PERSONNEL PSYCHOLOGY
2001, 54

THE CONTROVERSY OVER SCORE BANDING IN PERSONNEL SELECTION: ANSWERS TO 10 KEY QUESTIONS

MICHAEL A. CAMPION
Purdue University

JAMES L. OUTTZ
Outtz & Associates

SHELDON ZEDECK
University of California at Berkeley

FRANK L. SCHMIDT
University of Iowa

JERARD F. KEHOE
AT&T

KEVIN R. MURPHY
Pennsylvania State University

ROBERT M. GUION
Bowling Green State University (Emeritus)

A particular form of test score banding, in which bands are based on the reliability of the test and in which selection within bands takes into account criteria that are likely to enhance workforce diversity, has been proposed as an alternative to the traditional top-down (rank-order) hiring systems, but it has been hotly debated among both scientists and practitioners. In a question-and-answer format, this article presents three different viewpoints (proponents, critics, and neutral observers) on the scientific, legal, and practical issues. The article also attempts to seek some consensus among experts on this controversial procedure.

Attaining the dual goals of valid selection and a diverse workforce is often extremely difficult because the most valid selection procedures (such as mental ability tests) tend to have adverse impact. This is especially the case with traditional top-down (rank-order) hiring systems.

This has left selection practitioners with a dilemma. If they use traditional valid selection procedures (in particular, cognitive ability tests), and they observe the likely subgroup differences, they will have adverse impact (which may be against the law if it cannot be adequately de-

---

This article is based on a panel discussion by the same title presented at the meeting of the Society for Industrial and Organizational Psychology in Dallas, Texas (April, 1998).

Correspondence and requests for reprints should be addressed to Michael A. Campion, Krannert Graduate School of Management, Purdue University, 1310 Krannert Building, West Lafayette, IN 47907-1310; campionm@mgmt purdue.edu.

fended). The alternative may be to use less valid selection, which results in lower performing employees and less successful organizations. This is a choice between a "rock and a hard place" and may be the most perplexing problem facing the practice of personnel selection today. To make matters worse, this dilemma is most likely to occur if the selection procedures tap into mental ability either directly or indirectly, which many do, or if there is a low selection ratio, which is often the case.

In the past, it was possible to use "within-group norming" wherein the scores of candidates would be adjusted (often to a percentile) to reflect their relative standing compared to other candidates of their own race or sex. This allowed the employer to pick the best of each group and, at the same time, hire candidates in proportion to the population or other diversity goals. The Civil Rights Act of 1991 forbids within-group norming because some people viewed it as reverse discrimination. With protected groups often having lower means on selection procedures, a given relative (percentile) score will reflect a lower absolute score. Thus, it would be possible for a lower-scoring member of a protected group to be hired before a higher-scoring member of a nonprotected group.

Score "banding" has been proposed as an alternative to top-down selection and as one potential way of dealing with this dilemma (Cascio, Outtz, Zedeck, & Goldstein, 1991). As opposed to the traditional top-down method of selecting candidates, banding involves grouping the scores within given ranges and treating them as equivalent. It is based on the notion that small score differences may not be meaningful because they fall within the range of values that might reasonably arise as a result of simple measurement error. Banding reduces adverse impact because the bands are wide enough to include lower-scoring group members, and then hiring within the bands can be based on affirmative action or factors that do not show group differences (e.g., seniority, experience, or random draw).

Banding has been hotly debated within the field, however. There have been many follow-up articles on the topic in *Human Performance* since the Cascio et al. (1991) article (Aguinis, Cortina, & Goldberg, 1998; Cascio, Goldstein, Outtz, & Zedeck, 1995; Murphy & Myors, 1995; Sackett & Roth, 1991; Schmidt, 1991; Schmidt & Hunter, 1995; Siskin, 1995; Zedeck, Outtz, Cascio, & Goldstein, 1991). There have also been articles on the topic in the *American Psychologist* (Cascio, Zedeck, Goldstein, & Outtz, 1995; Gottfredson, 1994; Sackett & Wilk, 1994), *Personnel Psychology* (Murphy, 1994; Murphy, Osten, & Myors, 1995), the *Journal of Applied Psychology* (Truxillo & Bauer, 1999), and *Psychological Assessment* (Kehoe & Tenopyr, 1994), as well as a report on banding by the scientific affairs committee of the Society for Industrial and Organizational Psychology (Scientific Affairs Committee, 1994). The issues cover

the entire spectrum from statistical and scientific, to practical and legal, to moral and ethical.

The purpose of this article is to take stock of the current debate on the topic. The focus will not be on the general concept of banding because our field does that in many ways (e.g., cutoff scores and grades are forms of banding), but instead the focus will be on the special Cascio et al. (1991) form of banding. It has been several years since most of the research on the topic was conducted (especially when considering publication lags). Furthermore, substantial practical experience has accumulated as banding has been used in many situations in actual organizations.

Three types of experts have been amassed for this purpose. Each represents a key perspective on the topic. First, we have two proponents of this particular banding strategy, Sheldon Zedeck and James Outtz. They were two of the authors of the original article on the topic and all the rebuttals, and they have both been extremely active in using the technique in a variety of practical applications. Second, we have two critics of the statistical rationale for this particular approach to banding, Frank Schmidt, who has authored most of the articles criticizing the topic, and Jerry Kehoe, who will assume the role of critic by raising concerns with this strategy for defining bandwidths. Finally, we have two neutral observers, Kevin Murphy and Robert Guion, neither of whom has been identified as a strong proponent or critic of banding. Murphy has written several articles objectively examining the empirical consequences of banding, and Guion is a well-known commentator on personnel selection and author of the major handbook chapters and books on selection. All the authors have extensive expertise in the design and evaluation of selection programs, and thus they will not limit their comments to these roles.

The remainder of the article is organized around 10 questions on four major topics: (a) a summary of the scientific issues (because these issues have been addressed previously), (b) the legal issues and evidence, (c) recommendations whether banding should be used, including points of agreement and disagreement among the experts, and (d) practical advice for how to use banding. The practical advice follows the recommendations because some readers may decide not to use the procedure after weighing the pros and cons. For each question, comments will be presented (in order) by proponents, critics, and neutral observers.

Scientific Issues Surrounding Banding

1. In brief, what are the major psychometric and other scientific pros and cons of using banding? There are several subquestions here such as,

what is the psychometric rationale for treating scores within a band the same, and what are the psychometric issues associated with evaluating the choice of bandwidth?

Outtz and Zedeck: Banding is a tool that can be useful in certain situations. Scores on selection devices contain error. Strict rank-order selection (i.e., treating any difference in scores as meaningful) ignores measurement error. Measurement error manifests itself in different ways. As an example, validity coefficients typically range from .20 to .50. Therefore, selection devices typically account for 4% to 25% of the variance in the criterion. This means that 75% to 96% of the variance in the criterion is not accounted for. Yet, strict rank-order selection utilizes predictor scores as if they account for the total variance in job performance.

Most, if not all, referral methods in use today, even strict rank ordering, involve banding to some degree. Strict rank-order selection simply involves very narrow bands (e.g., one point). The differentiating factors are primarily the width of the bands and how scores are interpreted in light of the bands.

In strict rank-order selection, the band starts with the highest score and moves downward based upon any difference in scores until all vacancies are filled. This referral method is based on the inference that the average predicted criterion performance of the group selected will be higher than the average predicted criterion performance of any group selected in any other way. The accuracy of this inference depends upon the magnitude of the validity coefficient. Because validity coefficients are not perfect, we know that this inference will not be accurate for every sample selected. However, we theorize that "in the long run the mean criterion performance of samples selected via strict rank-order will be higher than those selected by any other method."

Other procedures (e.g., fixed bands or sliding bands) start from the highest score and move downward to include all scores within a score range determined on the basis of factors such as the reliability of the test, standard error of measurement or the standard error of the difference, or the consequences of prediction error. One rationale for these banding procedures is that the initial selection device may cover a limited portion of the criterion space with less than perfect prediction. Therefore, additional screening on a wider spectrum of KSAs is desirable before making final selections. It may be inappropriate to base access to additional screening on very small differences in scores on the initial selection device. Thus small differences are ignored.

The primary differences between strict rank-order selection and banding are:

1. Strict rank-order selection, in its purest form, results in the referral of fewer people per vacancy than selection methods that aggregate test scores.
2. Other banding procedures result in more applicants than vacancies, thereby necessitating selections from within a band.

It is important to note that all of the participants to this discussion now recognize that banding is a viable option, a position not always presented (e.g., Schmidt, 1991). The disagreement appears to be on what method or strategy to use for banding. It is our position that banding can be based on the notion of "reliability of measurement." If our measuring devices are not perfectly reliable, we should take into account the degree of unreliability in our interpretation of scores. In this way, we are relying on data that are part of the strategy for test development and validation. We view this as a better defense of banding than purely arbitrary decisions (e.g., that 90–100% is an A, 80–89% is a B, etc.) made exclusively by "experts" based on their assessment of the situation (i.e., judgments made by management "experts"). More information on our position can be found in the articles on banding presented in the reference list.

Schmidt: In answering this question, we must first make clear what we mean by "banding." There are two kinds of banding. First, there is traditional banding; a good example of this is the old-fashioned expectancy charts used in personnel selection since the 1930s. These are presented as bar graphs showing the probability of above average job performance for different bands of test scores. For example, it may happen that applicants with scores between 50 and 55, 80% are later rated as above average in performance. Taken together, the bands cover the entire test score range, and the result is a sort of histogram that has been found to be useful in conveying the meaning of validity to employers. Another example is the practice of the Gallup organization of dividing job interviewees, based on their interview scores, into A, B, C, D, and F groups, with the recommendation to the employer to hire only from the A group if possible. Traditional banding is used frequently in personnel selection. The only potential problem with such traditional banding is loss of utility resulting from treating all scores within each band as equal. But because the bands have typically been narrow, even this problem has been minimal.

