0196-0202/82/0302-0051$02.00/0
EAR AND HEARING

Vol. 3, No. 2
Printed in U. S. A.

A Quantitative Examination of the Sources of Speech Discrimination Test Score Variability

Harvey Dillon

National Acoustic Laboratories, Millers Point, Sydney, Australia

ABSTRACT

Several sources of variability inherent in any speech discrimination measurement are outlined. Limitations of the use of the binomial theorem to predict intrasubject (test-retest) variability are examined. First, the underestimation of variability that is caused by the inclusion of items of different degrees of difficulty is quantified by a reanalysis of published data. Second, it is shown that variability will be larger than expected if the subject's ability is different during test and retest sessions. Fortuitously, these two deviations from the model have opposing effects on the score variability. The estimates provided by the binomial theorem are thus better than if only one effect was present. The need for a clear distinction to be maintained between inter- and intrasubject variability and the effect of list differences on test-retest variability are also discussed.

Speech discrimination tests find many uses within audiology. In virtually all instances, it is necessary to compare the resulting test score with some other score. The comparison score may be another test score obtained from the client under different conditions, or may be an average value obtained by normal hearing persons, or may simply be the audiologist's estimate of what constitutes adequate discrimination ability. In every instance, however, it is necessary to know the accuracy (or uncertainty, or variability, or reliability, or precision) of each of the scores being compared.

Several authors have recently pointed out that the variability of speech discrimination tests can be estimated by using the properties of the binomial distribution. Among the predictions of the simple binomial model (upon which the confidence limit tables are based[19, 23]) are the statements that the variability depends only on the discrimination score obtained and the number of items in the test, and is independent of the particular subject and test (except to the extent that they determine the score). This paper examines some ways in which the assumptions involved in the simple binomial model may be violated, and attempts to quantify, where possible, the effect these violations have on the observed variability. It is shown that, once list equivalence problems are overcome, the simple binomial model can significantly underestimate or overestimate the variability applicable for a particular client and test. Clinical implications of this are considered. For the purposes of this paper, the variability associated with speech test scores will be represented by the size of the S.D. of a group of such scores. Although this is a convenient measure when assessing either the accuracy of any particular score or the importance of different sources of variability, it cannot be used directly to test whether two scores are significantly different from each other. This is because test scores obtained from a client under identical conditions do not form a normal distribution, but are grouped asymmetrically around the mean score. Furthermore, if the experimental conditions are altered, and as a result the mean score changes, then a different S.D. results. Thus, comparisons between scores can be made only after the scores have been mathematically transformed to a system where the S.D. is independent of the mean score and where reported scores are normally distributed around the mean. For clinical applications, tables have been prepared which indicate the critical difference required between two scores before they can be accepted as being significantly different.[19, 23] Naturally, the critical difference depends on the number of items in the test and the actual scores obtained.

This paper also discusses some erroneous conclusions that have been drawn when inter- and intrasubject variability estimates have been confused. An overview of the sources of variability is as follows.

FACTORS INVOLVED IN TEST VARIABILITY

List Differences

To avoid learning effects, speech discrimination tests usually consist of several "equivalent" lists. If the differ-

ent lists are not truly equivalent in difficulty (for all subjects measured on an individual basis, not just when averaged across subjects), then differences in performance between various test sessions will clearly contain a component due to list inequivalence. Campbell[3] has shown that these inequivalences can be minimized for the CID W-22 lists by rearranging the existing items into new lists. These new lists not only have the same overall discrimination score, but also have similar quartile points, indicating that the range of difficulty of the items in each list is also similar. To achieve the same ends, Hood and Poole[11] have taken a different approach for the Medical Research Council word lists. They showed that the articulation functions for the different lists can be brought closer together by appropriately redefining the reference intensity for each list, and deleting from the test battery five "rogue" lists whose articulation functions have slopes which are different from the rest. It is again emphasized that adjustments such as these make the lists equivalent only for the average subject, although the Hood and Poole[11] article shows that considerable improvement in the homogeneity of the test for individual subjects will also result when the group adjustment is made.

Problems caused by list differences can be largely overcome by the use of a closed response set format, because the same words can be repeated with only a small amount of learning occurring.[8, 14] The effect of list differences will not be discussed further in this article.

