Otology & Neurotology
27:110–116 © 2005, Otology & Neurotology, Inc.

Using Audiometric Thresholds and Word Recognition in a Treatment Study

Chris Halpin and Steven D. Rauch

Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts, U.S.A.

Objectives: First, to examine a possible limit on significant results imposed by a progressive floor effect for hearing threshold improvement in a treatment study. This floor effect for hearing recovery suggests that if inclusion criteria are not set sufficiently high, the superiority of a treatment group may not be detectable. Second, to examine the outcomes when using two different types of criteria for significant change in a subject's word recognition score.
Methods: Several single-number criteria (e.g., 15 percentage points) are compared with the 95% (p = 0.05) criteria from the binomial critical difference table for monosyllables. Critical differences for binomial variables change depending on whether the starting value lies in the middle (near 50% correct) or at either extreme of the range of scores (0 or 100%). Different judgments of significant word recognition improvement (or decrease) using binomial versus single-value criteria are presented.
Data Source: A recent treatment study of sudden sensorineural hearing loss (n = 318) is used to illustrate these effects.
Conclusion: First, there is a progressive floor effect of presenting severity that covaries with the outcome measure hearing threshold recovery. In some designs, this may act to constrain the ability to detect a significant difference. Second, in the example data set, the use of single-value criteria for significant within-subject change in word recognition (e.g., 15 percentage points) introduced a miscategorization error rate of approximately 9% when compared with the result of the binomial 95% critical difference table. **Key Words:** Audiometry—Pure-tone average—Hearing thresholds—Word recognition.
Otol Neurotol **27:**110–116, 2006.

This article explores two specific issues that were seen during the analysis of audiometric data in a treatment study. The study compared the presenting versus the 3-month follow-up outcomes of patients with sudden sensorineural hearing loss (SSNHL) (1). The main variables were the recovery of audiometric thresholds and word recognition scores. The analysis of hearing threshold improvement brought forward consideration of a progressive floor effect that constrains this variable on the basis of the initial presenting severity of the patient. Analysis of a different variable, improvement in word recognition score, allowed the quantification (at least for this study) of the differences seen using Thornton and Raffin's (2) binomial critical difference table versus fixed criteria (i.e., a 15-point change).

Address correspondence and reprint requests to Chris Halpin, Ph.D., Department of Audiology, Massachusetts Eye and Ear Infirmary, 243 Charles Street, Boston, MA 02114, U.S.A.; E-mail: cfhalpin@meei.harvard.edu

Supported by National Institutes of Health/National Institute on Deafness and Other Communication Disorders grant UO1 DC062960 1A1.

A FLOOR EFFECT INFLUENCING AUDIOMETRIC INCLUSION CRITERIA

Pure-tone thresholds and pure-tone averages (PTAs) are subject to their own characteristic behaviors and limitations. When pure-tone thresholds or PTAs are used to evaluate recovery with treatment, it is important to recognize the "one-way" behavior of these variables when recovery occurs. The expected, normal values (near 0 dB hearing level [HL]) do not reflect the center of the audiometric range, but the most sensitive (healthy) extreme. Disease effects are characterized by departure from these values in only one direction, and the best possible effect of any treatment is to return the elevated thresholds to their normal levels. Because recovery goes in the direction of a fixed (healthy) value, the positive effect of treatment is limited by the severity of the presenting hearing loss. Put simply, it is not possible to show an improvement substantially greater than the initial loss. Conversely, more severe cases have progressively more relaxed limits on recovery magnitude. For example, a patient entering a study with a PTA of 60 dB could possibly recover "twice as much" as one entering the same study with a PTA of 30 dB. In statistical terms, there is a progressive floor

effect (threshold recovery limit), covarying with the outcome measure (improvement in threshold or PTA).

When evaluating a treatment using thresholds or PTA, a probability is set for a judgment that the treated group outcome is better than the untreated group. When, for example, the effect is tested using a *t* test for significance at the $p < 0.05$ level, this means that the probability that the mean improvement for the treated and untreated groups are the same must be less than 5%. One way of expressing this would be to use the subjects' scores to form a range, called the confidence interval of the mean, in which one can be 95% confident that the true mean must fall. If this were done for the untreated and the treatment groups, and if the two ranges (confidence intervals) did not intersect, the means could be said to be different at the $p < 0.05$ level.

