"Test Score Banding: The Good, The Bad, and The . . ."

A Review of the Research Literature

Lawrence F. Lipson

Michigan Employment Security Agency

Northern Assessment Research & Development Center

Detroit, Michigan

Presented at the 21st Annual Conference of the

IPMA Assessment Council on Personnel Assessment

Hyatt Newporter Hotel

Newport Beach, California

June 25, 1997

<u>FOCUS OF THIS PRESENTATION</u>

I will be talking with you this afternoon about test score banding. Before I begin, I want to tell you that I don't profess to be a expert in this area. This session will be for information only - to impart to you some of what I have learned about banding from two perspectives:

[SHOW SLIDE #1]

(1) a <u>review of the literature</u> and (2) information gleaned from a couple of <u>practical applications of banding</u> in state government.

By way of quick background, I am a member of IPMAAC and this is the fourth national conference I have attended since 1993 and the third time that I have been a presenter. I was with you in Sacramento in '93, New Orleans in '95, Boston last year, and now here in beautiful Newport Beach this year. I've worked for the State of Michigan, Michigan Employment Security Agency for 22 years. And for the past 15 years, I have been part of a national assessment research and development program funded by the U.S. Department of Labor (DOL).

Currently, I am a Senior Test Research Analyst. Today's topic on test score banding for IPMAAC '97 is in no way related to any DOL project and has no association with any current or anticipated national policy from DOL. The objectives of my presentation are as follows:

[SHOW SLIDE #2]

(1) to <u>describe what banding is</u>, (2) to talk about various <u>types of banding</u>, (3) to discuss some <u>advantages of using banding</u>, (4) to discuss some <u>disadvantages of using banding</u>, and (5) to present some information about <u>how banding is currently being used</u> in practical applications. This will be from the eyes of researchers and practitioners in the field.

I will address the following questions:

[SHOW SLIDE #3]

Is banding a legitimate scoring mechanism to combat adverse impact?

Or is banding "fatally flawed" in a logical sense, as one researcher claims (Schmidt, 1995)? There are various strategies of banding. The permissibility of these strategies under the Civil Rights Act of 1991 has not been resolved. What, then, is the bottom line on use of banding?

WHAT IS BANDING?

First of all, what is this thing called banding anyway? Banding refers to a set of approaches to the use of tests where people whose scores fall within a specific range, called the ***band***, are regarded as having the same score. It's a way of breaking ties in test scores. Whatever construct, dimension, trait, or characteristic is measured by a particular test, everyone whose scores fall within a particular band are regarded as having the same score on that construct, dimension, trait, etc.

An example of banding is the use of stanine scores. This is where raw scores on a test are converted to single digit scores ranging from 1 through 9 (Sackett & Wilk, 1994). These scores of 1 through 9 represent deciles in the range of scores. They also each represent a ***band***. All of the individual raw scores within each band comprise the ***bandwidth.***

Paul Sackett and Steffanie Wilk describe the basics of banding as part of an article in *American Psychologist* (Sackett & Wilk, 1994). In talking about bandwidth, they say that bandwidth could be determined arbitrarily, but "it is common to define bandwidth in terms of the standard error of the difference between scores. . . . A band could be defined to include scores [that are] not significantly lower than the top score in the band, with bandwidth dependent on the level of significance chosen." Bandwidth depends directly on the measurement precision of the test (Murphy, Osten, & Myors, 1995).

Dr. Kevin Murphy of Colorado State University (Murphy, 1994) reiterates that “the width of a band is a function of the standard eror of measurement (SEM), more specifically the standard error of the difference bewteen two test scores (SED).” He also points out that the percentage of candidates who are included in a given band depends on where that band is positioned relative to the total distribution of test scores.

For example, “if the band is near either tail of the distribution, rather than being centered at the mean, it will include fewer applicants . . .” (Murphy, 1994). The effects of banding also depend upon what selection ratio is used. “When the [selection ratio] is either very high or very low, banding will affect a smaller group of applicants. Conversely, the effects of banding are greatest (in terms of the number of individuals whose selection might be affected) when the [selection ratio] is near .50. [That’s] where the applicant distribution is the densest” (Murphy, 1994).

If candidates for a job or a training program are to be ranked, all people whose test scores fall within the same band could then be ranked for selection on some other basis, such as relevant prior experience or training. And if this is the case, the chances of selection for members of the lower scoring group can increase substantially (Murphy, Osten, & Myors, 1995).