Statistical Fluctuations

If a single test presented many times to the one subject (with no learning or fatigue present) is considered, the same score will not be obtained at each presentation of the test. Such differences may be classified as statistical fluctuations. Under the assumption that all items in the test have an equal probability of attracting a correct response, P, the scores obtained from such a test will form a binomial distribution with the only parameters being P, and N, the number of items in the test. Lyregaard,[15] Hagerman,[10] and Thornton and Raffin[23] have all pointed out the applicability of the binomial distribution to speech discrimination scores. The latter two have also shown that the expected value of the S.D.,

$$\sqrt{\frac{P(1-P)}{N}},$$

was a good approximation to that obtained in the speech discrimination tests they considered. Practical speech tests, of course, do not consist of items which are equally easy to discriminate. Lyregaard[15] and Hagerman[10] have shown how the binomial theory may be extended so that the S.D. of tests containing items of different degrees of difficulty can also be predicted. It is worth repeating here that the statistical fluctuations in any speech discrimination test are entirely predictable once the difficulty and number of the test items are known. Even quite recent articles, e.g., Hughes et al.[12] and Penrod,[17] do not seem to have taken account of this fact. Indeed, were it not for the third source of variability, there would be no point in measuring the variability of scores obtained from a subject.

Subject Consistency

The above theory assumes that each presentation of the test to a subject attempts to measure a single, fixed attribute—the subject's discrimination ability. Many factors will affect this discrimination ability, and some of these may vary with time. For example, motivation, fatigue, amount of learning achieved in previous tests, and even the attitude of the experimenter or the amount of sleep the subject had the night before could all have an effect on the "true" ability of the subject at the time of any particular test. If the above factors (or any others) cause the subject's ability to change from test to test, then a variability greater than that expected on the basis of the binomial distribution would be measured. It is for this reason that estimates of the reliability of a particular test may have some value. As an example, a very long test may produce additional variability if an individual is more prone to fatigue on some days than on others, while a short test may not cause additional variability.

Subject Differences

When scores obtained from different individuals are compared, the spread of scores will contain both intra- and intersubject variability components. Thus, intersubject comparisons do not provide a valid indication of test-retest reliability, although they are sometimes used that way. The scatter of scores can be expected to be largest when it arises from subjects with a range of hearing losses and smallest from a more homogeneous group (such as normal-hearing persons). Data on intersubject variability are presented later in this paper and compared with intrasubject variability estimates for different types of populations.

QUANTIFYING VARIABILITY

There have been almost as many different methods used for quantifying test variability as there have been investigations. The procedures fall into two broad classes, however: those which lump intra- and intersubject variability together, and those which separate them. The following discussion will review some of the measures of test variability that have been made, and will attempt to quantify the relative size of the different sources of variability.

Intrasubject Variability

As has been mentioned earlier, the statistical fluctuation component of intrasubject variability may be calculated once the number of items and the degree of

difficulty of each item in the test is known. Use of the formula

$$\sigma_1 = \sqrt{\frac{P(1-P)}{N}} \quad (1)$$

predicted by the simple binomial distribution (all items of equal difficulty), will in general overestimate the true value when items of mixed difficulty are present in a test. The more exact theoretical formula

$$\sigma_2 = \sqrt{\frac{1}{N} \sum_{i=1}^{N} P_i(1-P_i)} \quad (2)$$

can be obtained from the properties of the subnormal binomial distribution.[10, 15] The values, P_i, correspond to the probability of obtaining a correct response for each item in the test.

Expected Variability for Common Tests Data on the degree of difficulty of individual items in a test have been presented by various authors for a number of tests. Equations 1 and 2 have been applied to each of these sets of data to calculate the theoretical S.D. under the assumption of both equal difficulty items, and mixed difficulty items, respectively. The results are shown in Table 1. Scanning down the right-hand column, one can see that the simple binomial formula overestimates the more exact theoretical value by a factor of between 1.05 and 1.43. Notice, however, that the ratios obtained from the Byrne and Walker[2] experiment are considerably higher than those obtained from the others. Their data are different from the others in that the difficulty of each item is calculated individually for each subject. If the data from each of their subjects are combined before calculating the expected variability, the ratio of the two estimates (σ_1 and σ_2) becomes 1.11, about the same as the average value obtained from the other experiments. The explanation for this is that words which are very easy (or difficult) for one subject are not necessarily so for another. Thus, averaging data across subjects leads to the impression that the words in a test are of more uniform difficulty than is actually the case for any individual subject. We may conclude that for the other speech tests in Table 1, the effects of mixed difficulty items, when measured for individual subjects, is to reduce the S.D. predicted on the basis of the simple binomial theory by a factor of approximately 1.4 also. A score of 70% on a 50-item test, for example, would have its test-retest S.D. estimate reduced from 6.5% to about 4.6%, a fairly large change.