To relate this approach to hearing improvement, if the untreated group in a study had a mean spontaneous recovery of 15 dB with a 95% confidence interval from 0 to 30 dB, the patient entering the treatment arm with a 60-dB PTA could recover by more than 30 dB (or not), which would allow enough freedom of improvement to compare the treatment outcome to the untreated group. In contrast, a patient entering the treatment arm with an initial PTA of 30 dB can only recover about ''half as much'' in PTA terms, and the best outcome expected (30-dB improvement) is not sufficient to clearly separate these results from the upper limit of the confidence interval of the untreated group (also 30 dB in this example). The inclusion of subjects who are not free to vary beyond the upper limit of the 95% confidence interval of the comparison group introduces a drag on the ability of the treatment group as a whole to significantly exceed the performance of the untreated group. This drag is generated not by lack of treatment effectiveness but by the limits imposed by the floor effect for audiometric recovery.

An Example Study

Figure 1 shows the outcome of a retrospective study of oral steroid treatment of patients with SSNHL (1). The initial severity versus the outcome (both expressed in decibels PTA) are plotted using filled symbols for 266 cases receiving steroids and open symbols showing 52 untreated cases. All cases were monaurally affected, and each affected ear is plotted such that the starting (or pretreatment) PTA is shown by the location along the horizontal axis, and the amount of eventual recovery is shown by the location relative to the vertical axis. On the vertical axis, the presenting or pretreatment PTA (in this case, the average of 0.5, 1, 2, and 4 kHz) is subtracted from the final PTA such that a large recovery will result in a large negative number. A triangular shaded area is shown, illustrating that none of the recoveries, either spontaneous or with treatment, would be expected to be larger than the severity of the initial presentation. Although it is true that audiometric thresholds can improve beyond expectation strictly because of their inherent variability (i.e., to –5 dB HL), it is rare for this to occur using PTA, and it did not occur in this study.

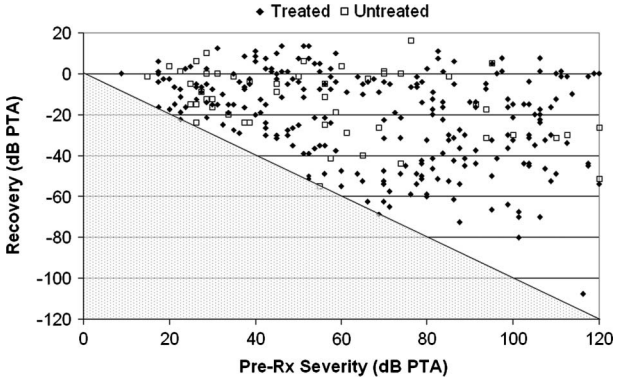

FIG. 1. Hearing threshold outcomes in a retrospective study of SSNHL (n = 318). The *horizontal axis* shows the initial severity of loss in decibels of PTA (0.5, 1, 2, and 4 kHz). The *vertical axis* shows recovery by subtracting the initial PTA from the final PTA (after 3 mo). As a result, cases with larger improvements have larger negative values on the *vertical axis*. The *shaded triangle* shows the progressively narrow area where results are not expected to be found because the final thresholds would have to recover to better than 0 dB HL.

Untreated Cases Versus an Absolute Recovery Criterion

Figure 2 shows the same data and axes as Figure 1. In Figure 2, the darker shaded region (''untreated'') shows the mean (13 dB) recovery and 95% confidence interval (3–23 dB) of the untreated cases in terms of the vertical (recovery) axis. Because this region describes the performance of the untreated cases, it is reasonable to argue that any treated outcomes lying within such an area do

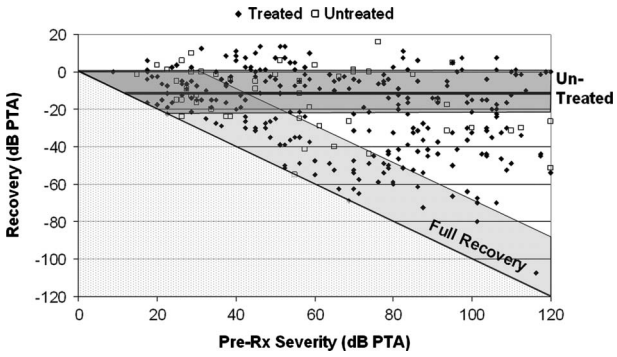

FIG. 2. Untreated cases versus an absolute recovery criterion. Axes and data are the same as in Figure 1. Here, the *dark horizontal line* shows the mean and the *darkly shaded area* shows the 95% confidence interval of the mean of the untreated group. The *lightly shaded diagonal* (marked Full Recovery) shows the behavior of a normal-referenced criterion (here, PTA ≤ 30 dB HL) versus the progressive floor effect for recovery. The intersection of the regions below approximately 40 dB HL starting PTA indicates that recovering cases cannot be separated from the untreated group mean.

