Applications of Generalizability Theory to Measurement of Individual Differences in Speech Perception

Marilyn E. Demorest
University of Maryland Baltimore County

Lynne E. Bernstein
Center for Auditory and Speech Sciences
Gallaudet University

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides an integrated framework for evaluating sources of variability in test scores. Two examples of its application to testing speech perception are reviewed: a study of NU-6 word lists (Demorest & Cord, 1993) and an experiment on the Speech Perception in Noise test (Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984). Generalizability coefficients (which are analogous to reliability coefficients) are derived from the data of Demorest and Cord (1993) and the impact of different types of test score interpretation on generalizability are outlined. The relation between test length and generalizability is illustrated with speechreading data from Demorest and Bernstein (1992). Results of generalizability analyses can be used to design more reliable and efficient test procedures.

Assessment of individual differences in speech perception requires standardized tests that are sensitive to relevant sources of variability in test scores and insensitive to irrelevant, extraneous sources of variability. The former characteristic is considered evidence of test validity, whereas the latter is concerned with reliability. Because reliability is necessary, but not sufficient, for validity, investigation of unwanted sources of variability in test scores is critical in the development and evaluation of psychometrically sound measures. Without estimates of reliability and its counterpart, measurement error, one cannot know whether differences in scores obtained under different testing conditions (e.g., aided vs. unaided) or at different times (e.g., before and after training) are truly different, or whether observed differences are just a byproduct of random test-score variability.

Reliability has traditionally been evaluated by examining extraneous sources of variability independently of one another. For example, *retest reliability* evaluates the consistency of test scores over time, with test occasion being the ex-

traneous variable. *Alternate-form reliability* evaluates the consistency of scores over different test forms, with test form being the extraneous variable. *Split-half reliability* and *internal consistency reliability* evaluate consistency of performance over items within a single test form, and *interscorer reliability* reflects consistency across scorers.

GENERALIZABILITY THEORY

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is a statistical theory of sources of variability in behavioral observations that permits estimation of the effects of several extraneous variables, and their interactions, within a single experiment. A *generalizability study* is an experiment in which potential sources of variability in test scores are manipulated. A statistical model for a single observation and an analysis-of-variance model appropriate for the experimental design are specified. Next, expected values of the mean squares from the analysis of variance are determined and used to estimate the variance component for each source of variability in the observations.

Generalizability of NU-6 Word-Recognition Scores

As an example, consider a study conducted by Demorest and Cord (1993) in which four monosyllabic word lists (Auditec recordings of NU-6) were administered on each of 2 days to a sample of 40 hearing-impaired adults. Subjects were recruited from the Aural Rehabilitation Program at Walter Reed Army Medical Center and typically had mild-to-moderate bilateral sensorineural hearing loss. The sources of variability were the test list and the test occasion. The statistical model for the score of one subject on a given list on a given day is:

$$\chi = \mu + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \epsilon, \quad (1)$$

where μ is a grand mean, the α parameters represent the effects of Subject (α_1), List (α_2), Day (α_3), List $\times$ Day (α_4), Subject $\times$ List (α_5), and Subject $\times$ Day (α_6), respectively, and ϵ is random, residual error. Given this model for a single score, the variance of observed scores is:

$$\sigma_\chi^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2 + \sigma_6^2 + \sigma_\epsilon^2, \quad (2)$$

where the subscripts on the variances correspond to those of the alpha-parameter and error effects in Equation 1. The goal of generalizability analysis is to estimate each of these variance components and their contribution to the total observed variance. Of the seven variance components in Equation 2, only the first, that for subject (σ_1^2), produces relevant variance; all other sources are extraneous to the purposes of testing. Thus, ideally, all other variance components should be zero. For example, if there were no differences among lists, the variance component for lists (σ_2^2) would be zero. Similarly, if subjects' mean performance was the same from one occasion to the next, there would be

no variance attributable to the day of testing ($\sigma_3^2 = 0$). Because these extraneous effects on test scores are typically not zero, it is important to estimate their magnitude and to determine their impact on test score interpretation.