The other type of banding, advocated by Cascio et al. (1991), is based on statistical significance testing. That is, it is based on the assumption that different scores should be viewed as equivalent unless they are statistically significantly different. Determination of whether any two scores are significantly different is based on the standard error of the difference (SED) between scores; hence, the name SED banding.

Unlike traditional banding, SED banding suffers from an internal contradiction: in order to actually use SED banding, one must violate the foundational assumption that scores that are not statistically significantly different must be treated as equivalent (Schmidt, 1991; Schmidt & Hunter, 1995). This is the key problem that has made SED banding controversial.

Obviously, one way to avoid this controversy and still use banding is to use traditional banding, which is not based on significance testing. What, then, is the appeal for some of SED banding? It is the illusion that significance testing provides a *scientific* basis for determining the width of bands. In traditional banding, bands are set based on professional judgment and administrative convenience. The advocates of SED banding maintain that significance testing provides an objective, scientific basis for banding (Cascio et al., 1991), and is therefore superior to traditional banding. However, this claim has been shown to rest on a logical contradiction and therefore to be false (Schmidt, 1991; Schmidt & Hunter, 1995).

There is no controversy surrounding traditional banding. The controversy surrounds SED banding, and that is why SED banding is the subject of this article.

I would now like to examine some troubling properties of SED banding. The introduction to this article states that SED banding "is based on the notion that small score differences are not meaningful." Similar statements are made by Cascio et al. (1991). But these score differences are not really "small" in SED banding.

Suppose our test has a reliability of .80 and we use a bandwidth of 2 SEDs, as recommended by Cascio et al. (1991). If the test distribution is approximately normal and if we take the highest score as being the one at the 99.9 percentile, the resulting 95% SED band will contain 38% of the score range and about 25% of the job applicants. That is, we are saying that scores are equal in the top 38% of the score range, and we are now treating the top 25% of all observed scores (or applicants) as having equal true scores for purposes of selection. This is a very wide band, yet it is typical of SED banding.

Is it true that the top 25% of all test scores have equal true scores? Psychometric methods tell us that the square root of the reliability coefficient is the correlation between observed scores and true scores. The square root of .80 is .89. So the observed scores on this test are linearly correlated almost .90 with true scores. This tells us that the conclusion that the top 25% of scores have equal true scores is false, because if that were true, the linear correlation between observed and true scores would certainly not be as high as .89! It would be much lower.

Now let us look at the difference in expected job performance between the top and bottom scores in our band. If the criterion-related validity of our test for predicting job performance is .50, that difference is .64 of a standard deviation in job performance. *This is a large difference.* A difference this large has substantial practical utility implications. For example, suppose that on a particular job *SDy* is $20,000–a fairly typical value. Then 64% of $20,000 is $12,800 per year in reduced output. In addition, this difference is often considerably larger than .64 *SDy*s, as shown in the results presented by Siskin (1995). So not only are true scores not equal for the observed scores in the band, expected job performance is also very unequal.

Here is another way to look at banding. We know from thousands of studies that the relations between ability, aptitude, and knowledge tests and job performance are *linear*—a straight line relationship (e.g., Coward & Sackett, 1990). This finding is so well established that it is stated in the SIOP *Principles*—which is unusual for any research finding.

Now imagine such a linear regression line of job performance on test scores. Imagine that you start at the top of this regression line and go about 40% of the way down the line, and you bend that top 40% of the line downward until it is horizontal. This represents what SED banding does. Banding considers the top 38% of the score range as equal for selection purposes. This means it acts as if the regression line were horizontal in almost 40% of the score range. In that range, the predicted job performance is the same for all scores—a prediction that is false. As we saw, the difference in expected job performance within the band between the top and bottom scorers is .64 *SDy*.

It should be noted here that non-SED bands do not assume horizontal regression lines within a band. This is because non-SED bands are not based on the statistical significance testing rationale that entails the statistical assumption of horizontal regression lines within bands. Instead, the rationale for non-SED bands is administrative convenience. Unlike SED based bands, these bands do not entail assumptions that deny the within band regression of performance on test scores. Instead, they accept the within band relationship between scores and performance and then make the decision to ignore this relationship in the interests of administrative convenience. In addition, the fact that non-SED bands are typically much narrower than SED based bands means the job performance losses are much smaller.

Despite these facts, banding advocates state that the top and bottom scores in the band are "psychometrically indistinguishable." What does this phrase *mean*? It means that if you use a statistical significance test with low power, you will not be able to detect as statistically significant the true score differences that you already know are there!

Banding is a way of testing differences between scores for statistical significance. The basic principle in banding is that if two scores are not statistically significantly different, then they are the same and should be treated as equal. Statistical significance is typically defined as a difference larger than 2 SEDs (as it was in our example). This test has very low statistical power to detect the differences we *know* are there. The power is low because the sample size is $N = 2$! (The sample size is $N = 2$ because we are comparing only two scores at a time.)

We know the differences are there from the linear model and from the high correlation between observed scores and true scores. Because we know the differences are there, we know that the failure to detect them is due to low power.

This illustrates a fundamental intellectual problem in SED banding. SED banding, in effect, says the following: We know from the linear model that there are real differences in observed and true scores in the top 38% of the score range, as there is in the rest of the range. However, you are not allowed to use these differences in selection, because a very low power significance test cannot detect them.

There is a more general principle that underlies the above point: Statistical significance testing is alien to the basic linear model underlying prediction in selection and in any other area. When regression is used in the prediction of any variable, it is *never* the case that scores on the independent variable can be or should be tested to see which are statistically significantly different from each other. The scores are taken as given, and predictions are made using the scores as given. Regression-based prediction models do not include significance testing on individual scores. That is inappropriate under the regression model, which is the prediction model in selection. In fact, statistical significance testing is an undesirable data analytic procedure under any circumstances (Cohen, 1994; Hunter, 1997; Loftus, 1996; Schmidt, 1996; Schmidt & Hunter, 1997), but space does not permit us to pursue that point here.

Finally, I want to reiterate a point from my 1991 article: There is a fatal *logical* contradiction in SED banding. The fundamental principle in SED banding is that scores that are not statistically significantly different cannot be considered different for selection purposes. They must be treated the same and considered interchangeable. This is the fundamental principle of SED banding, and SED banding violates this very principle in every operational application of banding. In every application of SED banding, there are large numbers of scores below the bottom of the band that are not statistically significantly different from most of the scores *that are in the band*. These scores are being treated differently even though the basic principle underlying banding says they are equivalent to those in the band ("psychometrically indistinguishable").

If this basic principle of banding is adhered to, *all scores* must be placed in the band, as I showed in my 1991 paper. That is, the band includes all the observed test scores, and the process of selection becomes entirely random. So there is no point in having a test or any other selection procedure.

The only way SED banding can avoid this is by violating its own foundational principle. Advocates of SED banding state that this principle will be applied only to differences between the top score and all other scores. This creates the logical contradiction: They say that scores should not be treated differently unless they are significantly different, but they violate their own principle in implementing banding for operational use.

This is the fundamental conceptual problem of SED banding based on the criterion of statistically significant differences between test scores. The obvious implication of this problem is that SED banding should not be used.

Kehoe: Although this debate will improve our understanding of banding and its various applications, the focus of this debate is not banding in general. Rather, the debate is about the specific rationale that score differences within a band are completely unreliable. As all authors note, banding is a very common practice, and in many cases (probably the vast majority of cases) does not rely on the premise of unreliability within bands. This premise is unique to the banding strategy proposed by Cascio et al. (1991).

The specific psychometric issue at the heart of the banding debate is whether observed scores within a band defined following the Cascio et al. (1991) procedure, SED banding, are reliably different or not. Cascio et al. argue that such score differences are completely unreliable. Kehoe and Tenopyr (1994) and Schmidt (1991) argue that they are reliable, even if they are small. To many this may seem to be a trivial point given that all agree small differences, *even if reliable*, can be unimportant. But this is an important point to organizations even if the statistical and psychometric issues appear to some to be trivial. The importance derives from the potential negative consequences if organizations can be compelled by case law to treat valid differences as completely unreliable.

The Cascio et al. (1991) SED banding practice relies on classical significance test theory applied to observed score differences resulting from some measurement process. Their rationale is that scores less than 1.96 standard errors of score differences (*SD*diff) apart are not significantly different at the $p < .05$ level. (Just as two means that are less than 1.96 standard errors of mean differences apart are not significantly different in a classical t-test inference process.) From a psychometric

and statistical point of view there is one fundamental flaw and two other problems with this rationale.

The fundamental psychometric and statistical flaw is the belief that reliable, valid scores can produce completely unreliable score differences. The Cascio et al. (1991) rationale is that scores within 1.96 *SD*diff of each other are not statistically significantly different. That is, the differences between observed scores within bands are completely unreliable. But the differences between observed scores can be completely unreliable only if the observed scores themselves are completely unreliable. But, of course, the observed scores are known to be reliable. The fact that the scores themselves are already determined to be reliable and valid means that the differences between randomly selected pairs of them must, to some extent, also be reliable and valid. Another way of expressing this point is that once the statistical conclusion is reached that a measure is reliable, it is no longer meaningful from a hypothesis testing perspective to then test a subsequent null hypothesis that differences among those reliable scores are unreliable.

