the error. No consistent pattern could be

Table 4
Mean RMSE Using Paradigms 2 and 3, the Three Equating Methods for the Five Chains, and Single-Link Equating (SL)

	Paradigm 2			Paradigm 3		
Chain	Tucker	Levine	VC1	Tucker	Levine	VC1
GK						
SL				.48	.60	.49
1	1.07	1.46	1.45	.96	1.49	1.47
2	3.06	4.09	4.13	3.48	3.85	3.89
3	.22	.38	.25	.16	3.43	2.29
4	3.91	4.43	4.69	3.55	1.90	2.32
5	1.67	2.29	2.08	1.31	1.54	1.62
FO						
SL				.41	.17	.17
1	.30	.71	.76	2.36	.21	.32
2	1.37	.99	1.09	.91	.93	1.24
3	.78	4.10	3.71	1.96	3.13	.37
4	3.15	4.59	4.68	.55	3.03	3.61
5	3.06	5.46	5.11	.58	4.35	4.19
RE						
SL				1.25	.51	.33
1	2.16	2.67	2.70	5.26	4.36	3.55
2	4.18	4.99	5.45	6.58	5.63	5.59
3	1.06	2.97	2.66	2.75	3.95	3.16
4	2.67	5.11	5.64	5.75	6.43	6.81
5	1.24	1.43	2.55	4.37	2.24	2.41
MA						
SL				.63	.28	.16
1	1.41	1.46	1.82	.45	2.02	2.44
2	.46	.69	.58	1.30	.29	.42
3	3.35	2.08	2.59	4.51	1.75	2.04
4	.20	.46	6.41	1.32	.96	5.05
5	.86	1.19	7.30	1.72	1.39	5.61
EN						
SL				.82	.67	.63
1	1.20	1.68	1.68	2.62	2.20	2.12
2	1.39	1.47	1.49	2.42	1.81	1.75
3	2.09	2.53	2.55	1.17	2.15	2.29
4	1.61	1.34	1.64	2.78	2.12	2.09
5	1.35	.77	.70	2.51	1.66	1.52
PET						
SL				10.90	3.03	1.82
1	1.00	1.40	1.81	11.81	3.74	2.79
2	8.83	11.57	11.55	17.62	12.06	11.22
3	2.58	7.48	7.76	9.29	10.33	7.28
4	5.90	8.87	14.24	15.74	9.46	14.41
5	2.04	3.67	5.62	10.95	4.47	4.01

detected, suggesting an interaction of the datasets with the different subtests.

The PET general score is a linear composite of the five subtest scores; thus an examination of the equating error for that score is not theoretically appropriate. However, the error in that equation was of interest because that score is operationally used for selection. Results for the

PET general score showed that the errors yielded by Paradigm 3 were larger than those yielded by Paradigm 2 for all chains and for all methods.

The question remaining is why the error accompanying the PET score was so much larger for the Tucker method when using Paradigm 3. This was probably due to an accumulating effect of the BIAS, which was more consistent in its direction for the Tucker method than for the other two methods. For example, all five BIAS estimates related to the subtests in Chain 2 were negative for the Tucker method, but only three of the five were negative for the Levine or the VCI methods. Thus, the errors yielded by these two methods tended to cancel each other by yielding an overall smaller error estimate, while the errors yielded by the Tucker method (having the same sign) accumulated over the five subtests.

It is not clear if such a consistency is an advantage or a disadvantage for an equating method. Lawrence and Dorans (1988) used consistency of equating results as a criterion for comparing various equating methods. They found that results based on the Tucker method were consistent across representative conditions and matched-sample conditions. Because the PET score was the one used for selection in this study, consistency of the error direction across the subtest scores was a disadvantage of the method; however, inconsistency of the error direction was an advantage, as occurred in the case of the Levine method.