The sources of variability in the experiment by Demorest and Cord (1993) are shown in Table 1. The mean squares symbolized in the third column of the table would be used in an analysis of variance to form *F* ratios for testing whether the first six sources of variability were statistically significant (i.e., non-zero).[1] The expected value of each mean square in an analysis of variance (i.e., the quantity that the mean square estimates) is a weighted sum of the variance components. For example, the expected value of the mean square for the interaction of Subject × Day, MS_6, is $4\sigma_6^2 + \sigma_\epsilon^2$, whereas the expected value of the mean square for residual error, MS_E, is σ_ϵ^2. The mean square for subject, MS_1, estimates not only the individual differences among subjects, but also variance attributable to the interaction of Subject × Day.[2]

Table 1

Analysis of Variance Table With Expected Values of Mean Squares

Source	*df*	*MS*	Expected value of *MS*
1. Subject (S)	39	MS_1	$8\sigma_1^2 + 4\sigma_6^2 + \sigma_\epsilon^2$
2. List (L)	3	MS_2	$80\sigma_2^2 + 40\sigma_4^2 + 2\sigma_5^2 + \sigma_\epsilon^2$
3. Day (D)	1	MS_3	$160\sigma_3^2 + 4\sigma_6^2 + \sigma_\epsilon^2$
4. L × D	3	MS_4	$40\sigma_4^2 + \sigma_\epsilon^2$
5. S × L	117	MS_5	$2\sigma_5^2 + \sigma_\epsilon^2$
6. S × D	39	MS_6	$4\sigma_6^2 + \sigma_\epsilon^2$
7. Error	117	MS_E	σ_ϵ^2
Total	319		

Note. Subject and day are random effects, list is a fixed effect, and the interaction of Subject × List × Day is assumed to be zero.

The next step in the analysis is to obtain estimates of the variance components in Equation 2. By equating each mean square to its expected value, estimates of the variance components can be obtained. For example, $(MS_6 - MS_E)/4$ provides an estimate of σ_6^2, and $(MS_2 - MS_4 - MS_5 + MS_E)/80$ gives an

[1] In this design, both Subject and Day are considered random effects. List is a fixed effect, and the interaction of Subject × List × Day is assumed to be zero. Given this model, some effects must be tested using a quasi *F* ratio (*F* ′) (Winer, 1971).

[2] The weights that appear with the variance components reflect the numbers of observations contributing to each mean upon which the mean square is based. For example, the Subject × Day interaction is based on 80 means (40 subjects × 2 days), and each of these means is computed across four lists.

Table 2

Estimation of Population Variance Components

Source	Estimator	Fixed-effect correction
1. Subject (S)	$(MS_1 - MS_6)/8$	1
2. List (L)	$(MS_2 - MS_4 - MS_5 + MS_E)/80$	3/4
3. Day (D)	$(MS_3 - MS_6)/160$	1
4. L × D	$(MS_4 - MS_E)/40$	3/4
5. S × L	$(MS_5 - MS_E)/2$	3/4
6. S × D	$(MS_6 - MS_E)/4$	1
7. Error	MS_E	1

Note. Subject and day are random effects, list is a fixed effect, and the interaction of Subject × List × Day is assumed to be zero.

estimate of $\sigma_2{}^2$. An algorithm for deriving the expected values of mean squares in analysis of variance is given in Winer (1971). One additional computational adjustment must be made for those sources of variability that involve fixed effects. If a fixed effect has k levels, each component involving that effect must be multiplied by (k − 1)/k. The formulas in Table 2 were generated following these principles.

Results from analysis of the Demorest and Cord (1993) data are shown in Table 3. Inspection of the *F* ratios suggests statistically significant effects for subject and list. Subject is the largest source of variability, accounting for an estimated 81.1% of the total variance of observed scores. The list effect, although non-zero, contributes very little to the variance of observed scores. This illustrates an important point: *The magnitude of an effect cannot be directly inferred from statistical significance.* Statistical power is high when effects are estimated from a large number of observations (e.g., 80 observations per list), yet the magnitude of the effect may be extremely small and of no practical significance. Effects for day and the interaction of List × Day produce negative variance estimates,[3] which have been set to zero. The Subject × List interaction is not statistically significant ($p > .05$), which implies that the differences among lists are the same from one subject to the next. The Subject × Day interaction is larger and is significant both statistically ($p < .01$) and clinically. Day-to-day variability in scores differs somewhat from one subject to the next (or equivalently, the individual differences among subjects differ somewhat from one day to the next). This effect accounts for an estimated 6.8% of the total variance. Residual error variance, the final component, accounts for 9.5% of

[3]Negative variance estimates occur when larger mean squares are subtracted from smaller ones (see formulas in Table 2).