Essentially, banding is an alternative to top-down selection of candidates by test score. There are various methods of banding. Depending on which method is used, and other factors that I’ll get into in a little while, banding can reduce the adverse impact of tests that are used for personnel selection (Cascio, 1995; Murphy, 1994; Murphy, Osten, & Myors, 1995; Sackett & Wilk, 1994). We all know that within-group norming, more commonly known as “race norming” of test scores was rendered illegal under the Civil Rights Act of 1991 (CRA ‘91) and cannot be used in today's employment selection practices. In the post-CRA' 91 era, the intent of banding, i.e., to reduce adverse impact, is just as politically correct as within- group norming was back in the 1980's (Schmidt, 1995).

METHODS OF BANDING

In the simplest and purest sense, there are only two methods of banding, fixed bands and sliding bands.

[SHOW SLIDE # 4]

However, within these two broad methods, are several more specific variations. Let's explore some of these variations. (The following information can be found in Sackett and Roth, 1991.)

[SHOW SLIDE # 5]

First, **Fixed Bands with Random Selection**. With this method, each candidate in a band has the same chance of being selected. If the first band is exhausted before enough candidates have been selected, then candidates are selected randomly from the second band, and so on through the bands until enough candidates have been selected.

A variation of the fixed bands with random selection method is **Fixed Bands using Minority Preference with Top-Down Selection.** In this variation, minority candidates (or the candidates whose scores fall in the lower group, as compared to a group of higher scoring candidates) are selected top-down within the first band until the band is exhausted of candidates in the lower scoring group. Then, candidates in the higher scoring group are selected from the band in a top-down order. This variation has also been called *Fixed Band- Diversity Based* (Cascio et al., 1991).

A third variation of using fixed bands is **Fixed Bands using Minority Preference with**

<u>Random Selection</u>. In this approach, candidates in the lower scoring group are *randomly* selected from the first band until either: (a) the band is exhausted of candidates in the lower scoring group, or (b) the selection ratio for candidates in this group was equal to the proportion of such candidates in the applicant pool. After going through this step, candidates in the higher scoring group are selected <u>*at random*</u> from the band. If the first band is exhausted before enough total candidates have been hired, then the same process is continued in the next band, beginning with candidates in the lower scoring group, as I have just described.

Now we come to the second major method of banding, <u>sliding bands</u>. As with fixed bands, there are the same types of variations, as follows.

[SHOW SLIDE #6]

First, there is the **Sliding Bands using Random Selection** method. In this method, all candidates are <u>*randomly*</u> selected from the first band until the top scoring candidates in the band have been exhausted, i.e., hired. Then, a new band is formed based on the candidates who remain. In other words, the band <u>*slides*</u> down to include those candidates who were not selected the first time. This process is repeated, and the band continues to <u>*slide*</u> downward. In effect, when this happens, new bands are continually being established as the band continues to slide.

A variation of this approach is **Sliding Bands using Minority Preference with <u>Top-Down Selection</u>**. This is very similar to the process I described when talking about fixed bands using minority preference with top-down selection. The only difference is that, whenever the top scoring candidates of both the lower, and then the higher scoring groups are exhausted, as by being selected or hired out of the band, the band then <u>*slides*</u> down, and a new band is formed.

Another variation on this theme is **Sliding Bands using Minority Preference with <u>Random Selection</u>**. Again, this is basically the same principle as with fixed bands using minority preference with random selection. Minority candidates are selected first on a random basis within the first band, then majority candidates are selected from the same band on a random basis. Then, the band *slides* down to form a new band based on those candidates who remain in the applicant pool. And the process is repeated again with the next wave of hiring.

The key to the notion of sliding bands "is that scores [that are] not significantly lower than the highest remaining score should not be treated different. One motivating factor for [using] the sliding band approach is increased minority selection" (Sackett & Wilk, 1994).

In their November, 1994 article in *American Psychologist*, Paul Sackett and Steffanie Wilk say that "the crucial issue for sliding bands is the selection of the majority group members with the top score in the band. Until those members have been selected, the band cannot slide and make additional minority group members eligible for selection" (Sackett & Wilk, 1994).