Equating Stability and Equating Chain Length

It was expected that the longer the equating chain, the larger the error involved. An examination of the error relating to the five subtests in Table 4 shows that this was not necessarily true. A comparison of the error yielded by the three equating methods using Paradigms 2 or 3, with the error yielded using Paradigm 1, revealed that single-link equating generally yielded a smaller error than the error yielded by longer equating chains. A comparison of the errors yielded by the different chains, however, did not show a consistent pattern. Thus, Chains 2 and 3 were identical, except for an additional form (Form 4) in Chain 3, but the error for the longer chain (Chain 3) was sometimes smaller for the three equating methods (e.g., RMSEs for GK and RE). Generally, the longer chains were not necessarily accompanied by larger errors. The inconsistency increased when examining PET, but again there did not seem to be a relationship between chain length and the resulting error.

One of the factors that was considered to have a possible effect on the size of equating error was the degree of variability in test length in the different chains. It has been shown (Gafni & Melamed, 1989) that for single-link equating, the error resulting when equating tests of different length (e.g., two FO forms containing, respectively, 27 and 22 items) can be much larger than when equating tests of the same length. However, in the present study there was no consistency found with respect to the size of equating error as related to variability in length of the tests included in the different chains. For example, all forms of GK in Chains 1 and 2 were of identical length; yet, the equating error for Chain 2 was larger than for Chain 1 (although Chain 2 was a shorter chain). Chain 3, which included one form of 60 items and three forms of 45 items each, generally yielded smaller errors.

A more consistent relationship between variability in test length and error size was found for FO. A comparison of the equating error in chains of the same number of links (i.e., Chain 2 with Chain 5, and Chain 1 with Chains 3 and 4), which consisted of forms of identical lengths (e.g., Chain 2, 27 items), as opposed to varying test lengths (e.g., Chain 5, 22, and 27 items), showed that less variability in test length was related to a smaller error.

It seems that many factors operate simultaneously in determining the amount of error, among which are the equating method, the type of test, the number of items, the relationship between the test and its anchor for each group involved, the date of administration, and the characteristics of the groups and their differences.

Conclusions

The results of this study suggest that the use of the conventional circular paradigm is not always appropriate for the estimation of equating error. Although the Levine and VCI methods tended to yield smaller errors using Paradigm 3, the Tucker method usually yielded smaller errors using the same paradigm. These differences could not clearly be accounted for by factors such as group differences, anchor item characteristics, or subtest characteristics. The Tucker method seemed to be more consistent than the other two methods across the different subtests in terms of the direction of error, and this resulted in a larger cumulative error overall for the general score on PET.

The belief that equating error grows as the number of links in the chain increases was only partially supported by the current findings. Although larger errors were yielded for Paradigms 2 and 3 than for single-link equating, within each paradigm a longer chain was not necessarily accompanied by a larger error.

Some of the larger error estimates are probably an overestimation of the real error, because equating of a test is usually based on two previous forms and not only one. However, it is strongly recommended that any means possible be used to minimize equating error: lengthen the anchor test, select items that are a good representation of test content in its entirety, keep test length stable, or select populations as similar as possible for equating. The quality of equating depends on a plethora of factors, each of which can contribute to error; hence, quality checks should be performed at frequent intervals in order to avoid large drifts and surprising results.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington DC: American Council on Education.

Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291-300.

Beller, M. (in press). Psychometric issues in admission procedures of Israeli universities. *Educational Measurement: Issues and Practice.*

Brennan, R. L., & Kolen, M. J. (1987a). Some practical issues in equating. *Applied Psychological Measurement, 11,* 279-290.

Brennan, R. L., & Kolen, M. J. (1987b). Reply to Angoff. *Applied Psychological Measurement, 11,* 301-306.

Gafni, N., & Melamed, E. N. (1989). *Equating error as a function of population properties, type of anchor test, and equating method* (RR-104). Jerusalem, Israel: National Institute for Testing and Evaluation. [In Hebrew].

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11,* 263-277.

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (RR-88-23). Princeton NJ: Educational Testing Service.

Melamed, E. N. (1989). *Predictive validity of the psychometric entrance examination under different equating methods.* Unpublished master's thesis, Hebrew University, Jerusalem, Israel. [In Hebrew].

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71-135). New York: Academic.

Acknowledgments

The authors thank the editor and two anonymous reviewers for their helpful and insightful comments on an earlier version of this paper. They are also grateful to Jerry Levinson for his assistance in preparing this paper.

Author's Address

Send requests for reprints or further information to Naomi Gafni, 4 Rav Ashi St. Jerusalem 93593, Israel.