A Reply to Angoff

Robert L. Brennan and Michael J. Kolen
American College Testing Program

We would like to thank Angoff (1987) for his thoughtful and extensive review of the Kolen and Brennan (1987) and Brennan and Kolen (1987) papers. His comments were very helpful to us in clarifying our thinking about a number of issues. Although we find ourselves in agreement with most of his comments, there are two issues that we believe merit further consideration—synthetic population weights and the circular equating paradigm. In retrospect, our initial discussion of these topics probably should have been more extensive. We hope that the following reply will clarify our position with respect to these two issues.

Synthetic Population Weights

The Tucker and Levine linear equating equations that are summarized in Table 3 of Kolen and Brennan (1987) employ the weights w_1 and w_2 for the new and old populations, respectively. The only constraints on these weights are $w_1 \geq 0$, $w_2 \geq 0$, and $w_1 + w_2 = 1$. As such, these weights define the "synthetic population," a term introduced by Braun and Holland (1982, p. 21). For example, if $w_1 = w_2 = .5$, then the synthetic group consists of an equal weighting of the two populations, regardless of the actual size of either population or the size of the equating samples from either population. By contrast, if n_1 and n_2 are the equating sample sizes, then using $w_1 = n_1/(n_1 + n_2)$ and $w_2 = n_2/(n_1 + n_2)$ means that the two populations are weighted relative to the equating sample sizes.

Although our basic equations are for the general case ($w_1 \geq 0$, $w_2 \geq 0$, and $w_1 + w_2 = 1$), we suggest that using $w_1 = 1$ and $w_2 = 0$ often leads to the most simple and direct score interpretations. We state that this is especially true when only one form is administered on any test date, a form is retired immediately after use, and the focus of score interpretation typically is on the group that just took the new form. (In effect, the equations provided by Gulliksen, 1950, pp. 299–304, also apply to this case.) We do *not* suggest routine use of $w_1 = 1$ and $w_2 = 0$, but we do recommend that practitioners consider this weighting scheme more often than is currently the case.

It seems clear, however, that Angoff sees little merit in using $w_1 = 1$ and $w_2 = 0$. He points out, for example, that different results would be obtained if $w_1 = 0$ and $w_2 = 1$ were used, even though this weighting scheme also leads to relatively simple score interpretation. He sees this difference in results as problematic because "the ultimate aim of equating is to provide an equation that describes the nature of the conversion of units from one instrument to another, without regard to the nature of the particular groups to whom these instruments were administered in the equating study" (Angoff, 1987, p. 295). To buttress his argument, he points out that a study by Angoff and Cowell (1986) sup-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 301–306

0146-6216/87/030301-06$1.55

ports the assumption of group invariance with respect to equating, and he claims that using $w_1 = 1$ and $w_2 = 0$ will likely lead to biased equating results.

As stated above, our recommendation for consideration of $w_1 = 1$ and $w_2 = 0$ is based upon the fact that these weights lead to simpler equations, as well as our judgment that results obtained using these weights are simpler to interpret, and are often (not always) more directly relevant to decisions made in certain contexts, especially licensure and certification testing. We offer no further defense for these particular weights. However, we infer from some of Angoff's comments that his criticisms might extend to any weighting system that departs from weighting relative to equating sample sizes or, perhaps, equal weighting. Therefore, many of the comments that follow are framed in the context of any set of weights.

We are in complete agreement with Angoff's statement about the "ultimate aim of equating," and his excellent study with Cowell lends considerable credence to the assumption of group invariance. Our disagreement with Angoff centers around our interpretation of the role of w_1 and w_2 in the equations in Kolen and Brennan (1987).