ADVANTAGES OF BANDING

[SHOW SLIDE #7]

The biggest advantage for banding of test scores, as opposed to a strict top-down procedure for selecting job candidates, is the possibility to <u>reduce adverse impact</u> on minority candidates. Generally, though, banding will "produce substantial increases in minority group selection only when minority preference within bands is in place." This, in effect, produces the same outcome as formal score adjustments, but "the relationship [between] banding and score adjustment is not commonly recognized. The permissibility of banding under the Civil Rights Act of 1991 is an issue yet to be resolved" . . . but a 1992 court case from the 9th Circuit Court of Appeals held that banding "was ruled as lawful under the Civil Rights Act of 1991" (Sackett & Wilk, 1994).

Further, Murphy, Osten, & Myors (1995) go on to say that "if test scores within a band are essentially equivalent, *any* system of selecting individuals from within the band might be defensible. If any system is appropriate, the criterion for developing a good system for within-band selection should be the extent to which it contributes to achieving key goals such as reducing adverse impact."

A second advantage, if you can call it an advantage, is the fact that "the <u>apparent inconsistency in banding is . . . common</u> in many measurement systems," such as in grading systems used by schools (Murphy, 1994). For example, a school teacher may use a **band** of 90 to 100 to represent an "A." Any score in this band is considered to be the same. Then, if the second **band** representing a "B" is 80 to 89, each of those scores are considered to be the same. Notice, however, that a score of 89 will be treated <u>differently</u> than a score of 90. A student scoring 89 will be assigned the grade of "B" whereas a student scoring 90 will be assigned the grade of "A." The same will be true for scores on the upper and lower limits of each succeeding band for grades of "C", "D", etc. Hence, the inconsistency. There are many measurement and decision-making systems that share this same type of basic inconsistency that banding can be criticized for (Murphy & Myors, 1994).

A third advantage of banding is that it has "an <u>intuitive appeal</u>, at least in the legal and political arenas" (Murphy, 1994).

<u>DISADVANTAGES OF BANDING</u>

There is an argument afoot by Dr. Frank Schmidt of the University of Iowa that all banding procedures for personnel selection are logically flawed. In his view, and I quote, "all banding models are simply attempts to reduce adverse impact while minimizing utility losses. This is their only rationale and justification; they have no legitimate statistical or measurement justification" (Schmidt, 1995). Sounds like a pretty powerful statement, doesn't it? Well, let's go on a bit further to see what he is talking about and why he is so strongly opposed to the use of banding of test scores for personnel selection.

Dr. Schmidt begins with the following explanation of the linear model: Scores on "valid employment tests bear a linear relationship to job performance. This implies that over a large number of job applicants, higher test scores will lead to higher job performance, and this will be true even if the test score difference is small and statistically nonsignificant." . . . Further, "it logically follows that, in the absence of other valid information about applicants, the optimal hiring decision policy is always to prefer higher scoring applicants to lower scoring applicants. This means that ranking with top-down selection is the optimal selection strategy for maximizing selection utility." (Schmidt, 1995).

Dr. Schmidt mentions a study by Cascio et al. (1995) that did not reject this argument, but nevertheless promoted the idea of banding test scores together that are less than two standard errors of measurement apart, or more specifically, two standard errors of the difference between the scores. The position of Cascio et al. (1995) "appears to be based on the belief that if two scores are not statistically different, then the best estimate is that they are equal, or at least that they should be treated as equal. [According to Frank Schmidt], this belief is incorrect," (Schmidt, 1995).

Dr. Schmidt gives the following hypothetical scenario of a candidate whose score is below the bottom of the band that was used in a hiring decision. The candidate's argument could go something like this: **[SHOW SLIDE #8]**

> There is nothing sacred about the top score that gives it a special status. After all, people near the center of the band and even people near the bottom of the band were hired. Their scores differed from mine by less than 2 Standard Errors of Difference (SEDs) and, therefore, are not meaningfully different from mine, under the basic banding criterion of what constitutes a meaningful

score difference.

[SHOW SLIDE #9]

> You cannot adopt this criterion and then by fiat refuse to apply it to any score differences except those between the top score and other scores. If it is a valid decision rule, then it can be applied to any score differences. Therefore, there is no reason why my score should not be in the band. In fact, it has to be in the band. Also, there can be no valid reason (other than chance) why they should have been hired and I was not (Schmidt, 1995).