not support a significant treatment effect. A lightly shaded diagonal region (marked ''full recovery'') shows the application of one possible recovery criterion: that of return of PTA to less than or equal to 30 dB HL. This criterion was not used in the original study, but absolute criteria versus normal (0 dB HL) results such as this can be considered when designing a treatment study. The effect of the progressive floor effect on such an outcome criterion is included here as an example. The achievement of the 30-dB HL absolute recovery criterion requires the greatest threshold improvement in the severe cases and progressively less improvement for cases starting with smaller PTAs. Indeed, for cases whose initial loss is less than 40 dB PTA, the full recovery and the untreated areas overlap. In these less severe cases, then, the criterion for both full recovery and for no effect of treatment are met simultaneously. This should not be construed as invalidating the use of criteria of this type, only that it may be useful to consider the possible implications of the floor effect for hearing threshold recovery.

PTA Inclusion Criterion

Figure 3 Shows the outcome in the same study when the principle of threshold recovery limit was applied to the setting of the inclusion criterion. As in Figure 2, the mean and confidence interval are shown for the untreated group. An additional shaded area below shows the mean (28 dB PTA) and 95% confidence interval (24–36 dB) for the improvement in the treated patients. This interval is narrower than that of the untreated group because of the inclusion of many more cases (n = 266). In this analysis, the lower limit of the 95% confidence interval for treated patients does not intersect the upper limit of the confidence interval for untreated patients, resulting in a judgment of a significant difference between the two groups. For this to happen, the treatment group members must be free to vary not just enough to achieve the group mean value (28 dB HL) but also to allow outcomes to be greater than that mean such that the average can separate itself from the untreated group. In Figure 3, it can be seen that an inclusion criterion of 40 dB PTA of initial severity should allow outcomes on both sides of the resulting treatment group mean.

The graphic approach in Figure 3 is an example for consideration and does not mean that there is only one defensible place to locate the inclusion criterion. Other methods (i.e., increasing the number of subjects) also act in the direction of addressing this limitation. However, on a practical basis, increasing the number of subjects is expensive and may not be the best way to solve the purely statistical ''drag'' introduced by a low inclusion criterion. Picking the inclusion criterion in this manner does not imply that data are discarded until the desired result is found. If there is no difference, one is not expected to appear by changing the inclusion criterion (although redoing any statistical test many times is ill-advised). Selecting an inclusion criterion as in Figure 3 will only act to allow the possibility that an effect can be seen. In contrast, including all cases without addressing the recovery floor effect may result in failing to detect an effect that is actually there.

Showing Further Effects of an Alternate Treatment

Figure 4 Shows a hypothetical further study, testing whether an alternative treatment is superior to the treatment applied in Figure 3. The mean and 95% confidence interval of the real treatment data from the study (Rx1, same as Treated in Fig. 3) are again shown. A similar sized, hypothetical mean and confidence limit of an alternative treatment (Rx2) is plotted to illustrate a group recovery outcome that would show a significant improvement using an alternative treatment. Figure 4 suggests that both the

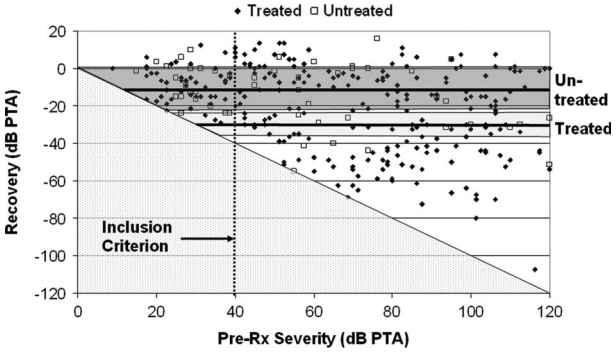

FIG. 3. Group comparison and the inclusion criterion. The data and axes are the same as the previous figures. In addition, the mean and 95% confidence interval of the untreated group remains. The *lightly shaded box* (Treated) shows the mean and 95% confidence interval for the treatment group. The *vertical dotted line* (Inclusion Criteria) shows an example of one such criterion (40 dB PTA) that would allow outcomes both above and below a significantly different treatment mean and its 95% confidence interval.