Specifically, we argue that w_1 and w_2 *define* the population of interest and, as such, they may be viewed as (synthetic) population *parameters*. Once this population is defined (i.e., w_1 and w_2 are specified), then (and only then) can the invariance of any estimated equating relationship be examined for various subsets of this (synthetic) population. Indeed, as Angoff has pointed out, the very form of the Tucker and Levine equations (see Table 3 of Brennan & Kolen, 1987) dictates that they are *not* invariant with respect to the choice of the w_1 and w_2 weights used to define the synthetic population. In our view, this fact is not an indictment of any particular weighting scheme; it is merely a statement of fact about equating with nonequivalent populations. Consequently, the issue of invariance is primarily an issue for subsets of examinees from the synthetic population, not for the potentially infinite number of ways the synthetic population could be defined.

Of course, it is possible to examine the extent to which choice of weights leads to different equating relationships with the common-item nonequivalent-populations design. The conclusion reached will be a consequence of the extent to which the populations that form the synthetic population are nonequivalent. The nonequivalence of the populations in conjunction with different weights guarantees different equating relationships, even though the differences are often very slight.

Note also that, given the role of w_1 and w_2 in the equations in Kolen and Brennan (1987), w_1 and w_2 should be defined a priori—that is, before equating samples are selected or sample sizes are specified. This point is analogous to the a priori specification of a population before drawing a sample in the classical theory of statistical inference. From this perspective it is somewhat difficult to justify the weights $w_1 = n_1/(n_1 + n_2)$ and $w_2 = n_2/(n_1 + n_2)$, which are based on equating sample sizes, *unless* the relative values of n_1 and n_2 are chosen to reflect the intended, relative weights of the two populations in the synthetic population.

In practice, $w_1 = n_1/(n_1 + n_2)$ and $w_2 = n_2/(n_1 + n_2)$ are likely the most commonly used weights, in part because their use leads to relatively simple equations, such as those in Angoff (1971). Moreover, these weights often yield a synthetic population that practitioners would claim is of some interest to them if they were specifically asked to define the synthetic population. However, in our experience, equating sample sizes do not always reflect the relative, intended weight of the separate populations in forming a synthetic population. The problem becomes especially complicated when a new form is linked to two or more old forms and different sample sizes are used for the various links. The extent to which practitioners are truly interested in the specific synthetic population they have implicitly defined for a specific equating is a matter of speculation.

Finally, we strongly endorse Angoff's concern for characterizing and eliminating bias in equating. However, it is our contention that the issue of bias can be addressed only after the population has been clearly and unambiguously defined (i.e., only after

a *single* choice has been made for w_1 and w_2). That is, in our view, an estimated equating relationship may be biased relative to the "true" relationship for the prespecified population of interest, but bias does not occur simply because of the method chosen to define a population a priori. Similarly, we would not find it helpful (and perhaps not meaningful) to say that an observed equating relationship is biased using $w_1 = 1$ and $w_2 = 0$ but unbiased using some other choice of weights. It is true, of course, that bias will result if $w_1 = 1$ and $w_2 = 0$ are actually used when the intended weights are, say, $w_1 = w_2 = .5$, but this would be a clear misuse of the weights $w_1 = 1$ and $w_2 = 0$.

The above comments probably should be evaluated primarily for their theoretical/philosophical worth, rather than their practical implications in most contexts. As stated and illustrated in Kolen and Brennan (1987), different choices of weights usually lead to only slight differences in equating relationships.

Circular Equating Paradigm

Angoff (1987) states that the circular equating paradigm, as a process for checking on equating stability, is useful because "it provides advance knowledge of what the errorless result should be, and this knowledge can be used in evaluating the error in an actual equating chain" (p. 298). Consequently, he suggests that the concerns about this paradigm that are expressed by Brennan and Kolen (1987) "seem not to be fully justified" (p. 298).

We agree with Angoff that the errorless result for equating in a circle is clear. Specifically, for linear equating with three forms, the errorless result is a slope of 1 and an intercept of 0 for Equations 5, 6, and 7 in Brennan and Kolen (1987). However, in practice, Brennan and Kolen state that the form of Equations 5 through 7 reveals that, although the slopes will be equal (not necessarily 1), the intercepts will be *unequal*. Angoff disagrees. He states that "if the values of the slope and intercept parameters in their Equations 5, 6, and 7 were specified in detail . . . the intercepts in all three equations would be zero" (p. 298).