Comparison of Measured and Expected Values To compare the theoretically determined estimates of variability with experimentally determined values, we require experiments where each subject has heard a given list two or more times, with the differences in score between test and retest computed individually for each subject. Suitable experiments have been performed by Hagerman[10] and Thornton and Raffin.[23]

A comparison of the theoretical and experimental S.D.s from the data of Thornton and Raffin[23] is given in Figure 1. For each subject whose result is given in Thornton and Raffin's Figure 1 (p. 513), the discrimination score was used to calculate the theoretical S.D. assuming items of equal difficulty (equation 1).

As can be seen from Figure 1, the simple binomial model provides a reasonable fit to the data. A perfect fit would be obtained when the points all fall on the solid line (with unity slope). Similar computations were applied to the data given by Hagerman[10] in his Figures 2 and 3 (pp. 223, 224), and the results are shown here in Figure 2. A reasonable fit is again obtained, although there is a tendency for the model to underestimate the variability for low scoring subjects. Recall, however, that

Table 1. Theoretical S.D.s, σ_1 and σ_2, based on formulas 1 and 2 (all variability estimates have been normalized to a 50-word list)

Speech Test	List or Subject	Data Source	Average Score (%)	σ_1 %	σ_2 %	Ratio, σ_1/σ_2
CID W-22	L.1	Campbell[3]	75	6.1	5.3	1.17
CID W-22	L.2	Campbell[3]	75	6.1	5.5	1.13
CID W-22	L.3	Campbell[3]	78	6.6	5.4	1.09
CID W-22	L.4	Campbell[3]	74	6.2	5.7	1.09
CID W-22	L.1	Thornton and Raffin[23]	86	5.0	4.3	1.15
CID W-22	L.2	Thornton and Raffin[23]	83	5.3	5.0	1.06
CID W-22	L.3	Thornton and Raffin[23]	82	5.4	5.1	1.05
CID W-22	L.4	Thornton and Raffin[23]	79	5.7	5.5	1.05
CID W-22	L.2	Penrod[17]	79	5.7	5.4	1.07
WIPI		Sanderson-Leepa and Rintelmann[20]	94	3.2	3.0	1.05
PB (Swedish)	L.12	Hagerman[10]	52	7.1	6.0	1.19
PB (Swedish)	L.12	Hagerman[10]	72	6.4	5.7	1.12
PB (Swedish)	L.12	Hagerman[10]	84	5.2	4.7	1.11
PB (Swedish)	L.12	Hagerman[10]	93	3.6	2.8	1.28
NST	S.1	Byrne and Walker[2]	71	6.4	4.5	1.43
NST	S.2	Byrne and Walker[2]	70	6.5	4.5	1.42
NST	S.3	Byrne and Walker[2]	64	6.8	5.1	1.32

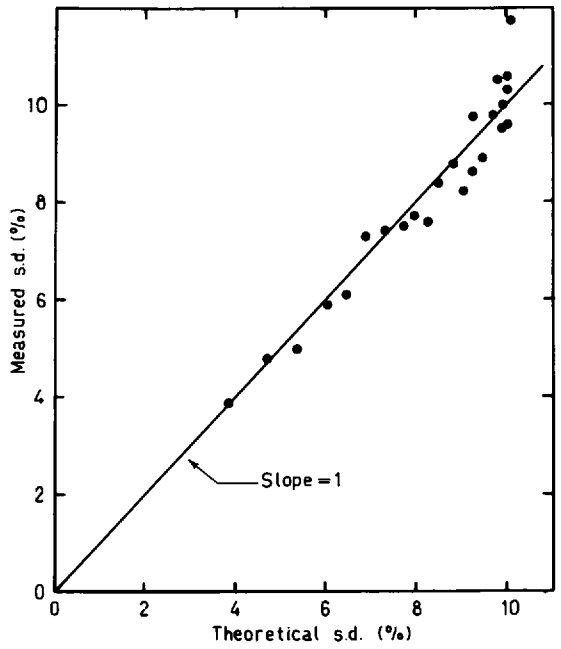

Figure 1. Measured test-retest S.D. versus predicted S.D. (equation 1). Data are from Thornton & Raffin.[23]

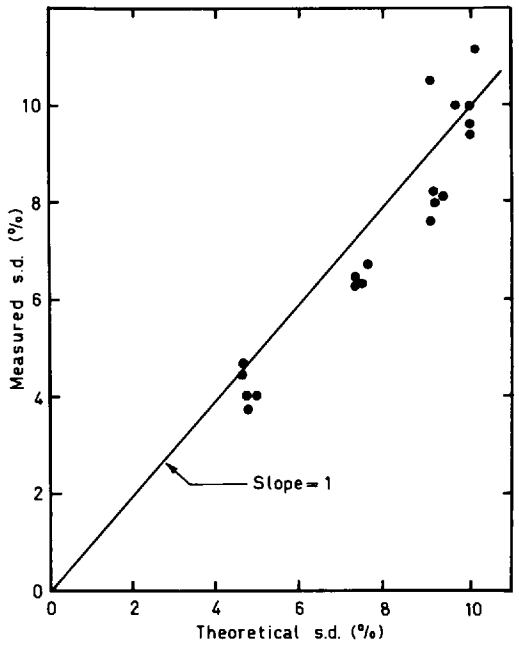

Figure 2. The same as for Figure 1, but data are from Hagerman.[10]

the simple binomial model was expected to overestimate the variability due to the mixture of item difficulties in these tests. Thus, the goodness of fit of the simple binomial model must be due in part to there being variability in addition to that expected on binomial theory grounds.