This can be shown in a formulaic fashion based on the classical psychometric theory that is the premise for this debate. Classical psychometric theory holds that the variance of observed scores on a measure, X, can be expressed as the sum of two components, true score variance and error variance. That is,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \quad [1]$$

Based on the definition of the variance of a composite:

$$\sigma^2(X_1 - X_2) = \sigma^2 X_1 + \sigma^2 X_2 - COV X_1, X_2 \quad [2]$$

Combining [1] with the observation that the covariance between randomly selected pairs of scores is zero yields

$$\sigma^2(X_1 - X_2) = 2 * (\sigma_T^2 + \sigma_E^2) \quad [3]$$

Cascio et al. (1991) argue that for scores within 1.96 *SD*diff, $\sigma_T^2 = 0$. But the prior conclusion that measurement process X is reliable has already established that $\sigma_T^2 > 0$. Both conclusions can't be true.

Assuming for the sake of argument that the 1.96 *SD*diff bandwidth constitutes a statistically valid inference procedure, a significant problem with the SED banding approach is that it has very low power to detect small-to-moderate true differences. As Kehoe and Tenopyr (1994) showed, the 1.96 *SD*diff bandwidth is frequently equal to or larger than the standard deviation of observed scores on the measurement proce-

dure (*SD*). This means, for example, that the Cascio et al. (1991) procedure will frequently conclude that true score differences as large as one *SD* do not exist.

A related problem is that score differences within 1.96 *SD*diff of each other can be important and large both statistically and practically. As Kehoe and Tenopyr (1994) showed given typical levels of reliability and validity, the likelihood of scorers at the top of an SED band scoring higher on a retest than scorers at the bottom of the same band can be 25 times more likely than bottom scorers outscoring top scorers. Similarly, scorers at the top of a SED band can be twice as likely to outperform bottom scorers on subsequent job performance than vice versa. Finally, top scorers can easily be twice as likely to perform above some standard of job success as bottom scorers.

In summary, there are two key technical issues in this debate. First, the conclusion that scores within a 1.96 *SD*diff band are not reliably different is inconsistent with the conclusion that individual test scores are reliable. Second, score differences less than or equal to 1.96 *SD*diff can be large and important. All forms of banding ignore some amount of score difference. Cascio et al. (1991) argue that differences within an SED band can be ignored (indeed, should be ignored) because they are unreliable. Other forms of banding argue that reliable differences within a band can be ignored because they are less important than other organizational objectives.

Murphy: If there is a large number of applicants, it is common to find that some applicants have identical scores on a test. For example, the Wonderlic Personnel Test reports scores ranging from 0 to 50; virtually all applicants will fall in a range of 30 or fewer points. With hundreds of applicants, you could be certain of encountering tied scores, and it might be necessary to make decisions to select some applicants and reject others even though their scores were identical. Dealing with ties requires you to decide what criteria other than test scores might discriminate among people with identical test scores. Decisions about whether or not to use some form of banding essentially boil down to the question of whether similar considerations should be taken into account when scores are highly similar rather than absolutely identical (Murphy & Myors, 1995).

It is common practice in many areas of measurement to aggregate scores (i.e., to group together people whose scores are similar but not identical into categories). For example, college grades are reported in terms of aggregated scales (e.g., A, B, etc.), not in terms of numerical averages on the various tests and papers graded during the term (e.g., 91, 92, etc.). Percentile ranks, stanines, and a variety of other categorization systems are used in personnel selection and classification, and

these often involve aggregating scores into bands of one sort or another. Grouping scores into broader categories always involves the loss of some information, but the users of test scores often find grouped or aggregate scores easier to interpret and work with. If you accept that aggregating scores in this way is both a common and reasonable thing to do, the question becomes one of how to aggregate.

One distinction between banding and other methods of aggregating scores is that decisions to treat scores as similar or different are tied to the measurement precision of the test (Murphy, Osten, & Myors, 1995). If a test is highly reliable, it might make sense to make fine-grained distinctions, but with less reliable tests, small differences in test scores might not be seen as meaningful. The advantage of banding is that the width of the band is tied directly to the reliability of the test. Virtually any alternative is likely to group scores on some arbitrary basis.

There are two arguments against banding that are often put forward. First, it is obviously suboptimal to group scores. Because most relationships between test scores and criteria of interest are likely to be linear (or at least monotonic), any scoring system that ignores even the most trivial difference between test scores entails some loss of information. Judged strictly from the criterion of statistical optimality, it is always better to treat the most trivial difference in test scores as if they were meaningful. However, it is hard to argue for the optimality criterion in the face of the numerous practices in personnel selection that plainly violate this criterion. The most obvious violation has to do with the way that we score tests. It is well known that number-right scores are rarely optimal for extracting information from tests, but these are the norm in virtually all applications of selection testing (some computerized tests now use scoring systems that are more nearly optimal). Similarly, the widespread practice of relying most heavily on selection methods that show poor track records for reliability and validity (e.g., poorly structured interviews) suggest that statistical optimality is not a criterion by which important decisions are made in personnel selection.

A more important criticism of banding is that it is an indirect and perhaps even an ingenuous method of inserting criteria that are not normally part of the selection process into selection decisions. Most applications of banding that have been discussed in the literature exist primarily as a means of providing better opportunities for applicants from some protected group. That is, by broadening the definition of a "tie" from its literal sense (i.e., comparisons among candidates who are truly identical on all normal criteria) to its broader sense (i.e., comparisons among candidates whose scores are judged to be similar), it becomes easier to introduce specific criteria (e.g., group membership) into selection decisions without appearing to violate anyone's rights or interests.

It is both more honest and more efficient to make an explicit decision about the value placed on specific selection criteria (e.g., you might articulate a good rationale for valuing diversity in an organization), and to include them formally as predictors in a multiple-component selection system. If the criteria that are used to differentiate among individuals within a band are legitimate ones, it should be possible to use them more generally as part of an overall decision-making process. The most damaging critique of banding is probably that it often looks like an effort to introduce selection criteria that seem to favor some groups over others through the back door.

Any attempt to make explicit statements about the value an organization places on diversity, or on other criteria that are used to discriminate among members of a band is likely to be a source of controversy, and possibly, as a source of legal trouble. Traditionally, organizations and I-O psychologists have been content to make broad but meaningless statements (e.g., this organization value's diversity) without coming to grips with the difficult question of how much value is attached to criteria of this sort, or why. However, failure to make hard decisions about what criteria are valued, and about how much value is attached to different criteria does not solve the problem, it merely pushes into the background. For example, an organization that chooses to use a cognitive ability test that costs $5 per person rather than an equally valid but less discriminatory structured interview that costs $100 per person has effectively made a value decision—that is, efforts to increase workforce diversity are worth less than $95 per applicant. Any choice an organization makes about criteria such as diversity is a concrete statement of values, and everyone would be better off (although many people may be less comfortable or less satisfied with a specific concrete statement than with meaningless boilerplate) if these statements were open and explicit.

Guion: There is another category: the management pros and cons. My position is that it is wrong to think of banding in narrow psychometric or statistical terms. What counts is the idea that some score's differences do not make a difference to management or, more broadly, to the organization. This is not a matter of statistical significance; it is a matter of indifference. If there is enough statistical sophistication in management thinking (and that is unlikely), one might ask specifically how much of a difference in the predicted criterion, using whatever regression equation seems fitting, is enough to make a choice between candidates an interesting choice. The con, I suppose, is that this view of banding (with or without banding on the criterion itself) requires managers to think about the question of how big a difference makes a difference and how they arrive at an answer. That is also the pro.

There is also the psychometric issue, but I consider it basically irrelevant. Most discussion of the psychometrics of banding has concentrated on some variant of the standard error of measurement. To make sense, at least where criterion-related validation has been done, the concentration should be on the standard error of estimate. One need not be either a rocket scientist or a psychometrician to know that candidate's scores on tests or other assessment procedures can vary by chance. One need not be either a brain surgeon or a psychometrician to know that even potentially explainable (statistically significant) differences may not matter to organizational effectiveness. This is the logical basis for grouping people as if the differences among them were trivial (a practice now called banding).

We should stop thinking of this as a scientific, statistical, or psychometric matter, although scientific, statistical, or psychometric data, clearly developed and clearly understood by organizational policy makers, could help inform the management decisions that are made.

2. What other research or scientific questions need to be answered about banding?

Outtz and Zedeck: There are three areas that we would like to see explored. Strict top-down selection is based on the premise that if the predictor—criterion relationship is linear, the mean criterion performance of applicants selected in strict rank-order will be higher (in the long run) than the mean criterion performance of applicants selected via any other method. However, this does not necessarily indicate that the mean total job performance of the strict rank-order selections will be higher than that of a group selected via another referral method such as banding.

Little if any research has been devoted to comparisons of samples selected via strict top-down referral and banding in terms of total job performance where a given predictor does not represent the total criterion space (which is almost always the case). Research of this nature would point out the importance of addressing the multidimensionality of job performance as well as the contextual factors that influence it. As an example, less than optimal performance on one job dimension may be compensated for by exceptional performance on one or more other dimensions. The question of interest would be, "To what extent is the utility of a given predictor limited by performance on important job dimensions unrelated to that predictor?"