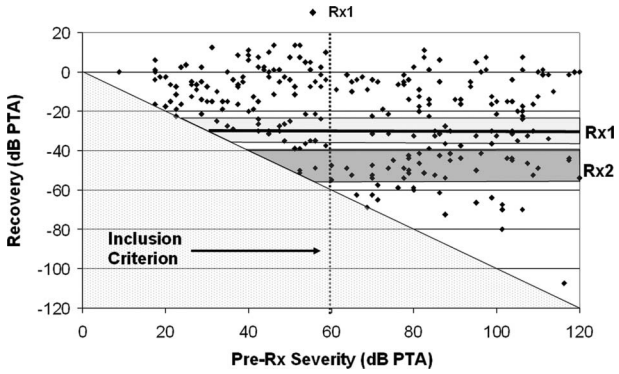

FIG. 4. Comparing two treatments. In this figure, the mean and 95% confidence interval of the original treatment group remain from Figure 3 (*lightly shaded box* marked Rx1). A box is added (*darkly shaded*, marked Rx2) estimating the possible performance of a treatment group whose hearing recovery effects can be shown to exceed those of Rx1 at the $p < 0.05$ level. The intent of this figure is to show that a better outcome may require cases that are more free to recover audiometrically, and so only cases above a higher inclusion criterion (here 60 dB PTA) may be able to show significance when analyzed in this way.

treatment groups be composed of patients who are able to improve by a larger amount than that required to differentiate the treatment group from the untreated cases as in Figure 3. Here, the inclusion criterion is moved up to 60 dB PTA to conservatively ensure that cases that cannot show this type of improvement do not act to obscure the possible superiority of the second treatment. This is a substantial hearing loss, and many legitimate cases of both disease and recovery will be excluded by this approach. This article is not intended to promote this or any method as optimal for analysis but rather to recognize the possible effects of the hearing recovery floor effect on studies with these particular design features.

COMPARISON OF FIXED VERSUS BINOMIAL WORD RECOGNITION CRITERIA

When two word recognition tests are compared, as in a within-subject change with treatment, both clinicians and clinical researchers may wish to know whether the scores are different at a set level of probability. In general, the expected distribution of a variable governs the increasing probability that two scores are different as those scores grow farther apart. This results in two scores, one above and one below the starting value, which are known as the critical differences at the selected probability level (i.e., $p = 0.05$). This article compares the results when two approaches to setting the critical difference are applied to the SSNHL treatment study (1) data where within-subject improvement in word recognition was evaluated in 281 cases.

Clinical Word Recognition

Clinical word recognition, for purposes of this article, corresponds to the recommendation for standard methods originally described by Egan in 1948 (3). The patient is presented a list of monosyllables, preceded by a carrier phrase (“You will say…camp/say/king…”) and the result is an accumulation of correct and incorrect responses, expressed as percentage correct. This result is interpreted with reference to the level and audibility in that particular condition, with the most common paradigm being a single, high level where the maximum score is expected (4). The nature of the variable is an accumulation of binary responses (correct/incorrect). Statisticians have published detailed treatments of the expected behavior of such scores, which they refer to as binomial variables (5).

The Binomial Model of Variance

The distribution of binomial variables differs from the normal (bell curve) model such that the distribution is wider in the center of the range and narrower at the extremes. For word recognition tests, the 95% critical difference near 50% is expected to be larger than that near 0% or 100% (6). This has practical significance because different word recognition ranges are expected in different studies in otology. For example, studies of cochlear implant patients will often show scores clustered at the low extreme of the range (0–20%), Studies of advanced Ménière’s disease may show many scores near the center (30–60%) and studies of conductive hearing loss, including surgical outcomes, may cluster at the high extreme (near 100% correct). The effect of the binomial distribution of word recognition scores is that quite different criteria for critical differences are expected in these studies, based strictly on mathematics, before the specific behavior of the patient or the disease is even considered.