One limitation of these two studies is that the test and retest consisted of the two halves (or five fifths) of a single 50-word speech test. Thus, it was only required that the subject maintain a consistent performance level over a time of a few minutes. For clinical purposes, however, it is often necessary to compare scores obtained on different days. Data obtained by McConnell et al.[16] indicate no additional variability when test and retest are separated in time. However, their variability estimates were so large (S.D.s of over 9% for a 50-word list) that some other factor clearly dominated the total variability. The additional factor was possibly the live-voice testing procedure used by them.

To check the validity of the binomial theory for scores obtained over a time interval of more than a few minutes, some data obtained for other purposes in this laboratory (National Acoustic Laboratories) were reanalyzed. The original experiment[2] involved the measurement of the speech discrimination of three subjects under six different experimental conditions. For each of these conditions, a 55-item nonsense syllable test[14] was presented 10 times over a period of three weeks. For the present analysis, the scores from the first three tests were discarded because a learning effect was present (despite the closed response set format). The mean and S.D. of the remaining seven tests were calculated. Because for each subject the mean scores for the different conditions were quite similar, the S.D. estimates could be averaged to produce a single value for each subject. These values are shown in Figure 3. The theoretical values shown in this

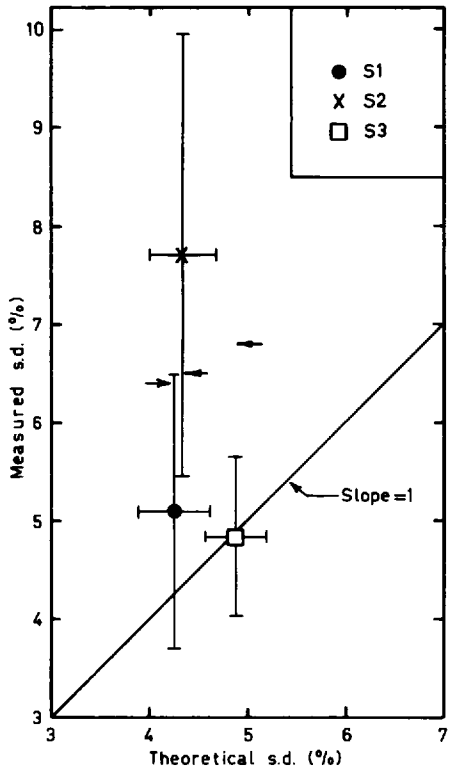

Figure 3. Measured test-retest S.D. versus predicted S.D. The predicted values are based on the more exact equation 2. The *error bars* show the 95% confidence limits for each value and the *arrowheads* show the values that are obtained when the simple binomial formula (equation 1) is used. Data are from Byrne & Walker.[2]

figure were obtained by analyzing the scores for individual items for each subject and then using formula 2.

It is evident from Figure 3 that the measured variability is quite close to that expected on the basis of the mixed difficulty model for two out of the three subjects. The error bars drawn correspond to ± twice the S.E. of the estimate and so correspond to the 95% confidence limits. Subject 2 is clearly more variable than the other two, and the obtained variability is significantly greater than that expected. The arrowheads indicate the variability predicted by the simple binomial theorem. Notice that the variability for subject 3 is significantly below this, and that for subject 1 above (although not significantly). Because of the different behavior of these subjects, the effect on variability of extending the test-retest period cannot be definitely stated. However, it clearly has not caused any additional variability for at least two of the subjects.

The data presented so far may be summarized as follows:

1. When applied to common speech discrimination tests, the simple binomial assumption predicts S.D. values that are 1 to 2% higher than those predicted when the mixed difficulty items are taken into account.

2. Despite this, the simple binomial model works quite well for the "average" subject (as shown by the large-scale experiments reported by Thornton and Raffin[23] and Raffin and Schafer.[18] However, when more detailed measurements are made on individual subjects, the reasons for this can be seen.

3. Some subjects (e.g., S3 of Fig. 3) display no apparent variability additional to that predicted by the more exact mixed-difficulty binomial model. For such subjects, the simple binomial model thus overestimates their speech discrimination score variability.