A second area that warrants further research is determining the most appropriate factors to be considered in deciding the bandwidth. As an example, what is the best way to go about collecting and utilizing management input, statistical data (e.g., SED), and legal perspectives in determining the most appropriate bandwidth? In other words, we are

suggesting that determination of how to set the bandwidth needs to take into account: (a) statistical data such as the reliability of the test; (b) the nature of the job (risk involved to people and property); (c) the legal and political environment as well as the business necessity pertaining to having a diverse work force; (d) long term implications for the image of the organization; and (e) other such social and political factors. The ways in which these factors are considered, integrated, and combined are beyond the present discussion, but the key factor to keep in mind is that the decision maker can control the bandwidth by choice of the critical value, C, in the SED formula.

Finally, most of our emphasis has been on the conceptual underpinnings of banding, particularly the role of reliability of measurement and the standard error of estimate in determining differences between scores. Our examples of banding have been based on content validity strategies. We would like to see more research on the criterion side, which implies the study of the impact of banding on predicted performance. Starts in this direction have been made by Aguinis et al. (1998) and Siskin (1995). Hanges, Grojean, and Smith (in press) have also commented on the focus on job performance as opposed to test scores.

Schmidt: We need to know more about how SED banding performs when the bands are formed based on predicted job performance rather than on test scores. Job performance is clearly more important than observed scores or true scores on selection instruments. I recently testified in a court case centered on SED banding in which the judge implied that for SED banding to be relevant it should focus on job performance, not on true scores. This implies a banding method that is based not on the standard error of measurement (SEM) or the SED, but on the standard error of estimate (SEE) for the regression line that predicts job performance from test scores.

Let us take our example where validity is .50. Instead of working with test scores to create the band, we are now working with predicted job performance scores, but the procedures are otherwise identical. The analogous standard error of the difference (SED) is the SEE multiplied by the square root of 2. Let us refer to $\sqrt{2}$SEE as SED*. The bandwidth for the 95% band is 2 times the SED*. And again, the band extends downward 2 SED*s from the highest score, which in this case is the highest predicted job performance level (the highest $\hat{y}$). The highest $\hat{y}$ is the $\hat{y}$ predicted from the highest test score, which in this example I took as the test score at the 99.9 percentile.

What we find is that the resulting band includes 97% of all the predicted job performance scores! Only the bottom 3% of applicants will be excluded from the band. The top 97% of applicants must all be con-

sidered equivalent for purposes of selection. If the validity of the test is less than .50, this band will be *even wider.*

What is wrong with this picture? The key variable in selection is job performance, not true scores on tests. Yet when SED banding is applied to predicted job performance, the band includes virtually all the applicants. Aguinis et al. (1998) examined this question. Because they used the 68% band rather than the 95% band, their bands were less wide. However, the 68% band does not correspond to an acceptable alpha level in significance testing. Recall that significance testing is the basis for SED banding. With a 95% band, all scores outside the band are significantly different from the top score at (at least) the .05 level. This is not true for a 68% band. For a 68% band, the corresponding alpha level is .32.

More research is needed on SED based banding for the case in which the focus is on job performance.

One of the advantages claimed for SED banding is that for the applicants within a band, it provides the opportunity to consider other relevant factors in addition to test scores. These may include job experience, seniority, past job performance (in promotion decisions), professional preparation, and so forth. However, this proposition must be evaluated in relation to the research literature on statistical versus clinical combination of information. That research literature has consistently shown that humans are poor information integrators and that statistical combination of information is both more consistent and more valid. This means that if factors such as those listed here are relevant to selection decisions, they should be included statistically in the final scores, and not weighted in subjectively when personnel decisions are made. For example, if past job experience is considered relevant, then points should be given for such experience, according to a scoring plan, and those points should be included in the total score. We need research on practices currently being used so that we can make sound recommendations based on the research findings on human decision making.

Kehoe: Other valuable research relating to the use of banding (not just SED banding) includes:

1. In various applications, what score ranges are frequently regarded as narrow enough to ignore in favor of other considerations such as administrative ease? For example, academic letter grading is a banding strategy that ignores certain ranges of numeric score information in return for the administrative ease of a small number of letter grades. Compared to the test's *SD*, how wide are these ranges? Similarly, in many employment applications, employment test scores are grouped into priority clusters again for administrative ease. In practice, what are typical ranges for these clusters?

2. Research that translates social values into bandwidths should be continued. Much of the previous banding research could be viewed as this type of research in that the goal of the research has been to show the outcome consequences of various bandwidths. Another version of this research would be to describe the social values of important stakeholders such as regulatory agencies, business leaders, advocacy leaders, and the like and translate these values into definitions of bandwidths and bandwidth applications that support those social values.
3. Research on decision rules under uncertainty should be directed at employment applications. Rather than seek statistical arguments that mask uncertainty, there is likely to be more value in describing and evaluating decision rules and processes that explicitly account for uncertainty. This research would necessarily need to address the manner in which social values can be incorporated into decision processes in the face of psychometric uncertainty. For example, Kehoe and Tenopyr (1994) described procedures for defining bandwidths based on social values relating to group selection.

Each of these three suggested areas of research is based on the premise that bandwidths (for employment selection applications) should be based on some explicit evaluation of the inevitable tradeoffs between the loss of score information and other benefits that may be gained by using bands. SED banding masks these tradeoffs by assuming that there is a statistical rationale for defining complete uncertainty implying that the use of SED bands results in no loss of score information.

Murphy: The width of a band is a function of the reliability of the test and the degree of confidence desired before deciding that two scores are meaningfully different. The major scientific question is how we should go about making decisions about the sort of confidence interval desired. Several authors have noted that this is an important decision, but to date, few principles have been articulated for deciding whether bands should be based on 95% confidence intervals, 90% intervals, 50% intervals, or some other number.

The applications of banding I have seen have relied for the most part on 90% or 95% confidence intervals. That is, the decision to label two scores as different would occur only if the researcher was 90% or 95% confident that the apparent difference in test scores could not be explained in terms of the imprecision of the test. These figures, in turn, appear to be holdovers from widely followed conventions in significance testing, but other than simple force of habit, it is hard to discern any argument for these values rather than other possibilities.

In significance testing, the choice of a criterion for statistical significance (i.e., alpha level) involves a trade-off between the possibility of making Type I errors (i.e., labeling scores as different when they are not)

versus the possibility of making Type II errors (i.e., labeling scores as essentially identical when they are in fact different). This tradeoff is rarely explicitly examined in significance testing, and authors who analyze this tradeoff often reach surprising conclusions about the best choice for alpha levels (e.g., Murphy & Myors ,1998, show that decisions to use stringent alpha levels, such as $p < .05$ or $p < .01$ make sense only if Type I errors are seen as substantially more serious than Type II errors).

The choice of a wide versus a narrow confidence interval for defining bands (e.g., bands could be based alternatively on 95% confidence intervals or on 50% confidence intervals, with the latter indicating that two scores would be accepted as different if the probability that they came from different populations was at least as strong as the probability that they could have come from the same population) is by definition the choice to balance the risks of two sorts of errors (i.e., the error of inappropriately treating them as similar vs. the error of inappropriately treating them as different) in a particular way. To my knowledge, these errors have never been explicitly evaluated or compared in any actual application of banding. Similarly, none of the applications of banding I have seen have attempted to determine whether the balance between Type I and Type II errors implied by the choice of a particular confidence level makes sense. The most pressing research need in the area of banding is to develop useful methods of incorporating well-considered decisions about the risks one is willing to accept in forming bands of different widths.

Choices about the appropriate balance between Type I and Type II errors probably cannot be made in the abstract, but rather are likely to depend on the context in which decisions are made. For example, in some occupations (e.g., information technology workers), the demand for qualified applicants often exceeds the supply. Employers who make choices that result in wide bands (e.g., choosing a .95 confidence level rather than some lower figure) may find it very difficult to identify and pursue the relatively small number of high-potential candidates. On the other hand, if most applicants are well qualified, and if differences between applicants are known to be relatively small (e.g., applicants for graduate school who survive some initial screening often have highly similar qualifications), wide bands might be fully appropriate.

Legal Issues Surrounding Banding

3. Has banding been subject to legal challenge? What has been the outcome?

Outtz and Zedeck: We are not aware of any court decision that has outlawed banding. In three cases in which banding has been at issue

(*Bridgeport Guardians v. City of Bridgeport*, 1991; *Chicago Firefighters Union Local No. 2 v. City of Chicago*, 1999; *Officers for Justice v. Civil Service Commission*, 1992), the pr..mise and logic of banding have been upheld. In contrast, however, what has been successfully challenged is how candidates are selected from within the band. In particular, specific preference for minority candidates has not been supported.

Schmidt: Gutman and Christiansen (1997) have addressed these questions in detail, as have Sackett and Wilk (1994) and Barrett, Doverspike, and Arthur (1995). The two major court opinions on banding seem to make two things clear. First, if banding is not used with minority preferences, the courts appear to have no objections. Second, the courts appear to reject banding with systematic minority preferences.

As long as there are no minority preferences within bands, courts do not distinguish between traditional banding and SED banding. Employers may use either. On the other hand, the rejection of banding with minority preferences is consistent with broader court trends against racial preferences in university admissions, government contracts, and other areas. In fact, even when employers have introduced banding as a means of complying with pre-existing consent decrees, the courts have greatly circumscribed the use of banding with minority preferences (Gutman & Christiansen, 1997). Courts appear to have little interest in banding per se and even less interest in conceptual or logical distinctions between traditional banding and SED banding; instead, their focus is (appropriately) on the question of legally impermissible distinctions based on race.

As pointed out by the three articles cited above, banding without minority preferences does little to reduce adverse impact. So the current stance of the courts appears to block achievement of what is perhaps the major objective of SED banding.