Basically, Frank Schmidt's bottom line is that banding cannot be logically defended and may be difficult to defend legally in court or in administrative challenges (Schmidt, 1995). I'm not here to judge that statement one way or the other, just to give you both sides of the issue and to explain what banding is and how it works. Most of us here at this conference are public sector employees from cities, counties, states, the federal government (or heaven help us -- from more than one of the above at the same time!). And many of us IPMAAC people work in civil service departments or in human resource departments of public agencies. So, I'm sure that many of you sitting in this room today use some form of banding procedure in your jurisdictions for hiring of job candidates.

With that profound and astute observation, I'll wrap up my presentation with a couple of examples of how banding is currently being used by civil service departments in the state of Michigan (where I am from) and in the state of New York.

PRACTICAL EXAMPLES OF HOW BANDING IS BEING USED

Michigan

Michigan uses what they call a broad band certification process for hiring into state government jobs. The Department of Civil Service in Lansing establishes and maintains employment lists of eligibles. The lists may be divided by geographic area, organizational unit, occupational specialty, type of appointment (i.e., Full-time, Limited-Term, etc.), or other criteria. Eligible candidates are placed on these employment lists in one of three bands, corresponding to the following examination scores:

[SHOW SLIDE #10]

First Band:	Scores of 100 - 90
Second Band:	Scores of 89 - 80
Third Band:	Scores of 79 - 70

All hires and promotional appointments must be made from the top available band, appearing on a certified employment list. A Certified Employment List is a listing of eligibles for vacancies in a specific classification, employment type, and location. In the absence of names of laid-off employees from recall lists, all names contained in the first band of the employment list are referred for each request made by a state agency. If there are fewer than 15 names available in the first band, all names from the next lower band will be included in the referral to the employing agency. Names from successively lower bands will be included in the referral until the total names from the combined bands reaches a minimum of 15 names, if 15 names are available (Charles Blockett, Jr., Director, Bureau of Selection, Michigan Department of Civil Service, personal communication, May 5, 1997). [Mr. Blockett is a member of IPMAAC.]

New York

[SHOW SLIDE #11]

The state of New York uses a fixed band method where scores are grouped by using

Standard Error of Measurement formulas. Typically, their bands represent five-point clusters of test scores. After a passing score is determined, the raw scores of passing candidates are converted to a 70 to 100 point scale. Bonus points are added for seniority and veteran status. State agencies in New York use the "Rule of Three," meaning that agencies must make their personnel choices from one of the top three eligible candidates in the band who are willing to accept the position. Because of their use of this Rule of Three, New York does not use sliding bands.

All eligible candidates within a given band are equally reachable for appointment. The principle advantage of band scoring, in New York's eyes, is that it allows the agencies greater flexibility in selecting eligible candidates for appointment to state jobs. Further, it can serve to increase the validity of the selection system. For example, an agency may ask candidates about qualifications they have that are above the minimums required. In such case, band scoring may allow the agency to select individuals who have completed specialized coursework or who have specific experiences that would contribute to job performance. (Paul Kaiser, Principal Personnel Examiner, New York State Department of Civil Service, personal communication, May 5, 1997). Mr. Kaiser is a member of IPMAAC and is here this week at this conference.

So, now that I've told you all about test score banding, I'll turn the podium over to our next presenters who will talk to us about various cutoff score methods.

REFERENCES

Cascio, W. F., Outtz, J., Zedeck, and S, Goldstein, I. L. (1995). Statistical implications of six methods of test score use in personnel selection. Human Performance. 8 (3), 133- 164.

Murphy, K. R. (1994). Potential effects of banding as a function of test reliability. Personnel Psychology. 47. 477-494.

Murphy, K. R., and Myors, B. (1994). Evaluating the logical critique of banding. Unpublished manuscript, Colorado State University, Fort Collins.

Murphy, K. R., Osten, K., and Myors, B. (1995). Modeling the effects of banding in personnel selection. Personnel Psychology. 48 61-83.

Sackett, P. R. and Roth, L. (1991). A monte carlo examination of banding and rank order methods of test score use in personnel selection. Human Performance. 4 (4), 279-295.

Sackett, P. R., and Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. American Psychologist. 49 (11), 929-954.

Schmidt, F. L. (1995). Why all banding procedures in personnel selection are logically flawed. Human Performance. 8 (3), 165-177.