4. Some subjects (e.g., S2 of Fig. 3) display considerable variability additional to that predicted by the mixed-difficulty model. If sufficient additional variability exists, the simple binomial model will underestimate the variability.

5. Thus, although the simple binomial model provides a good estimate of the average variability, it will not necessarily give the correct estimate for any particular subject and test. (However, unless additional information is known about the subject and test, the simple binomial model probably provides the best estimate that can be made in a clinical situation.)

Composite Variability

When scores from different subjects are combined and treated as repeated measurements of the same subject, the dispersion of the scores so obtained cannot be attributed to the "variability" of the test. Such measurements are discussed here for two reasons. First, measurements made in this way are sometimes erroneously considered to be estimates of the test variability. Second, an indirect test of the binomial model is that such estimates should not be less than those expected on intrasubject grounds alone.

There have been two main methods of quantifying variability when both inter- and intrasubject differences have been lumped together. The first of these involves testing a group of subjects twice and then calculating a correlation coefficient based on the score each individual obtains in each test (e.g., Carhart and Tillman[4] and McConnell et al.[16]). This method is of no use to us in quantifying the relative importance of the different sources of variability because the correlation coefficient so obtained depends in a complex way upon the range of abilities shown by the subjects and the variability exhibited by each subject in the test (Beattie and Edgerton[1]). A high correlation between test and retest scores can always be obtained if the experimenter chooses (intentionally or otherwise) a sample population with a wide range of speech discrimination abilities. Conversely, low correlation coefficients can be guaranteed if the members of the sample population have very similar speech discrimination abilities. In addition, the correlation coefficient cannot be used to indicate the accuracy (or uncertainty) of any individual speech discrimination score. Thus, correlation coefficients are of no use as measures of speech test reliability.

The second method that has been used involves (e.g., Gengel and Miller[9] and Hughes et al.[12]) presenting the test to a group of subjects and then calculating the S.D. of the resulting distribution of scores. This procedure is more useful than the correlation coefficient method because the S.D. so calculated can be directly compared with that expected on the basis of intrasubject variability alone. Before such a comparison is made, one point should be emphasized.

The variation in true scores from subject to subject will be strongly influenced by the choice of subjects undergoing the test. A mixture of normal-hearing and profoundly deaf individuals, for example, would be expected to produce a greater variation in scores across individuals than a group of normals, or a group of similarly impaired individuals. Thus, statements about the second source of variability, subject differences, must be accompanied by a clear statement about the characteristics of the subjects involved.

Data suitable for estimating the size of typical intersubject differences has been provided by Elpern,[5,6] Kruel et al.,[13] Gengel and Miller,[9] Sanderson-Leepa and Rintelmann,[20] Penrod,[17] Schwartz and Surr,[21] and Sergeant et al.[22] Data from these studies will be presented in the same form as previously. Any significant increase in the difference between theoretical and experimental variability estimates will thus indicate that intersubject differences are at least comparable with the intrasubject dif-

ferences already discussed. Estimates of the S.D. of speech discrimination scores when measured across subjects have been obtained from the eight different studies mentioned above and are shown in Figures 4 and 5. The theoretical S.D. for both of these figures has been calculated from the properties of the simple binomial distribution only; i.e., no allowance has been made for mixed-difficulty items. As we have already seen, such an allowance would move each point somewhat to the left. The experimental points have been indicated by using closed symbols for those experiments involving normal-hearing subjects and open symbols for those involving hearing-impaired subjects. It is quite apparent that for the hearing-impaired subjects the measured intersubject S.D. is considerably larger than that expected from intrasubject considerations. On the basis of these data alone it is not possible to state whether this large variability is due to poor test-retest consistency for each subject, or to significant differences between the true ability of the various subjects. However, the results of several studies[2, 10, 18, 23] have shown that the test-retest variance for hearing-impaired subjects is similar to that expected from the binomial distribution. Thus, the large additional variability for hearing-impaired subjects (shown in Fig. 4) can probably be ascribed to differences in ability between the various subjects.