Table 3

Estimates of Population Variance Components for NU-6 Word Recognition

Source	Mean square	*F* or *F'*	Variance estimate	Proportion of total variance
1. Subject (S)	484.81	18.58	57.34	.811
2. List (L)	37.91	8.72	0.32	.005
3. Day (D)	13.61	0.52	0[a]	0[a]
4. L × D	0.31	0.05	0[a]	0[a]
5. S × L	10.77	1.60	1.52	.021
6. S × D	26.09	3.88	4.84	.068
7. Error	6.73		6.73	
Total			70.74	

Note. Data are from *Evaluation of Temporal and Interlist Sources of Variability in NU-6 Test Scores: A Generalizability Analysis* by M.E. Demorest and M. Cord, 1993, Manuscript in preparation. Reprinted by permission. Subject and Day are random effects; List is a fixed effect; and the interaction of Subject × List × Day is assumed to be zero.

[a]This variance estimate was negative and was set equal to zero.

the variance.

Based on the results in Table 3, total variance of observed scores in this experiment is estimated as follows:

$$\hat{\sigma}_x^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 + \hat{\sigma}_4^2 + \hat{\sigma}_5^2 + \hat{\sigma}_6^2 + \hat{\sigma}_\epsilon^2, \quad (3)$$

$$70.74 = 57.34 + 0.32 + 0 + 0 + 1.52 + 4.84 + 6.73. \quad (4)$$

From these estimates, it is possible to derive theoretical predictions about the reliability (or generalizability) of scores on these NU-6 word lists under a variety of clinical testing protocols (see Generalizability Coefficients below).

Generalizability of Scores on the Speech Perception in Noise (SPIN) Test

Bilger, Nuetzel, Rabinowitz, and Rzeczkowski (1984) performed a generalizability analysis of the Speech Perception in Noise (SPIN) test in which several variables were manipulated. A sample of 128 hearing-impaired adults responded to the 10 forms of the test, each of which contained high- and low-context items. Half the subjects were tested through headphones, half with loudspeakers. Half were tested in a single session, half in two sessions 2-4 weeks apart. In addition, there were two methods of recording responses: immediate write-down by the examiner (Marker 1) versus transcription by an independent observer from a recording of the subject's response (Marker 2). There were also eight orders of testing used.

Because of the large number of subjects in their study, the large number of

test forms, and the multiple conditions of testing/scoring, Bilger et al. (1984) found that many irrelevant sources of variability were statistically significant (i.e., non-zero). Their variance component analysis, however, revealed that the magnitude of many of these effects was trivial. They concluded that "SPIN scores were not influenced in an important way by the choice of transducer, the number of visits required to complete testing, or the order in which the forms were administered" (p. 36). In contrast, relatively large interactions were obtained for Subject × Context, Subject × Form, and Subject × Context × Form. This led Bilger et al. to conclude that the high- and low-context items produced two different types of speech test and that they should be scored, and evaluated, separately. Accordingly, they specified a simplified model for each context containing three independent variables (subject, form, and marker) and their interactions. Subject was considered a random effect, and form and marker were considered fixed effects. Results obtained with these models are shown in Tables 4 and 5.[4] For both contexts, Subject was the largest source of variance,

Table 4

Estimation of Population Variance Components: SPIN High-Context Sentences

Source	Mean square	*F*	Variance estimate	Proportion of total variance
1. Subject (S)	446.34	6468.03	22.31	.902
2. Form (F)	71.26	15.03	0.23	.010
3. Marker (M)	5.25	42.33	0.00	.000
4. S × F	4.74	68.73	2.10	.085
5. S × M	0.12	1.80	0.00	.000
6. F × M	0.12	1.81	0.00	.000
7. Error	0.07		0.07	.003
Total			24.72	