Although Angoff does not present a proof of his contention, we believe that the essence of our disagreement with him involves population considerations similar to those discussed above. To address this issue, consider the circular equating paradigm in Figure 1 of Brennan and Kolen (1987), and let the three constituent equatings be denoted X_1-to-Y_2, Y_2-to-Z_3, and Z_3-to-X_1, where the subscripts 1, 2, and 3 designate the groups of examinees who take Forms X, Y, and Z, respectively. If these three groups are isomorphic with the *same* population, then Angoff's statement is true. In practice, Angoff's zero-intercepts contention could be achieved, for example, by administering all three tests to the *same* group of examinees and equating the three pairs of tests *without* using common-item links.

However, Angoff's contention about zero intercepts is not true, in general, for Tucker and/or Levine equating with the common-item nonequivalent-populations design. When this design is used in the circular equating paradigm, three *different* synthetic populations are involved. They are the synthetic populations for X_1-to-Y_2, Y_2-to-Z_3, and Z_3-to-X_1. It is possible to demonstrate analytically that Equations 5 through 7 generally have nonzero and unequal intercepts for any a priori choice of weights, but the required equations are exceedingly complex. Therefore, provided below are illustrative results for circular equating with three forms of two certification tests, each of which has 125 items with a common-item section of 30 items. (Kolen, 1985, also considered these data, but from a slightly different perspective.)

For each test, Table 1 provides slopes and intercepts for the linear equatings of each of the three pairs of forms based on Tucker assumptions with weights proportional to sample sizes. (For purposes of subsequent comparisons, mean equating results are also provided.) Using the linear equating results in Table 1, Table 2 provides the slopes and intercepts for equating in a circle, conditional on the form used as the initial (and final) form. For example, using Equation 5 in Brennan and Kolen (1987) and starting with Form X, the result is

$$x' = a_x a_y a_z(x) + (a_z a_y b_x + a_z b_y + b_z)$$

$$= (.98717)(1.15537)(.95144)(x) + [(.95144)(1.15537)(1.80766) + (.95144)(-17.56786) + 7.52014]$$
$$= 1.0852(x) - 7.2075 \quad . \qquad (1)$$

For both tests, the linear (Tucker) equating results in Table 2 demonstrate that, under the circular equating paradigm, slopes are identical regardless of initial form but intercepts *differ*. (Table 2 also demonstrates that with mean equating, both the slopes and intercepts are unchanged by the choice of initial form.) Furthermore, the equatings are unequally accurate for the three starting forms, based on a root-mean-squared-error criterion defined as

$$\text{RMSE} = \left[\sum_{i=0}^{k} f_i(\hat{g}_i - i)^2 \Big/ \sum_{i=0}^{k} f_i \right]^{1/2} \quad , \qquad (2)$$

where f_i is the frequency of score i ($i = 0, \ldots, k$) of the initial form, and $\hat{g}_i$ is the estimated equated score associated with a raw score of i.

Clearly, under the circular equating paradigm with Tucker (and Levine) equating, results differ depending on the initial form that is specified. These differences are a source of ambiguity to us when we attempt to interpret results such as those in Table 2.

Recall that the linear equating results in Tables 1 and 2 are based on weights that are proportional to equating sample sizes (see footnote b in Table 1). It might be hypothesized that the intercepts and RMSEs would be identical under some other a priori choice of weights. This is not generally true. For example, with weights of .5 for the Test 1 forms in Tables 1 and 2, the Tucker linear equating intercepts are -7.1882, -7.2496, and -6.8834 when the starting forms are X, Y, and Z, respectively, and the corresponding RMSEs are 1.5421, 1.4441, and 1.6929. This illustrates that the crux of the matter is not the specific choice of weights; it is the fact that, implicitly or explicitly, weights are required because populations are nonequivalent.