Kehoe: In addition to the other panelists' comments, I offer only one point about the *Officers for Justice v. Civil Service Commission*, 1992 case. In that case, Judge Robert Peckham concluded, "The City has shown that the examinations can reliably differentiate only between the top scorer in a band and candidates below the band... We find the band to be psychometrically sound, and thus, the City has shown that banding is of equal or greater validity than strict rank ordering." Judge Peckham's "finding" that scores within 1.96 *SD*diff of one another cannot be reliably differentiated is a source of considerable concern to organizations that routinely base scientifically and psychometrically appropriate selection practices on the opposite conclusion that any two reliable (nonequal) scores are, to some extent, reliably different. In future cases challenging employment standards or, for that matter, admissions standards, Judge Peckham's conclusion could be used to compel organizations to ignore

reliable score differences that happen to be within 1.96 *SD*diff's of one another.

4. What are the possible legal risks and possible advantages from using banding? How can practitioners avoid these risks?

Outtz and Zedeck: The utility of a selection procedure must be evaluated within the context of overall organizational goals as well as the organization's legal and social responsibilities.

1. Strict rank-order selection may result in adverse impact. Adverse impact, in turn, may lead to:
- Less workforce diversity.
- Increased risk of legal challenge.
- Reduced legal defensibility.
- Increased likelihood that the selection procedure will be perceived as unfair.
- Less likelihood that subgroup differences in selection rates are the result of merit, since differences between subgroups on the selection device may be quite small.

2. Organizations often value diversity. Banding offers one way to achieve greater diversity with minimal if any sacrifice in utility. The cost of legal defensibility must be considered when determining the utility of a selection device. Reducing adverse impact can enhance legal defensibility (or sometimes prevent legal challenge).

3. Nonpsychometric factors have often been used in making selection decisions (e.g., residency requirements, veteran's preference, sons and daughters of alumni, etc.).

4. Banding does not necessarily constitute preferential treatment. Banding can result in preferential treatment if, for example, race or gender or ethnicity is used as the sole determinant for selection within a band.

Advantages of using banding include the following.
- Banding can give managers or selecting officials autonomy in making final selections while limiting their selections to the best applicants.
- Banding can give managers or selecting officials flexibility in finding the right persons (from among the best qualified) for specific jobs.
- Banding can allow a more comprehensive screening of applicants who are qualified based upon an initial limited screen.
- Banding can increase perceptions of fairness.
- Banding can reduce adverse impact (increase workforce diversity).

Risks of using banding include the following.
- Banding can create the false impression that validity and reliability are less important than minimizing adverse impact.
- Bands can be cumbersome to administer, thereby lessening the likelihood of organizational acceptance.

Schmidt: As I indicated, I recently testified in a court case involving SED banding. In that case, Chicago firefighters challenged the use by the City of Chicago of SED banding with minority preferences in promoting firefighters. We do not yet have a decision in that case, and it is clear that whatever the decision is, it will be appealed. However, my experience in this case, along with the three articles I cited in answering Question #3, convince me that SED banding with minority preferences is difficult to defend legally. In my opinion, in today's legal climate the use of banding with minority preferences is an invitation to a legal challenge.

Finally, a reviewer asked that we comment on the reactions of labor unions to banding, pointing out that labor contracts often have detailed stipulations about selection methods. In the court case I mentioned here, it was the firefighters union that mounted the legal challenge to SED based banding, with the contention being that SED banding violated the union contract. This same union has now brought a second court case challenging SED banding.

Kehoe: One aspect of Judge Peckham's decision characterizes both the risks and advantages of banding. Clearly, in this particular case banding was a solution that enabled the intent of the consent decree to be met, technical issues notwithstanding. On the other hand, the risk to other organizations is that Judge Peckham's decision codifies into case law the technical error that scores within a SED band cannot be reliably differentiated. This creates the potential legal argument that any cut score can be challenged on grounds that it cannot be reliably distinguished from scores well below it. The risk is that if case law establishes the legal defensibility of score indifference, organizations may be compelled to ignore certain score differences and to use lower selection standards that alleviate group differences in selection rates in spite of predictable loss in utility.

A major question about banding in general is whether its reception in courts depends on the rationale of complete indifference. Although Judge Peckham asserted that the rationale of indifference was the basis for concluding that banding did not sacrifice validity [sic], would he have been persuaded that reliable score differences could be small enough to warrant the same legal support? Such an outcome would have been more consistent with organizations' reliance on affirmative action plans that are based on the organization's values associated with equal employment opportunity.

Assuming that various banding strategies that acknowledge small real score differences such as described in Kehoe and Tenopyr (1994) can be legally defended, the following represents some of the advantages of banding solutions to selection problems (not limited to SED banding).

1. Banding can avoid the need to explicitly quantify group score differences and, in so doing, avoid the appearance of group-based selection decisions.
2. Bandwidths can be defined based on a variety of potentially competing organization values such as employment cost, equal opportunity, and performance utility to produce easy-to-administer selection processes.
3. Some forms of banding can be used to "govern" selection processes to maximize value while minimizing cost. For example, defining two or three score ranges above some minimum qualification standard can help hiring managers select the best available candidates given some real limit on employment costs.
4. As with the Cascio et al. (1991) application in San Francisco, banding has the potential to meet certain types of court-directed requirements in consent decrees particularly in public sector employment settings.

If the score indifference rationale is codified into case law, the most significant risk is that the score-indifference rationale can be used to compel organizations to ignore useful information and rely on lower than desired qualification standards.

A disadvantage of some forms of banding, such as sliding bands, is that they are difficult to apply in a typical private-sector employment context. Private-sector, high volume employment is usually continuous but with a ceiling on the costs. In this setting, banding strategies that interactively adjust the selection rule as a function of who has already been selected are not likely to be practical.

Murphy: The most obvious risk is that it would be relatively easy to exploit banding to make testing practically meaningless in most organizations. As psychologists, we have ethical obligations that might help limit the abuse of this strategy, but many other players in selection (e.g., unions, attorneys, interest groups) would have a strong incentive to use banding to essentially remove testing from selection. Unfortunately, it would be pretty easy to do this.

If relatively wide confidence intervals are used in conjunction with unreliable tests, it might be impossible to say that any pair of applicants (at least within a wide range of scores) is really different. For example, a good plaintiff's attorney could wipe out a testing program by insisting that people be careful in making distinctions (e.g., use a 99% confidence interval) and by insisting on a relatively unreliable test (Murphy et al., 1995).

A second risk is that concerns about the fairness of banding are likely to be linearly related to its effectiveness in advancing its goals (again, banding is most likely to be used in an attempt to increase diversity). The wider the band, the greater the opportunity to use things other than test scores to differentiate among applicants. However, the wider the band,

the less likely it is that people will accept the argument that scores within the band are not really different. On the other hand, a dogmatic insistence on top-down selection in contexts where the differences among applicants are relatively small runs the risk of overemphasizing trivial differences in test scores, and perhaps ignoring larger differences on other criteria.

Guion: To answer this one, I need to expand on the ideas that produced the answer to the first one. If it is decided that there is a range of indifference (i.e., a band of scores in which any differences among candidates are deemed trivial), then a further management decision must be made about a policy for choice within a band. If the number of people in the top band is fewer than the number of openings, then there is no problem; unless a member of that top band has some disqualifying characteristic (e.g., a felon applying for police work), everyone in that band will be offered a job. The problem emerges when there are more people in a band than there are openings. Those to be offered jobs need to be chosen on some basis. If the offers go to those at the top of the band, then the existence of bands is meaningless; it is top-down selection. Perhaps offers could go to randomly chosen candidates, but that would be a hard sell. I advocate making offers to people who have demonstrable strength in something relevant to job performance or important to the organization, and not making offers to people with demonstrable weaknesses or disqualifying (or nearly so) characteristics.

A search for diversity might be important to the organization, but diversity should not be restricted to demographic diversity (e.g., race, sex, age). Diversity can and should include diverse skills or background experiences that have some relevance to the job or organizational functioning, but which are identifiable only infrequently and therefore cannot be included in some overall multiple regression. (My personal favorite example is the freshman engineering student I encountered more than 50 years ago. He had dismal scores on all academic aptitude measures, but he was getting a monthly income of $400—that was 1949 dollars—from royalties on his patents.)

In short, my view is that people who look at banding only in terms of affirmative action for racial or sex minorities are too myopic. Now I can give my answer to the question.

The risk is that the bases for choices within a band are arbitrary and ad hoc and can lead to charges of unfairness that, even if not legally actionable, can give the organization a bad name. The solution is to develop, in advance of making a test or test battery operational, a set of variables to investigate for candidates within the band, or to establish procedures in advance for identifying unforeseen information as either positive or negative. Otherwise, if a piece of information is regarded

today as positive, but is not regarded as positive tomorrow, then a candidate tomorrow with that characteristic can call foul. If that person is in a protected class, the lack of system is actionable.

The "possible advantages" include the notion that the organization might actually get some good people that it might otherwise miss, and that it can develop a more functionally diverse workforce.

Recommendations Whether to Use Banding

5. Before recommending whether banding should be used, what are the possible points of agreement among this group of experts? For example, would we all agree that it is a common practice to aggregate similar scores for interpretation purposes (e.g., cutoff scores, percentile ranges, stanines, etc.), or that there are ranges of observed scores that are similar enough that other considerations should be taken into account?

Outtz and Zedeck: We would agree that it is a common practice to aggregate similar scores for purposes of interpretation. We would also agree that there are ranges of scores that are small enough to warrant consideration of other factors in making selection decisions. We agree, too, that the larger the score differences on a valid test or selection device, the greater the likelihood that there are meaningful differences in the capabilities of the applicants to perform those aspects of the job predicted by the test. We agree that statistical data need not be the only factor in determining bandwidths.