Note. Adapted from "Standardization of a Test of Speech Perception in Noise" by R.C. Bilger, J.M. Nuetzel, W.M. Rabinowitz, and C. Rzeczkowski, 1984, *Journal of Speech and Hearing Research, 27,* p. 39. Copyright 1984 by the American Speech-Language-Hearing Association. Adapted by permission. Subject is a random effect, all other effects are fixed, and the interaction of Subject × Form × Marker is assumed to be zero.

accounting for 90.2% and 81.7% of the total variance for high- and low-context sentences respectively. The main effect of test form and the Subject × Form interaction were statistically significant for both contexts. For high-context sentences the differences among forms were negligible, but there was a non-trivial interaction of Subject × Form, indicating that the form differences were not the

[4]For unknown reasons, variance estimates for the Subject × Form interaction differ slightly from those reported by Bilger et al. (1984).

Table 5
Estimation of Population Variance Components: SPIN Low-Context Sentences

Source	Mean square	*F*	Variance estimate	Proportion of total variance
1. Subject (S)	561.25	4165.63	28.06	.817
2. Form (F)	494.20	49.37	1.70	.050
3. Marker (M)	1.62	14.13	0.00	.000
4. S × F	10.01	74.30	4.44	.129
5. S × M	0.11	0.85	0.00	.000
6. F × M	0.09	0.64	0.00	.000
7. Error	0.13		0.13	.004
Total			34.34	

Note. Adapted from "Standardization of a Test of Speech Perception in Noise" by R.C. Bilger, J.M. Nuetzel, W.M. Rabinowitz, and C. Rzeczkowski, 1984, *Journal of Speech and Hearing Research, 27,* p. 39. Copyright 1984 by the American Speech-Language-Hearing Association. Adapted by permission. Subject is a random effect, all other effects are fixed, and the interaction of Subject × Form × Marker is assumed to be zero.

same from one subject to the next (or equivalently, the individual differences among subjects were not quite the same from one form to the next). These effects were even larger for the low-context sentences, and this led Bilger et al. to recommend reorganization of the SPIN items into new forms that would be more nearly equivalent (see Bilger, 1984). Of particular importance was the finding that, despite their statistical significance, differences between the two markers/methods of scoring were virtually zero, and this factor did not interact with subject or form. Thus, in clinical application, the two scoring methods could be used interchangeably.

GENERALIZABILITY COEFFICIENTS

Generalizability analysis yields *coefficients of generalizability* which are analogous to reliability coefficients. Each coefficient is based on a data collection model for obtaining test scores and a universe of generalization for test score interpretation. Together these determine which sources of variability affect observed scores and universe scores. (The latter are analogous to true scores in classical test theory.) The coefficient equals the ratio of universe-score variance to observed-score variance:

$$\rho^2 = \frac{\sigma^2_{universe}}{\sigma^2_{observed}}. \tag{5}$$

For example, consider a data collection model in which a single NU-6 word list is presented to a subject on a given day. All subjects are tested with the

same list. The variance of the observed scores in the population would be:

$$\sigma^2_{observed} = \sigma^2_S + \sigma^2_{S\times L} + \sigma^2_{S\times D} + \sigma^2_\epsilon. \quad (6)$$

That is, variance of the observed scores is influenced not only by individual differences among subjects ($\sigma_S{}^2$), but also by interactions of Subject × List ($\sigma_{S\times L}{}^2$) and Subject × Day ($\sigma_{S\times D}{}^2$), and residual error ($\sigma_\epsilon{}^2$).

Different universes of generalization result in different formulas for the numerator of Equation 6. One possibility is to define the universe score as the subject's expected score across all lists and across days. When universe scores are defined in this way, the formula for universe-score variance is:

$$\sigma^2_{universe} = \sigma^2_S. \quad (7)$$

Another way to define the universe score is: the subject's expected score across lists on the day of testing. Thus there is generalization across lists, but the score is interpreted as specific to that day. The formula for universe-score variance becomes:

$$\sigma^2_{universe} = \sigma^2_S + \sigma^2_{S\times D}. \quad (8)$$

The variance of universe scores contains two components: one for subjects (irrespective of the day of testing) and one for the interaction of Subject × Day. A universe score that is specific to a particular day contains not only the main effect of subject, but also the interaction of Subject × Day.