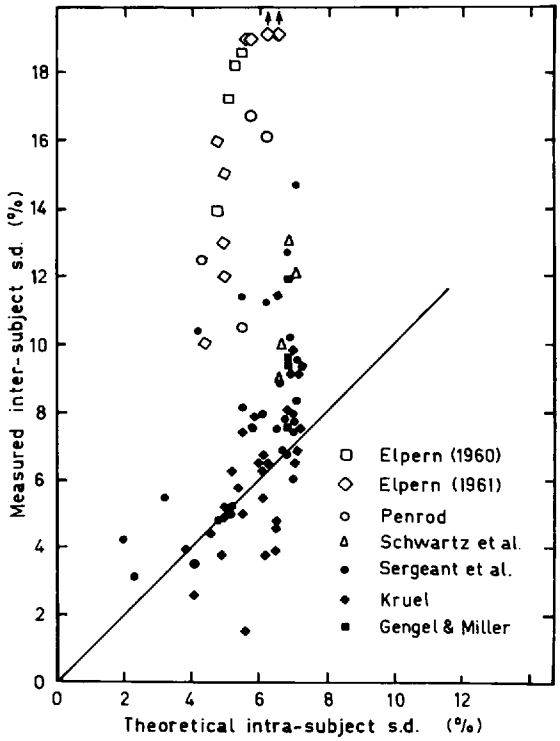

Figure 4. Measured intersubject variability versus predicted intrasubject variability (equation 1). *Open symbols* refer to data obtained from hearing-impaired subjects and *closed symbols* to normal-hearing subjects.

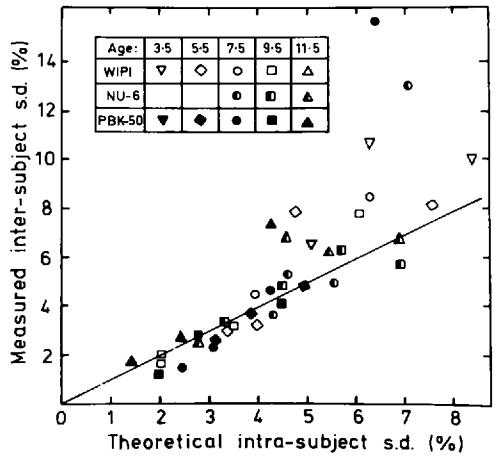

Figure 5. The same as for Figure 4, but data are from Sanderson-Leepa and Rintelmann[20] for children of various ages.

Intersubject variability for the normal-hearing subjects, however, seems to be of a size comparable with the intrasubject (or test-retest) variability, at least for these particular tests. The points shown in Figure 4 indicate a trend similar to that already observed for the test-retest variability. Largest deviations from the model (the line of unity slope) occur for the larger values of variability, which result from test scores in the range 30 to 70% correct.

Figure 5 shows a similar set of data, only this time obtained from children of various ages. There appears to be a slight trend for a greater intersubject variability to be measured for younger children than for older children. As before, this larger spread of scores cannot be interpreted to indicate a poorer test-retest consistency for these children, inasmuch as it is confounded with differences in the childrens' true abilities.

For all of the inter-subject variability data, the binomial model provides an approximate minimal variability. This minimal intersubject variability will be obtained only when all subjects in the group have equal speech-discrimination ability.

MISUSES OF VARIABILITY ESTIMATES

Variability estimates have been misused in a number of ways.

Variability of Different Populations

Subjects with a sensorineural hearing loss are often considered to have poorer test-retest reliability than normal-hearing subjects. To the author's knowledge, this has not been proven in any study, except on fallacious grounds. Comparison of the two types of subjects have not equated the groups for their overall level of performance (e.g., Engelberg[7]), one of the two chief determinants of intrasubject variability. Thus, under a given set of test conditions, impaired persons will score lower than normal hearers and so a higher variability will be appropriate (assuming that the normal hearers are obtaining

fairly high scores). The S.D. for a 50-item test (calculated from equation 1) is shown in Figure 6. It can be seen that the variability for the hearing-impaired subjects will approach that for the normal-hearing subjects only if the discrimination score for the hearing impaired approaches quite low values. The test conditions are usually arranged to avoid this situation (e.g., by testing at a higher sensation level or by using an easier test).

Similarly, Sanderson-Leepa and Rintelmann[20] have commented that for the WIPI test: "At a given sensation level, (intersubject) variability appears to bear an inverse relationship to the age of the subjects." Inasmuch as the test scores increased with the subject's age, this trend can be predicted simply because of the reduction in the intrasubject component of the total variability.

An extensive set of data has been gathered[18] which shows that the binomial distribution is an equally good predictor of score variability when it is applied to a hearing-impaired population as when it is applied to a normal-hearing population. (The normal hearers were tested under degraded signal conditions so that a range of test-retest scores was available for comparison with the binomial predictions.)

Test Reliability Validation

The use of shortened discrimination tests has been suggested by Elpern.[6] His data show that the same variability is obtained for 25-word lists as for 50-word lists, which is quite at variance with the result expected from binomial theory predictions. This occurs because his variability estimates are calculated across subjects and are dominated by intersubject differences. Thus, the change in the intrasubject component brought about by the shortened list remains undetected in his experiment.