We disagree with the proposition put forth by some that statistics such as the SED should not be used in determining how large of a difference in scores is meaningful. It would seem to us that statistical data should be considered in setting bands, if for no other reason than to set an outer limit for the bandwidth.

As stated above, it appears as though this group of experts agrees that banding is a viable method for test score use. The disagreement is on how to form the bands; whether there should be a psychometric rationale, a decision-theoretic rationale, or an arbitrary basis. Another area of perhaps disagreement would be on what secondary criteria should be used. For example, should minority status be one of the factors to use to select from within the band? It is our position that a diverse workforce is an economic and practical necessity in a number of situations, and that secondary criteria should include measures that lead to an increase in the workforce diversity.

Schmidt: I would think that we could all agree that traditional (non-SED) banding and related procedures have been used in personnel selection for decades and that their use has probably simplified understanding and administration of personnel selection systems for HR man-

agers and other managers. We could perhaps also agree that because these procedures depart at least somewhat from optimal top-down selection based on the linear model, they have entailed some loss of selection utility, although this loss has often been small. There is probably also general agreement that banding without minority preferences—whether traditional or SED banding—typically does little to reduce adverse impact.

However, some may not agree with this important conclusion: There is a fundamental difference between traditional, non-SED banding and SED banding. SED banding, unlike traditional banding, is logically flawed, as explained in my response to Question #1, and should not be used.

Kehoe: The fundamental premise of all banding, including SED banding, is that score differences can be small enough to be outweighed by other considerations in making selection decisions. I believe we all agree on this fundamental point. The question is, "How small is small enough that other factors can determine selections?" On this point, I believe we all also agree that any form of banding should be evaluated based on whether it achieves the goal of appropriately balancing competing interests. I believe we all also agree that there is no absolute or even conventionally accepted bandwidth that is "correct" in any sense for all applications. Even Cascio et al. (1991) apparently would not have pursued SED banding, technical issues notwithstanding, if it did not help the organization satisfy the requirements of the applicable consent decree. Their own evaluation of SED banding focused on the extent to which desired outcomes were achieved without sacrificing too much benefit. Similarly, Murphy and Myors (1998) who appear to embrace the basic concept that classical hypothesis testing conventions can be used to define bandwidths nevertheless ultimately treat the bandwidth decision as one that depends on the extent to which the desired outcomes are achieved.

Although this may appear to be a trivial point of agreement, it is not. I believe that it is an important point that all participants in this debate share the fundamental view that the appropriateness of any bandwidth is based on the extent to which resulting tradeoffs produce desired results. The primary contribution of our science is to provide the tools to measure the tradeoffs (and the forum and common language for this debate). Classical hypothesis testing conventions are not the primary contribution of our science to banding practices.

Murphy: Possible points of agreement include the following:

1. It is common practice to aggregate scores.
2. It is not optimal to do so, but very little else we do is driven by strict optimality criteria.

3. If you are going to aggregate, it makes more sense to take into account the reliability of the test than to ignore it.
4. You should be able to articulate the criteria used to make important decisions about banding (e.g., fixed vs. sliding bands, wide vs. narrow confidence intervals).
5. There are usually better ways than banding to accomplish the goal of articulating factors other than test scores into selection decisions.

Guion: I would agree to both propositions, but I have reservations about limiting the number of bands to two big ones, as is done when cut scores are established. In general, I think bands should be small—much smaller than most of the discussions have had them.

6. What would you personally recommend regarding whether or not to use banding?

Outtz and Zedeck: We would recommend the use of bands in most hiring and promotion decisions. The width of the band used should be determined by factors such as:

1. The criterion space covered by the selection device(s).
2. The amount and quality of validity evidence.
3. The consequences of errors of prediction.
4. The reliability of the selection device(s).
5. Concerns for diversity.

Narrow bands are preferable to wide bands.

Schmidt: I recommend that employers not use SED banding at all. I do not object to traditional banding in situations in which professional judgment indicates that the gain in simplicity and understandability of selection systems outweighs the loss of selection utility resulting from limited departure from top-down selection. In such situations, the traditional bands will be fairly narrow, because wide bands will usually result in substantial utility losses, given valid selection methods.

I recommend that if additional factors are to be considered in hiring that these factors be quantified and included as part of each applicant's total score. Research indicates that the alternative of incorporating these factors using subjective judgment will lead to lower validity.

Regardless of type of banding used, or width of band, I do not recommend minority preferences within bands because (among other problems) such use is an invitation to a legal challenge in today's legal climate.

Kehoe: I do not recommend *SD*diff banding as defined by Cascio et al. (1991) because of the flawed definition of the appropriate bandwidth.

Rather, in employment settings where banding would solve selection problems, I recommend that bandwidths be defined based on organization values including efficiency, performance utility, and social values.

(See Kehoe & Tenopyr, 1994, for a description of procedures for defining value-based bandwidths.)

Murphy: I would recommend banding only if I could make a credible argument that the band is not too wide. That is, I should be able to make an argument that is both scientifically credible and acceptable to decision makers and applicants that the differences between scores within a band are really small enough that they could be sensibly ignored. To do this, I think I would have to be able to articulate why I am giving more weight to some sorts of errors than others (see point #2).

I would also have to be able to make a credible argument that the criteria used to distinguish among individuals within the band were legitimate ones. For example, if I use banding to increase diversity, I ought to be able to articulate why increasing diversity is worthwhile to this organization (this explanation needs to go beyond platitudes about the general value of diversity). If I could make both arguments with a straight face, I would have no problem recommending banding.

Guion: I strongly recommend the use of banding where management can agree on narrow bands (ranges of indifference), where the bands are small, and where a variety of considerations are clearly articulated to guide choices within a band where there are more candidates than openings. There is no reason in this view to think that all bands (at least when using fixed bands) will be the same size; a region of indifference may be much larger in the middle of a distribution, or perhaps at the low end of the distribution, than at the high end.

Note that in this view, banding is not simply for hiring more minorities. Actually, it probably won't work for that anyway if bands at the high end of a distribution are small enough that differences within them actually do not matter. Banding is in my view a general tool to force more thought and data into staffing decisions.

7. In your opinion, what are the best alternatives to banding to reduce adverse impact (and enhance diversity) yet still have valid selection? Your answer might include a comparison to traditional approaches like top-down ranking and cutting scores, but also consider alternative ways of using test scores, alternative types of selection procedures, changes to other processes (e.g., recruiting practices), or any other advice for obtaining the dual goals of high validity and low adverse impact. Please be sure to indicate your preferences among these methods.

Outtz and Zedeck:

- Utilize several selection devices as opposed to a single selection device (particularly when the single selection device is a paper-and-pencil or multiple-choice test).

- Attempt to account for as many aspects of job performance as possible.
- Attempt to utilize assessment devices that cut across assessment methods in terms of the manner in which content is presented and the type of response required. We have found that selection devices that allow oral responses typically have less adverse impact than those that (a) require a written response, or (b) require the respondent to choose from among several written responses.
- Utilize targeted recruitment rather than the shot-gun approach of trying to test as many applicants as possible.
- When a selection device that has high adverse impact must be used, the adverse impact can be reduced by setting a cut point that minimizes adverse impact to the greatest extent, then conducting further screening based on selection devices that have less adverse impact.

This question presumes that the basic reason for banding is to reduce adverse impact. We argue (a) that banding may have been promulgated on the basis of seeking means to reduce adverse impact, and (b) that the results of banding may have in fact been less adverse impact in particular situations. Nevertheless, we propose that banding be considered in any selection or promotion situation in which you want to consider secondary characteristics of candidates. We state this because we recognize that even when we have the best validity coefficients, more than half of the performance variance remains unexplained.

Schmidt: John Hunter and I discussed this question in a recent article (Hunter & Schmidt, 1996). Recent research has shown that noncognitive measures, such as integrity tests and measures of the personality trait of conscientiousness, are valid predictors of job performance (e.g., see Schmidt & Hunter, 1998). Additional noncognitive measures are currently being created and evaluated. Unlike measures of general mental ability (GMA) and specific abilities, these measures typically show no racial or ethnic differences in mean scores. Our position is that employers have an *obligation* to use such valid noncognitive measures in selection along with GMA measures, because this approach reduces adverse impact while simultaneously increasing validity and increasing the job performance of those hired.

Although this approach reduces adverse impact, it does not typically eliminate it. Because of pre-existing job related differences between groups, there is probably no feasible way to completely eliminate all group differences in hiring rates through choice of selection methods while maximizing (or even maintaining) validity. Complete elimination will require wider social changes (Schmidt, 1988).

Kehoe: With the emergence in the past 5–10 years of substantial evidence of interview and personality validity, the most promising strategy for minimizing adverse impact is also the most promising strategy for maximizing validity. The most promising strategy is the inclusion of noncognitive predictors such as personality assessment and experience evaluation in some form of compensatory selection process with cognitive ability, job knowledge assessment, and/or other selection criteria that cause group differences in selection. This is most promising in the sense that it addresses the group difference issues by seeking to improve the job relevance of the overall selection strategy.

This strategy is not the same as substituting personality and experience for group-affecting selection criteria. The fact that similar validity values may be reported in meta-analytic studies of different types of selection procedures does not mean that those different selection procedures would produce the same benefit or be indistinguishable from an organizational perspective. The possibility that interview/experience validity is similar to that of cognitive ability should not be taken to mean that they are necessarily substitutable.