Using Single-Value Criteria

In many studies reporting a word recognition change within subjects, a criterion value (percentage point change) is set as the criterion for a critical difference. These values tend to vary from 10 to 15 percentage points. This level of change is well-supported by observation, and this article does not suggest that these values are incorrect. These values do, in fact, form the center of a *range* of values that describe the statistical critical differences between word recognition tests when the binomial model of variance is used. In contrast, the fact that the binomial criteria change (in raw score terms) to maintain the same probability raises the possibility that a fixed criterion could allow certain errors in detection of differences. Given a fixed criterion value, a fixed difference from an initial score of 50% could be found that did not, in fact, exceed the 95% probability requirement (false alarm) or that a difference from an initial score near 0%, although less than the fixed criterion, would nonetheless exceed the 95% critical difference (miss). Although these errors can be shown as theoretical possibilities, it does not prove that the magnitude of such differences is great enough to merit application of the binomial method. In this section, we illustrate this issue using data from the SSNHL treatment study (1), which included within-subject improvement in word recognition as a variable (n = 281) and report the differences using fixed criteria versus the binomial critical difference table.

The Binomial Critical Difference Table

The binomial model of variance for tests of this type indicates that the best method for determining a critical difference is not to apply any single value but to apply a changing set of values indicated by the binomial variance model. In practical terms, this can be done by referring to a table published for this purpose by Thornton and Raffin (2) (Table 1). For example, if the pretreatment word recognition score (for 50 standard CID-W22 monosyllables) is 52% and the posttreatment score is 68%, a criterion of a 15-point change would result in a judgment of a significant change. However, because the beginning score is in the center of the range, the expected variance is large and this subject would be required to exceed 70% (18-point difference) for a $p < 0.05$ level of confidence. This would then be an error of specificity (false alarm), because the 15-point criterion would produce a judgment of significance when none was actually

TABLE 1. *Upper and lower limits for the 95% critical differences for percentage scores, adapted from Thornton and Raffin (2)*

Score (%)	n = 50	n = 25	n = 10
0	0–4	0–8	0–20
2	0–10		
4	0–14	0–20	
6	2–18		
8	2–22	0–28	
10	2–24		0–50
12	4–26	4–32	
14	4–30		
16	6–32	4–40	
18	6–34		
20	8–36	4–44	0–60
22	8–40		
24	10–42	8–48	
26	12–44		
28	14–46	8–52	
30	14–48		10–70
32	16–50	12–56	
34	18–52		
36	20–54	16–60	
38	22–56		
40	22–58	16–64	10–80
42	24–60		
44	26–62	20–68	
46	28–64		
48	30–66	24–72	
50	32–68		10–90
52	34–70	28–76	
54	36–72		
56	38–74	32–80	
58	40–76		
60	42–78	36–84	20–90
62	44–78		
64	46–80	40–84	
66	48–82		
68	50–84	44–88	
70	52–86		30–90
72	54–86	48–92	
74	56–88		
76	58–90	52–92	
78	60–92		
80	64–92	56–96	40–100
82	66–94		
84	68–94	60–96	
86	70–96		
88	74–96	68–96	
90	76–98		50–100
92	78–98	72–100	
94	82–98		
96	86–100	80–100	
98	90–100		
100	96–100	92–100	80–100

n = the number of words presented during the test.

present. Conversely, a subject in the same study with a presurgical score of 94% and a postsurgical score of 80% would have had a drop in speech intelligibility exceeding the 95% binomial critical difference, whereas the 15-point criteria would not have been exceeded. This would be an error of sensitivity (miss). In the end, there is no way to remove these sorts of errors by changing the criterion to any single number. As is seen, the effect of raising or lowering any fixed criterion is to change the balance of errors between sensitivity and specificity without reducing the total amount of error. It is also important to note that the solution does not lie in abandoning the standard monosyllable tests (CID W22; NU#6). The variance of monosyllable tests is small and well understood as compared with other tests with more complex structures.

The Effect of Word List Length

Before examining more closely the outcomes of fixed criteria versus the binomial table, it is important to briefly examine a separate source of significant variability in word recognition scores. The variance of word recognition scores is very sensitive to the number of words presented to the subject. In essence, the expected outcome of a test using 25 words (a half-list) versus the original 50-word set is the same score; however, the variance using the half-list is much greater. This is not always reflected in the literature regarding the use of half-lists. The fact that both full and half-lists tended to cluster around the same score (which is true) was used to suggest that half-lists could be used for efficiency (7). In contrast, an important judgment for both the clinical and research use of these scores hinges on whether time or treatment has resulted in a critical difference. For this purpose, a full list of 50 words is better suited to the task.