Confusion about the various components of total variability seems to underlie one method that has been used for validating new speech tests. Sergeant et al.,[22] for example, say that "subject-by-subject variability is an index of test reliability" and thus seek to partially validate their test by showing that it has a small intersubject variability. This approach is based on the assumption that all people with hearing thresholds within a certain range will have exactly the same speech discrimination ability when measured on any test using speech material! It is thus hoped that the variability estimated is dominated by intrasubject factors. Since "normal" subjects vary to a greater or lesser extent on practically every human characteristic that is measurable, the assumption of equal speech discrimination abilities among "normally" hearing subjects appears to be quite unjustified. Of particular importance is the fact that peripheral properties of the auditory system (e.g., perception of temporal order, masking curves, and temporal integration) differ quantitatively from subject to subject. Because properties such as these are presumably involved in the speech recognition process, it seems likely that this ability also will vary. Thus, a test which results in all normally hearing subjects achieving about the same score may in fact be a test which is very insensitive to difference in speech discrimination ability—hardly a feature desirable in a speech discrimination test! Normally one desires that a test be maximally sensitive to changes in performance, so that different tokens of the system being tested (e.g., telecommunications channels, hearing aids, people) can be most efficiently measured and ranked. This requires that both the sensitivity and reliability of the test be considered before optimal performance is obtained. It may well be, for example, that the largest difference in the scores for normal and hearing-impaired subjects is obtained when different normal subjects achieve consistently different scores on a given test. A later paper will examine quantitatively the trade-offs between sensitivity and reliability that may be achieved in a speech discrimination test.

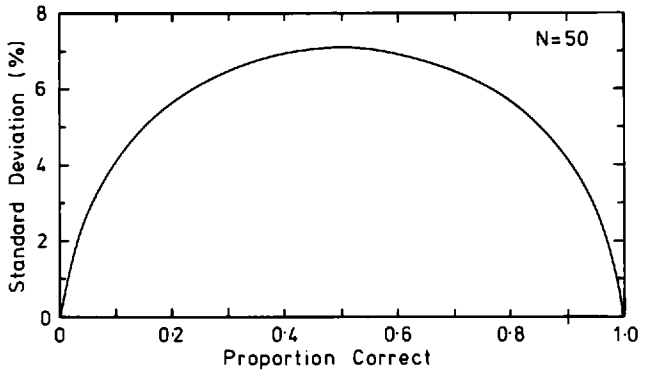

Figure 6. The test-retest S.D. for a 50-item test predicted by the simple binomial theorem (Fig. 1) as a function of the relative test score.

CONCLUSIONS

The intrasubject (or test-retest) variability of a speech discrimination score can be reasonably well predicted by assuming that the test scores arise from a simple binomial distribution. One of the consequences of this is that only the overall score and the number of test items need be known in order to estimate the accuracy of any particular test score. That the simple binomial model is successful in predicting variability is due to a combination of factors. First, speech discrimination tests contain items of mixed difficulty. This has the effect of decreasing the expected S.D. of test scores, typically by one to two percentage points. Second, the true performance of subjects may be different during the test and retest sessions (e.g., due to fatigue, learning, or a change in motivation or concentration). This has the effect of increasing the expected variability of test scores. These two factors thus combine to make the data fit the model better than if only one of them was present. Fortuitously, the amount of reduction caused by the first factor often (but not

always) happens to be about equal to the increase caused by the second factor. The fit to the binomial model has been shown to be excellent by several large-scale studies.[10, 18, 23] Unfortunately, each of these has analyzed data gathered in a single test session only, and has combined results across subjects in order to produce their variability estimates. The results reported in this paper (for three individuals tested many times over an extended period) show that the binomial model may underestimate the variability for some individuals.

One implication this has for clinical use is that the clinician should regard the variability predicted by the binomial theorem as the minimum likely to be obtained. Sometimes the clinician may be aware that a particular client is displaying greater than expected variability. (The first and second halves of a discrimination test may lead to widely different scores, for example.) Under these circumstances, stricter criteria for a significant difference between test scores should be adopted. If the variability of test scores is to be kept down to a value no greater than that predicted by the binomial theorem, it is clear that the experimental conditions (including the state of the subject) should be as similar as possible during the two tests the scores of which are to be compared. This may be more likely to occur if the results are obtained in a single test session, although an experimental validation of this hypothesis is required.

Although the results in this paper show that the binomial theorem may overestimate the variability for some subjects on some tests, this error is the more conservative one of not accepting a true difference as significant. The risk of incurring this error may be minimized by using a test comprised of items of similar difficulty. In the clinical situation, the properties of the binomial distribution can best be utilized through the use of published tables of critical differences.[19, 23]

The point is also made that intersubject variability should not be confused with intrasubject variability. Furthermore, a low value for intersubject variability when using normal subjects is not necessarily a desirable feature in a speech discrimination test because it could indicate a test which is insensitive to differences in speech discrimination ability.