A third way to define the universe score is: the subject's expected score on this list across days. The test score is interpreted as specific to a given list, but there is generalization across days. A universe score that is specific to a particular list contains not only the main effect of subject but also the interaction of Subject × List. The formula for universe-score variance is analogous to Equation 8, but contains a component for the interaction of Subject × List rather than Subject × Day:

$$\sigma^2_{universe} = \sigma^2_S + \sigma^2_{S\times L}. \quad (9)$$

It is, of course, possible to define the universe score without generalizing across either lists or days. The test score is then interpreted as specific to a given list and to the day of testing, and the universe score contains the effect of subject, as well as the interactions of Subject × Day and Subject × List. Universe-score variance for this definition is given by:

$$\sigma^2_{universe} = \sigma^2_S + \sigma^2_{S\times L} + \sigma^2_{S\times D}. \quad (10)$$

Comparison of Equations 7-10 shows clearly that the generalizability coefficient will be highest when test score interpretation is restricted to the same test list and test day and lowest when generalization is across both lists and days. This result occurs because testing was conducted on a single day with a single list. Generalizing beyond the conditions of test administration is less accurate than

restricting test score interpretation to the conditions under which the scores were obtained.

Generalizability coefficients estimated from the data of Demorest and Cord (1993) are shown in Table 6 for the four universes of generalization represented in Equations 7-10. Although the expected differences among the coefficients are obtained, all of the generalizability coefficients are relatively high. This indicates that the NU-6 scores for these lists in this client population are sensitive to individual differences in word recognition and relatively insensitive to extraneous variables such as list and day of testing. The generalizability coefficients could be raised even further if testing were conducted on more than one day or with more than one list.

Table 6

Generalizability Coefficients and Mean Reliability Coefficients for NU-6 Words Under Four Universes of Generalization

Universe of generalization	Generalizability coefficient	Mean reliability coefficient
Across lists and days	.814	.808
Across lists for a given day	.883	.877
Across days for a given list	.836	.832
A given list on a given day	.904	

Note. Data are from *Evaluation of Temporal and Interlist Sources of Variability in NU-6 Test Scores: A Generalizability Analysis* by M.E. Demorest and M. Cord, 1993, Manuscript in preparation. Reprinted by permission.

As noted in the introduction, it has been customary to estimate test reliability by calculating correlation coefficients for scores obtained under various test conditions. Generalizability coefficients for the first three universes of generalization in Table 6 are conceptually and theoretically equivalent to delayed alternate-form, alternate-form, and retest reliability coefficients, respectively. Mean values for these coefficients from Demorest and Cord (1993) are also shown in Table 6 and it is apparent that the agreement is quite good. However, generalizability theory also makes it possible to estimate immediate retest reliability (same list, same day), even though no immediate retests were given.

GENERALIZABILITY AND TEST LENGTH

Generalizability theory is especially useful for estimating the number of test items needed to achieve a particular level of generalizability. As a general rule, when test lists or forms are lengthened by a factor of k, variance components that involve test lists are divided by k. A generalizability function can be gen-

erated by substituting various values of k and determining the estimated generalizability. For example, Demorest and Bernstein (1992) performed a generalizability analysis on speechreading data from 104 subjects with normal hearing who viewed 100 video-recorded CID Everyday Sentences (Davis & Silverman, 1970), 50 for each of two talkers. The dependent variable was the total number of words correct on a single sentence. Generalizability coefficients were estimated for five models of data collection and generalization, three of which were defined as follows:

Model 1: Test with a single talker, generalize over all test items by this talker.

Model 2: Test with a single talker, but generalize over all test items and *both* talkers.

Model 3: Test some subjects with one talker, others with the other talker; generalize over all test items and both talkers.