Figure 5 shows the impact (from the binomial table) of using monosyllable lists of different lengths. A series of upper and lower 95% critical differences are shown, each around the same score of 50% correct. The horizontal axis is the number of monosyllables presented, showing the effect of list length on the width of the critical difference at the $p = 0.05$ level. Even the full 50-item list results in a sizable width between critical differences (32–68%) in the midrange, but even more striking is the finding that monosyllable lists of less than 50 items

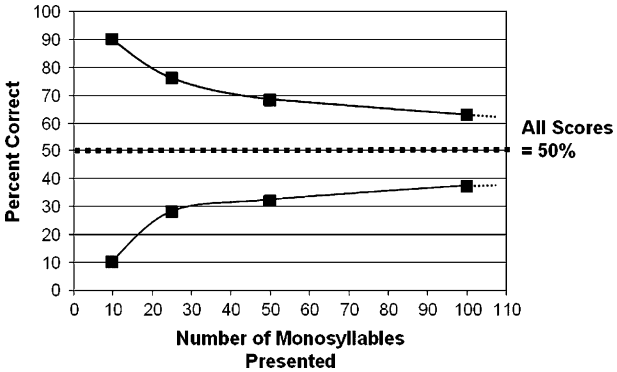

FIG. 5. Effect of word list length. A series of upper and lower 95% critical differences are plotted for tests in which scores are 50% correct (52% for the 25-word list). The *horizontal axis* is the number of monosyllables presented, showing the effect of word list length on the width of the critical difference between two word recognition scores.

produce an inherent variability so large that they may not be useful for most within-subject comparisons.

The Effect of Using a Fixed Difference as the Criterion

The question remains as to the actual amount or importance of errors introduced by using fixed rather than binomial difference table criteria when comparing two scores. Figure 6 shows errors of sensitivity (missed differences) and specificity (false alarms or 1 – specificity) for criteria from 10 to 18 points when these criteria were applied to the SSNHL (1) data set (n = 281). This graph is constructed to show a receiver operating characteristic as if the different single-value criteria were detectors, and the signal was significance as defined by the binomial critical difference table of Thornton and Raffin (2). The data set was analyzed once for each criterion value, and both error types were plotted in the standard sensitivity by 1 – specificity (receiver operating characteristic) space. The widest fixed criterion (18 points) will have no false alarms but is a wide enough criterion to miss approximately 9% of actual significant differences. This error rate is not solved by using a lower value. Although fewer misses are indeed seen, the occurrence of false alarms rises with lower values. The criterion of 15 points produced 9 false alarms (3.2% error of specificity) and 17 missed significant differences (6.0% error of sensitivity), for a total error rate of incorrectly categorized results of 9.3%. The total error rate in this example is roughly constant at approximately 9% and corresponds to the expected, curvilinear model of imperfect signal detection.

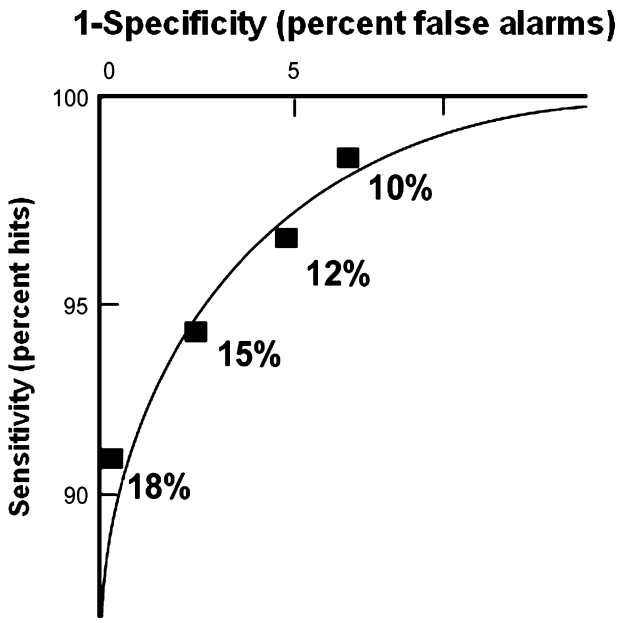

FIG. 6. The effect of using fixed criteria. The data set of a study using within-subject improvement in word recognition scores (n = 281) was analyzed four times using fixed critical differences of 18-, 15-, 12-, and 10-points change. Hits and false alarms are plotted using the binomial difference table as the standard. A fairly constant 9% rate of miscategorization was found in this study, with the balance of sensitivity versus specificity shifting with the size of the fixed criterion.