References

1. Beattie, R. C., and B. J. Edgerton. 1976. Reliability of monosyllabic discrimination tests in white noise for differentiating among hearing aids. J. Speech Hear. Disord. **16**, 464–476.
2. Byrne, D., and G. Walker. 1980. Effects of multiband compression and expansion on intelligibility and quality of speech. 4th Audiological Society of Australia Conference, Melbourne.
3. Campbell, R. A. 1965. Discrimination test word difficulty. J. Speech Hear. Res. **8**, 13–22.
4. Carhart, R., and T. W. Tillman. 1972. Individual consistency of hearing for speech across diverse listening conditions. J. Speech Hear. Res. **15**, 105–113.
5. Elpern, B. S. 1960. Differences in difficulty among the CID W-22 auditory tests. Laryngoscope **70**, 1560–1565.
6. Elpern, B. S. 1961. The relative stability of half-list and full-list discrimination tests. Laryngoscope **71**, 30–36.
7. Engelberg, M. 1968. Test-retest variability in speech discrimination testing. Laryngoscope **78**, 1582–1589.
8. Fairbanks, G. 1958. Test of phonemic differentiation: the rhyme test. J. Acoust. Soc. Am. **30**, 596–600.
9. Gengel, R. W., and L. Miller. 1976. A clinical pass/fail criterion for word discrimination in noise. Paper presented at the Annual Convention of the American Speech and Hearing Association, Houston.
10. Hagerman, B. 1976. Reliability in the determination of speech discrimination. Scand. Audiol. **5**, 219–228.
11. Hood, J. D., and J. P. Poole. 1977. Improving the reliability of speech audiometry. Br. J. Audiol. **11**, 93–101.
12. Hughes, E. C., R. H. Arthur, and R. L. Johnson. 1979. Test-retest variability in testing hearing of speech. J. Am. Aud. Soc. **5**, 17–20.
13. Kruel, E. J., D. W. Bell, and J. C. Nixon. 1969. Factors affecting speech discrimination test difficulty. J. Speech Hear. Res. **12**, 281–287.
14. Levitt, H., and S. B. Resnick. 1978. Speech reception by the hearing-impaired: methods of testing and the development of new tests. pp. 107–130. *in* C. Ludvigsen, and J. Barfod, eds. *Sensorineural hearing impairment and hearing aids.* Scand. Audiol. Suppl. **6**.
15. Lyregaard, P. E. 1973. On the statistics of speech audiometry data. *National Physics Laboratory Acoustics Report Ac 63*, Department of Industry, Teddington.
16. McConnell, F., E. F. Silber, and D. McDonald. 1960. Test-retest consistency of clinical hearing aid tests. J. Speech Hear. Disord. **25**, 273–280.
17. Penrod, J. P. 1979. Talker effects on word-discrimination scores of adults with sensorineural hearing impairment. J. Speech Hear. Disord. **44**, 340–349.
18. Raffin, M. J., and D. Schafer. 1980. Application of a probability model based on the binomial distribution to speech discrimination scores. J. Speech Hearing Res. **23**, 570–575.
19. Raffin, M. J., and A. R. Thornton. 1980. Confidence levels for differences between speech-discrimination scores: a research note. J. Speech Hear. Res. **23**, 5–18.
20. Sanderson-Leepa, M. E., and W. F. Rintelmann. 1976. Articulation functions and test-retest performance of normal-hearing children on three speech discrimination tests: WIPI, PBK-50, and NU Auditory Test No. 6. J. Speech Hear. Disord. **41**, 503–519.
21. Schwartz, D. M., and R. K. Surr. 1979. Three experiments on the California consonant test. J. Speech Hear. Disord. **44**, 61–72.
22. Sergeant, L., J. E. Atkinson, and P. G. Lacroix. 1979. The NSMRL Tri-Word Test of Intelligibility. J. Acoust. Soc. Am. **65**, 218–222.
23. Thornton, A. R., and M. J. M. Raffin. 1978. Speech discrimination scores modeled as a binomial variable. J. Speech Hear. Res. **21**, 507–518.

Acknowledgments: The author wishes to thank Denis Byrne and Gary Walker for the provision of the raw data referred to in this paper, and Denis Byrne for his helpful comments on an earlier version of this manuscript.

Address reprint requests to Dr. H. Dillon, National Acoustic Laboratories, 5 Hickson Road, Millers Point, Sidney, 2000, Australia.

Received February 13, 1981; accepted August 24, 1981.