Murphy: I think it is more honest and more efficient to make explicit judgments about the criteria that should be used to evaluate applicants and about the relative weights that should be assigned to these criteria. Banding is usually a roundabout method of introducing criteria such as diversity, under the sometimes dubious argument that they are used only when peoples' scores on selection tests are tied, or nearly tied. It may make sense to develop special rules for handling ties in an efficient and equitable fashion, but in most cases, we would all be better off if organizations articulated what they value in applicants and why, and if those values and preferences were made part of a formal selection scheme that applied equally to all applicants.

Guion: The best ways have been illegal since 1991, for example, separate rank-order lists, different tests, and so forth. These weren't good ideas anyway, but they might have served the dual purpose. I apologize, but I won't really answer this question because it requires data I have not seen. For example, it is asserted (usually without data) that certain kinds of noncognitive tests have high validity and little or no adverse impact. Research needs to be done to determine whether the inclusion of such tests along with the cognitive tests usually known to be valid will serve that dual purpose. Until we as a profession and the social advocates on both sides of the affirmative action mess look to data instead of ideology, we won't have a basis for advocating one method over another.

Practical Issues Surrounding Banding

8. If I were to use banding, how would I actually go about it? For example, how should I decide on the width of the bands and what criteria should be used to select people within the band?

Outtz and Zedeck: Regarding the width of the band, we recommend using the formula for the standard error of the difference (SED; see Cascio et al., 1991), but also to choose the critical value for the confidence interval based on the *risk* involved in an incorrect decision. That is, in high-risk situations, such as fire and police work, the band should be relatively narrower than for low-risk positions (e.g., clerical positions). We consider the hire of a candidate who turns out to be ineffective to be more problematic for the organization and society when the job involves police work, for example, than when it involves clerical duties such as filing, copying, and so on. To select the critical value to be used in the SED equation, increasing alpha (e.g., $p < .10$) results in a smaller band. The rationale and illustration of adjusting alpha to impact bandwidth is discussed in Zedeck, Cascio, Goldstein, and Outtz (1996).

Regarding secondary criteria, we recommend that the organization identify factors or characteristics that (based on job analyses) are relevant and desirable and for which information can be gathered from candidates. For example, candidate relevant prior experience, certification of relevant courses, useful additional skills such as facility in a foreign language, and other such factors can be used as secondary criteria.

Some have argued that the use of secondary criteria results in the use of less valid and reliable (or perhaps invalid and unreliable) measures. We stress that the measures must be job relevant, but might be measures that are not required for all in the organization (e.g., bilingual ability, special training as a paramedic). In this way, we propose use of banding with valid and reliable measures, but increasing our ability to measure more of the criterion space by use of *job-relevant* secondary criteria.

Regarding difficulties using banding, the most significant problem we have come across is employers who go to the trouble of setting up bands based upon valid and appropriate criteria, then attempt to hire or promote persons from a lower band ahead of persons in a higher band when they feel the need arises (e.g., affirmative action or legal defensibility). This practice can usually be discouraged by pointing out the possibility of charges of reverse discrimination.

Schmidt: See my answer to Question #6.

Kehoe: Like many organizations, we currently use a form of banding to optimize the combination of the competing values of administrative efficiency, performance utility, and group impact. This form of banding is very common among private sector organizations that use test-based

selection procedures. A single band is defined by the determination of a cut score. Candidates who score at or above the cut score are eligible to be selected; those who score below are not. This band is not based on any consideration of score differences. Rather it is based on the level of expected performance of candidates at the lower end of the band, the yield rate and cost of employment resulting from the choice of cut score, and the impact of the bandwidth on group differences in selection rates. An alternative approach is to first establish a minimum standard below which no candidate may be selected and then divide the range above the cut score into two or three (typically) bands of scores. Selection takes place from the top band until it is exhausted at which time selection takes place from the next lower band, and so on. Throughout this process the decision can always be made to refresh depleted bands so as to maximize the skills of new hires. When selecting within bands, score information is ignored in favor of other considerations.

If banding in either of the above forms or in some other form were implemented primarily to accommodate the organization's interest in equal employment considerations, I recommend defining the bandwidth based on a consideration of probabilities of job success as described in Kehoe and Tenopyr (1994) Bandwidth is defined by the range of observed scores predicting a range of job success probabilities that the organization is willing to treat as the same in return for the social value of reduced group differences in selection rates.

9. How would I explain banding to applicants, especially unsuccessful applicants? How would I explain banding to line managers and other decision-makers?

Outtz and Zedeck: The explanation we would give is simply that selection is not a perfect process. Therefore, applicants/candidates who are similarly qualified will be given further consideration on the basis of additional factors that are important to the organization. We make sure that these additional factors are clearly spelled out for all concerned. We also make sure that the additional factors are job related and consistent with organizational values.

Results of research on reactions to banding (Truxillo & Bauer, 1999) suggest that applicants should be made aware of the use and purpose of banding, and that the organization should emphasize the psychometric logic of banding and the need for diversity in the organization. Such information should be presented in the application materials or in sessions that orient applicants to the selection process.

It is our assumption that in situations where unions have contractual agreements with regard to hiring issues, the use of secondary criteria should pose no problem for their acceptance. Unions often want to see

factors such as seniority and experience rewarded, and using such factors as secondary criteria would meet their concerns. We are personally familiar with situations in which the union has agreed to banding, where its members sit on panels to review the candidates for their possession of the secondary criteria.

Schmidt: One problem with SED banding is that no legitimate explanation can be given to applicants who are unsuccessful because their scores fall outside the band (Schmidt, 1991). For example, consider an applicant who was not selected because her score was three points below the lower end of the SED band. That applicant would be told that she was not selected because her score was statistically significantly lower than the highest score. However, she could then point out that her score was not significantly different from the scores of *most* of the people in the band who were hired. So why was she not hired, too? Because of the fundamental logical contradiction within SED banding, there is no honest or truthful answer that can be given to this question. The basic principle of SED banding is that scores that are not significantly different should not be treated differently, and in fact should be considered as equivalent. (See my response to Question #1). Yet, if that principle is applied here, then the conclusion is that this applicant should have been hired.

How do advocates of SED banding handle this question? Essentially, they must say the following: "I understand your position, but in order to make SED banding workable we have decided to apply this fundamental principle only to comparisons between the highest score and other scores. Those are the only comparisons we make." However, this limitation on application of what is presented as a universal principle is purely arbitrary, and in fact contradicts the universal nature of the principle.

Hence there is no way that SED banding *can* be honestly explained to unsuccessful applicants, because no such explanation is possible.

On the other hand, traditional banding *can* legitimately be explained to unsuccessful applicants, along the following lines: "The scores of applicants for this job range from 12 to 110. From a practical point of view, this is a lot of scores to deal with. We have decided to simplify administration of this selection system by grouping people into 5-point intervals or bands. The reason you were not hired is that your score fell into a band lower than the bands we selected from." Note that in this case there is no counter argument similar to the one above that the unsuccessful applicant can make. This is true because traditional banding is not based on a statistical significance rationale.

The same principles apply in explaining banding to line managers and other decision makers: There is no noncontradictory explanation that can be provided for SED banding, and traditional banding can be

explained easily as a procedure for simplifying administration of the selection system.

Kehoe: I would explain to applicants, particularly internal candidates, that the selection process is based on a number of important considerations. Lower levels on one of these considerations may be outweighed by higher levels on one or more of the other considerations. I would not attempt to provide specific technical information about band definition or about the precise algorithm used, if there is one, to balance competing considerations.

Part of the challenge for selection program managers is to explain how and why other factors than the one measured by the banded scores influence selection decisions. As described above, applicants might get a general explanation. However, business leaders/managers whose positions are being selected for frequently demand more detailed explanations. A point to emphasize to this particular audience is that the importance of some factors that influence selection decisions depends on their job relevance. Attributes that are tested for, such as cognitive ability and personality attributes, are of this sort. For this discussion, I'll refer to these as "validity" factors. But other factors are important for other reasons than job relevance. For example, factors such as seniority, date of application, and diversity of workforce can be important to organizations for a variety of reasons. I'll refer to these as "organization" factors. In general, banding provides a fairly straightforward process for relying on more than one type of consideration. The general rationale is that among similarly qualified applicants (as measured by some composite of "qualifying" scores) within a band, selection decisions are made based on other types of considerations. This is generally not difficult for managers to understand. The hidden question that can be difficult is the decision about which factors are combined into a single composite score, or multidimensional scores, to define the bands within which other factors will determine selection decisions.

My recommendation about this point is that if "organization" factors are intended to influence individual selection decisions, they should not be combined with "validity" factors to determine bands. Rather, the "organization" factors should be the basis for selection decisions within bands based on "qualification" factors. Not only is this strategy relatively easy to explain, it also eases somewhat the problem of explicitly determining relative weights comparing the importance of "validity" factors and "organization" factors which may not be commensurate. The importance of "validity" factors can be reasonably expressed in utility terms whereas the importance of, say, seniority is in terms of social or organizational values that are not easily captured in units of utility.

Separately, labor unions are another constituency that some organizations must address when explaining selection procedures. Regarding selection issues, the primary advocacy of unions is that employee seniority be an important consideration, if not the only consideration. There is no inherent conflict between banding and the value placed on seniority. In fact, banding can provide an explicit mechanism for managing the role and importance of seniority in making selection decisions (e.g., a selection practice could incorporate the rule that within a band candidates are chosen in rank order of seniority). However, unions can also value simplicity and consistency of selection practices. A banding strategy, such as sliding bands, by which the minimum standard for selection varies over time or conditions, may cause concern with unions if it gives the appearance that the selection standards are not the same for all candidates.