We also express concern about the circular equating paradigm as a procedure for evaluating different equating procedures. Specifically, we state that under the circular equating paradigm,

> . . . an equating procedure involving the estimation of only one or two moments often

Table 1
Mean and Linear Equating Results for Three Forms in a Circular Equating Paradigm Using Tucker Assumptions

Test	Type of Equating	Statistic	Equating Results: X_1-to-Y_2	Y_2-to-Z_3	Z_3-to-X_1
1	Mean[a]	Slope	1.00000	1.00000	1.00000
		Intercept	.56873	-2.44229	2.91695
	Linear[b]	Slope	.98717	1.15537	.95144
		Intercept	1.80766	-17.56786	7.52014
2	Mean[a]	Slope	1.00000	1.00000	1.00000
		Intercept	.76581	-1.24437	1.08364
	Linear[b]	Slope	.94326	1.05589	.98757
		Intercept	6.04892	-6.49943	2.23874

[a]For mean equating, an equated score is $m(x) = x - [\mu_s(X) - \mu_s(Y)]$ where $\mu_s(X)$ and $\mu_s(Y)$, under Tucker assumptions, are defined in Table 3 of Kolen and Brennan (1987).

[b]The weights are as follows:
X_1-to-Y_2: $w_1 = .3072$, $w_2 = .6928$
Y_2-to-Z_3: $w_2 = .6988$, $w_3 = .3012$
Z_3-to-X_1: $w_3 = .4930$, $w_1 = .5070$

Table 2
Results for Circular Equating Paradigm Conditional on the Form Used as the Initial Form

Test	Type of Equating	Statistic	Initial Form X	Initial Form Y	Initial Form Z
1	Mean	Slope	1.0000	1.0000	1.0000
		Intercept	1.0434	1.0434	1.0434
		RMSE	1.0434	1.0434	1.0434
	Linear	Slope	1.0852	1.0852	1.0852
		Intercept	-7.2075	-7.2690	-6.9023
		RMSE	1.5420	1.4438	1.6926
	Equipercentile	RMSE	1.7171	1.6061	1.5514
2	Mean	Slope	1.0000	1.0000	1.0000
		Intercept	.6051	.6051	.6051
		RMSE	.6051	.6051	.6051
	Linear	Slope	.9836	.9836	.9836
		Intercept	2.1277	2.1062	2.1173
		RMSE	.6221	.5939	.6126
	Equipercentile	RMSE	.8663	.7628	.8574

Note. Mean and linear equating results are based on Tucker assumptions. Equipercentile equating results are based on frequency estimation.

> tends to appear better than one involving the estimation of many moments. For example, using this paradigm with real data, mean equating often appears to be better than linear equating, which in turn often appears to be better than equipercentile equating. (Brennan & Kolen, 1987, p. 283)

These statements are illustrated by the mean, linear, and equipercentile equating results in Table 2. Based on the RMSE statistics, in almost all cases mean equating appears better than linear equating, and linear equating appears better than equipercentile equating. There are exceptions (e.g., equipercentile vs. linear equating for Test 1 starting with Form Z), but not many. We suggest that this pattern of results is primarily a statistical artifact brought about by comparing procedures that involve estimating different numbers of parameters.

In short, we continue to have reservations about the circular equating paradigm as a procedure to examine equating stability or to evaluate equating procedures based on estimating different numbers of parameters. Consequently, we recommend that this paradigm be used with considerable caution.

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. (Reprinted by Educational Testing Service, Princeton NJ, 1984.)

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

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23,* 327–345.

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

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Kolen, M. J. (1985, April). *Comparison of methods for linear equating under the common item nonequivalent populations design*. Paper presented at the annual

meeting of the American Educational Research Association, Chicago.

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

Author's Address

Send requests for reprints or further information to Robert L. Brennan, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.