The results shown in Figure 6 are specific to the study from which they are drawn. As discussed above, the width of the 95% critical difference is expected to change with starting values near the midrange as opposed to those at the extremes. The starting values in this study covered the entire range from 0% correct to 100% correct, and the scores were not tightly clustered around any one value (mean, 41%; standard deviation, 39%; median, 31%;). The full range of starting values allows errors in both directions, and the number of initial word scores in the middle of the range indicates that this could be considered a conservative estimate of the resulting errors. For example, outcomes in studies of conductive losses may cluster more tightly near 100%, and this might result in a further reduction of sensitivity, for example, using a criterion of a 15-point change. If no substantial error rate was seen in the analysis in Figure 6, it could be argued that the mathematical argument regarding treating word recognition scores as a binomial variable makes little difference. The 9% miscategorization rate actually found allows investigators one example that can be used to quantify the possible error rates that may appear as they use within-subject comparisons of word recognition scores.

One example of using the table conservatively would be to analyze each pair of word recognition scores (i.e., before versus after) using the 95% critical difference table and report that 60% of subjects improved by more than the 95% critical difference, 10% got worse, and 30% did not exceed the critical difference. This is a different approach than grouping the pre- and posttreatment data and using the resulting standard deviation to evaluate whether the means of the groups are different. Grouping the raw scores assumes normal (not binomial) variance and may well be sufficiently robust to withstand the violation of that assumption if large numbers are used and the groups are carefully balanced. A compromise position has been proposed (8) in which the binomial variable is converted to one with a normal distribution using the ArcSin function, and then the groups are combined. Figure 6 shows that errors are introduced in both directions, and so use of the table is *not* less likely to show an effect. In fact, given the expected starting range of certain studies near the extremes (such as low scores with cochlear implants or high scores in conductive losses), the narrower binomial critical difference criteria will be more likely to show effects missed by the wider single criterion values.

DISCUSSION

This article illustrates two effects seen during the analysis of audiometric data in a treatment study. There is no intent to cover the range of statistical methods that may be considered when designing such studies. The

object is to raise and explore the recovery floor effect for hearing thresholds and the binomial critical differences, without suggesting that this is a complete treatment of either one. Further exploration of these issues using different data sets would likely be useful. The floor effect particularly is an issue specific to recovery of hearing thresholds, and there are many studies where recovery is not expected. The word recognition criteria effects relate to the comparison of two scores (e.g., before versus after treatment) and primarily apply to within-subject differences rather than to the means of groups. Nonetheless, it is hoped that this brief discussion will allow these issues to be recognized and evaluated as to their possible effects in studies using audiometric outcome measures.

Acknowledgments: The authors thank Aaron Thornton, Ph.D., and the American Speech-Language-Hearing Association for permission to make use of Table 1.

REFERENCES

1. Chen C, Halpin C, Rauch S. Oral steroid treatment for sudden sensorineural hearing loss: a ten year retrospective analysis. *Otol Neurotol* 2003;24:728–33.
2. Thornton A, Raffin M. Speech discrimination scores modeled as a binomial variable. *J Speech Hear Res* 1978;21:507–18.
3. Egan J. Articulation testing methods. *Laryngoscope* 1948;58:955–91.
4. Hirsh I, Davis H, Silverman E, Reynolds E, Eldert E, Benson R. Development of materials for speech audiometry. In Chaiklin, J Ventry I, Dixon R, eds. *Hearing Measurement: A Book of Readings*. New York: Addison-Wesley, 1952:183–96.
5. Freeman M, Tukey J. Transformations related to the angular and the square root. *Ann Math Statist* 1950;21:607–11.
6. Hagerman B. Reliability in the determination of speech intelligibility. *Scand Audiol* 1976;5:219–28.
7. Elpern B. The relative stability of half list and full list discrimination tests. *Laryngoscope* 1961;71:30–6.
8. Studebaker G, Gray G, Branch W. Prediction and statistical evaluation of speech recognition test scores. *J Am Acad Audiol* 1999;10:355–70.