Murphy: I would go back to themes raised under point #6. Your ability to explain banding hinges largely on your ability to make a credible and acceptable argument that test scores within a band are close enough that they can be treated as essentially identical. If you cannot sell this point, you will not be able to sell anything else about banding. On the other hand, if key stakeholders accept your argument that the band is reasonable in size, it should be possible to sell banding. You may still have difficulty explaining or justifying particular aspects of a proposed banding program, and if you do have this difficulty, that can be treated as a diagnostic test. Banding programs that cannot be explained or justified to most of the stakeholders in personnel selection are probably not good programs, and the more difficulty you have in putting together an honest and convincing explanation, the more likely it is that you have incorporated undesirable or arbitrary features into your banding system.

Guion: I wouldn't explain it to applicants because I would assign a different numerical value, or score, to each band. If there are 20 bands, I would assign a score of 20 to the top band, a score of 19 to the next, and so on. And if my views were unexpectedly to win out, there would be no need to explain them to managers and other decision makers because they would have developed the bands.

10. What is going on in practice right now? Are many organizations using banding, either in the private or public sector?

Outtz and Zedeck: Our experience is that most organizations in the public and private sectors use some form of banding, although they don't refer to it as banding. As an example, many organizations establish categories of candidates, such as "best qualified," "well qualified," "qualified," and "less qualified." Other organizations simply establish two bands based upon a pass/fail cutoff. Public sector employers tend to

have more structured and codified procedures for setting bands, such as a Rule of Three (meaning a band consisting of the top three scores) or a Rule of Five, and so forth.

Schmidt: My consulting experiences and conversations with people in organizations suggest to me that since the Civil Rights Act (CRA) of 1991 went into effect, there has been increased interest in banding and increased use of banding. My sense is that this increase is greater in the public sector than the private sector. The 1991 CRA made the adjustment of test scores based on race or ethnicity illegal. Many state, local, and municipal governments (such as Chicago) had been using such score adjustments to eliminate disparate impact. When this procedure had to be discontinued, state and local governments looked to banding as an alternative way to reduce disparate impact. Without some such alternative, they would have suddenly gone from zero disparate impact to a high level of disparate impact, especially for Blacks and Hispanics.

This process has been weaker and less important in the private sector because in the private sector selection systems typically allow for the injection of subjective factors—including "diversity considerations"—into the selection process. These factors are subjectively combined with score information in a way that is functionally equivalent to score adjustments, but since there is no written record of score adjustments, this process technically does not run afoul of the 1991 CRA. In the public sector, on the other hand, such unseen subjective "mental score adjustments" are forbidden by Civil Service System rules that call for objectivity in examinations and evaluations. Hence, the greater attraction of banding in the public sector.

Public sector jurisdictions that have been attracted to banding have typically been attracted to SED banding, not to traditional banding. My observations lead me to believe they are attracted to SED banding because it exudes an aura of being scientific due to the fact that its foundation is statistical significance testing. Many naïve people believe that significance testing is "scientific" (Schmidt & Hunter, 1997; Schmidt, 1996). In public sector selection, if an alternative procedure is going to be weighty enough to force the jettisoning of the traditional rule of top-down selection, then it has to be a real heavyweight (i.e., have a lot of "scientific" weight). Traditional banding methods, not being based on significance testing, appear much more lightweight to such people.

A second reason is the appearance of objectivity. A major purpose of the reforms that produced Civil Service systems of selection was the elimination of subjectivity and subjective bias in selection decisions. SED banding, because of its use of significance testing, gives the impression of being objective and hence not being subversive of merit hiring principles. In fact, SED banding may appear to be as objective as, or

even more objective than, the top-down hiring rule. Hence it is a useful tool in undermining that rule because one can ague that one has not abandoned objectivity. This is ironic in that within the wide bands thus created, subjective nonmerit factors are then allowed to strongly influence who is hired. This is most obvious in the case of SED banding with minority preferences, but is also true even in the absence of minority preferences.

A third reason for the preference for SED banding is that SED banding produces very wide bands, allowing a great deal of discretion in selection decisions. Traditional banding, by contrast, usually produces narrow bands.

This trend toward SED banding is currently being met with legal challenges, as discussed earlier in response to Question #3.

Kehoe: I doubt we know the answer to this. My general impression is that few if any private sector organizations are using SED banding in the manner proposed by Cascio et al. (1991). Likely the most common forms of banding are the types I described in response to Question #9.

Guion: I don't really know, but I doubt that there is much banding being done under that name. There are lots—too many—organizations using 2-band cut score systems, and there may still be some who set up expectancy charts (which usually have four or five bands), but I don't know of many that have even thought of using bands defined by SEMs (for which I'm grateful).

REFERENCES

Aguinis H, Cortina JM, Goldberg E. (1998). A new procedure for computing equivalence bands in personnel selection *Human Performance, 11*, 351–365

Barrett GV, Doverspike D, Arthur Jr W (1995) The current status of judicial review of banding: A clarification *The Industrial-Organizational Psychologist, 33*(3), 39–41.

Bridgeport Guardians Inc v. City of Bridgeport (CA 2, 1991) 933F.2d 1140.

Cascio WF, Goldstein IL, Outtz J, Zedeck S. (1995). Twenty issues and answers about sliding bands. *Human Performance, 8*, 227–242

Cascio WF, Outtz J, Zedeck S, Goldstein IL. (1991) Statistical implications of six methods of score use in personnel selection. *Human Performance, 4*, 233–264.

Cascio WF, Zedeck S, Goldstein IL, Outtz J (1995) Selective science or selective interpretation? *American Psychologist, 50*, 881–882.

Chicago Firefighters Union Local No. 2 v. City of Chicago. (1999). WL 1289125.

Cohen J. (1994). The earth is flat ($p < 05$) *American Psychologist, 49*, 997–1003.

Coward WM, Sackett PR (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75*, 295–300

Gottfredson LS. (1994) The science and politics of race-norming. *American Psychologist, 49*, 955–963.

Gutman A, Christiansen N. (1997) Further clarification of the judicial status of banding. *The Industrial-Organizational Psychologist, 35*, 75–81

Hanges PJ, Grojean MW, Smith DB. (in press) Bounding the concept of test banding: Reaffirming the traditional approach *Human Performance*

Hunter JE. (1997) Needed A ban on the significance test. *Psychological Science, 8*, 3–7
Hunter JE, Schmidt FL. (1996). Intelligence and job performance· Economic and social implications *Psychology, Public Policy, and Law, 2*, 447–472
Kehoe JF, Tenopyr ML. (1994). Adjustment in assessment scores and their usage A taxonomy and evaluation of methods. *Psychological Assessment, 6*, 291–303
Loftus GR (1996) Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–171
Murphy KR. (1994). Potential effects of banding as a function of test reliability PERSONNEL PSYCHOLOGY, *47*, 477–495.
Murphy KR, Myors B (1995) Evaluating the logical critique of banding *Human Performance, 8*, 191–201
Murphy K, Myors B (1998) *Statistical power analysis· A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ Erlbaum
Murphy KR, Osten K, Myors B (1995) Modeling the effects of banding in personnel selection, PERSONNEL PSYCHOLOGY, *48*, 61–84
Officers for Justice v Civil Service Commission (CA9, 1992) 979 F 2d 721, cert denied, 61 U S L W 3667, 113 S Ct 1645 (March 29, 1993)
Sackett PR, Roth L (1991) A Monte Carlo examination of banding and rank-order methods of test score use in personnel selection *Human Performance, 4*, 279–295
Sackett PR, Wilk SL (1994) Within-group norming and other forms of score adjustment in preemployment testing *American Psychologist, 49*, 929–954
Scientific Affairs Committee (1994) *An evaluation of banding methods in personnel selection* Arlington Heights, IL The Society for Industrial and Organizational Psychology
Schmidt FL. (1988) The problem of group differences in ability scores in employment selection *Journal of Vocational Behavior, 33*, 272–292
Schmidt FL (1991). Why all banding procedures are logically flawed *Human Performance, 4*, 265–277
Schmidt FL (1996). Statistical significance testing and cumulative knowledge in psychology Implications for the training of researchers. *Psychological Methods, 1*, 115–129
Schmidt FL, Hunter JE (1995) The fatal internal contradiction in banding Its statistical rationale is logically inconsistent with its operational procedures *Human Performance, 8*, 203–214
Schmidt FL, Hunter JE (1997) Eight common but false objections to the discontinuation of statistical significance testing In Harlow L, Mulaik S, Steiger JH (Eds.), *What if there were no significance tests?* Hillsdale, NJ Erlbaum
Schmidt FL, Hunter JE (1998) The validity and utility of selection methods in personnel psychology Practical and theoretical implications of 85 years of research findings *Psychological Bulletin, 124* (2), 262–274
Siskin BR (1995) Relationships between performance and banding *Human Performance, 8*, 215–226
Truxillo DM, Bauer TN (1999) Applicant reactions to test score banding in entry-level promotional contexts *Journal of Applied Psychology, 84*, 322–339.
Zedeck S, Cascio WF, Goldstein IL, Outtz J (1996) Sliding bands An alternative to top-down selection In Barrett R (Ed), *Handbook of fair employment strategies* (pp 222–234) Westport, CT· Quorum
Zedeck S, Outtz J, Cascio WF, Goldstein IL. (1991) Why do "testing experts" have such limited vision? *Human Performance, 4*, 297–308