As can be seen in Figure 1 (adapted from Demorest & Bernstein, 1992), all three functions begin to plateau at about 30-40 items, suggesting that for these recordings of the CID sentences (Bernstein & Eberhardt, 1986) individual differences among subjects with normal hearing can be estimated with about 40 sentences. Generalizability is highest for Model 1 and worst for Model 3. As with the examples in Table 6 from Demorest and Cord (1993), generalizability is

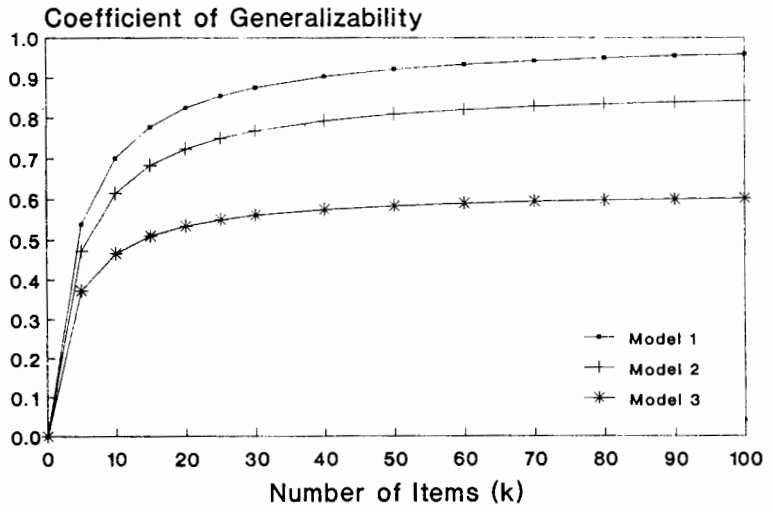

Figure 1. Generalizability as a function of test length (number of items) for three models of data collection and interpretation. Adapted from "Sources of Variability in Speechreading Sentences: A Generalizability Analysis" by M.E. Demorest and L.E. Bernstein, 1992, *Journal of Speech and Hearing Research, 35,* p. 882. Copyright 1992 by the American Speech-Language-Hearing Association.

most accurate when test score interpretation coincides with the conditions of testing. If a single talker is used for testing, it is better to restrict test interpretation (i.e., generalization) to that talker (Model 1 vs. Model 2). Moreover, it is better to have constant test conditions for all subjects. Testing different subjects with different talkers (Model 3) produces much greater variability in observed scores and lowers generalizability even further. Yet this model corresponds to what occurs clinically when different recordings are used by different clinics or when live-voice testing is done.

CONCLUSION

Generalizability theory provides an integrated framework for evaluating multiple sources of variability in behavioral observations and for deriving implications for test development and test score interpretation. Issues of test length, interlist equivalence, temporal variability, stimulus presentation conditions, scoring methods, and talker effects can all be examined within a comprehensive model of measurement error. After significant sources of variability have been identified, steps can be taken to control and/or estimate their effects. Testing protocols used for individual clients or for program evaluation can be designed with these goals in mind. For example, it is clear that relatively minor changes in testing procedures, such as the use of standardized, recorded materials, can have a significant impact on test score variability and generalizability. At the same time, it is also important to know which potential sources of variability do *not* have large effects, because this justifies flexibility in test administration procedures and test score interpretation. Generalizability theory has only recently begun to be applied in the domain of speech perception, but, as illustrated above, it can provide valuable insights about the magnitude of the impact of extraneous variables on measurement of individual differences. This knowledge can be used to improve the reliability, validity, and efficiency of testing, issues that are critical in the current climate of concern with the cost and effectiveness of clinical services.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant DC00695. The authors gratefully acknowledge the very helpful comments of three anonymous reviewers.

REFERENCES

Bernstein, L.E., & Eberhardt, S.P. (1986). *Johns Hopkins lipreading corpus I-II: Disc I* [Videodisc]. Baltimore: The Johns Hopkins University.

Bilger, R.C. (1984). Speech recognition test development. In E. Elkins (Ed.), *ASHA Reports 14: Speech recognition by the hearing impaired* (pp. 2-15). Rockville, MD: American Speech-Language-Hearing Association.

Bilger, R.C., Nuetzel, J., Rabinowitz, W.M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research, 27,* 32-48.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Davis, H., & Silverman, S.R. (Eds.). (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart, & Winston.

Demorest, M.E., & Bernstein, L.E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research, 35,* 876-891.

Demorest, M.E., & Cord, M. (1993). *Evaluation of temporal and interlist sources of variability in NU-6 test scores: A generalizability analysis.* Manuscript in preparation.

Winer, B.J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.