Chapter 16

Treatment Efficacy in Adult Audiological Rehabilitation

Allen A. Montgomery
University of South Carolina

In this chapter the problems of defining and measuring treatment efficacy in audiological rehabilitation for the adult are addressed from several points of view. First, a definition and commentary are presented, focusing on the need for outcome measures based on real world performance. This is followed by a discussion of outcome measures. Statistical problems found in improvement scores and test development are considered and an experimental clinical outcome measure with some potential for allowing ecologically valid inferences

is presented. Finally, suggestions for future research and general research directions are made.

DEFINITION OF TREATMENT EFFICACY

Presently the measurement of treatment efficacy is an important topic in almost every profession (ASHA has shown recent interest [Frattali, 1992; Olswang, Thompson, Warren, & Minghetti, 1990]; ASHA, 1992) and it is especially important in audiological rehabilitation where no large scale investigations of efficacy have been conducted.

Treatment *efficacy* has been defined as "the probability of benefit to individuals in a defined population from a medical technology applied for a given medical problem under ideal conditions of use" (Office of Technology Assessment, 1978, p. 16). Treatment *effectiveness*, on the other hand, is the probability of such benefit "under ordinary conditions by the average practitioner for the typical patient" (Office of Technology Assessment, 1978, p. 16). Thus, efficacy is an idealized concept – what one can expect of a particular clinical procedure at its best, whereas effectiveness refers to the results of the procedure applied in everyday practice. Most attention, not surprisingly, has been on establishing efficacy, because if a clinical procedure is not efficacious, it cannot be effective in routine use.

The definition of efficacy implies several things:

1. It implies that large numbers of subjects must be evaluated if one is to establish the probability of benefit.

2. It assumes the capability to classify clients accurately into discrete diagnostic categories. Ideally, the categories should reflect prognosis and also produce relatively homogeneous groups of clients. That is, they should be more specific than the general audiometric severity levels ranging from mild to profound loss. Thus, for example, two individuals, one with and one without dementia, or one with and one without English as the primary language should fall into different diagnostic/prognostic categories because they, presumably, would respond differently to clinical procedures.

3. It implies that benefit must be carefully defined. There are three aspects to measuring benefit that must be considered. First, benefit is based on an outcome measure that identifies and quantifies the presence and extent of the benefit. Second, a level of performance on the outcome measure must be set so that clients scoring above that level may be counted as receiving benefit, and the probability of such benefit can be calculated. This, in turn, implies that our profession must begin to agree on those levels and go beyond the practice of merely reporting treatments that produce statistically significant results. Significance at the .05 level is largely influenced by the number of subjects and the type of experimental design (such as repeated-measures designs) and may bear little relation to the actual benefit produced by the treatment. Third, benefit must be defined in a frame of reference. It is quite reasonable to consider benefit as the maintenance of a client's current communication performance level (in

the presence of other deterioration) or as performance levels that decrease more slowly than the average for a specific diagnostic group. Alternatively, benefit could be defined as being relative to the effects of no treatment or relative to a standard or previously used treatment.

Treatment efficacy was defined narrowly above. A final working definition of treatment efficacy translated into audiological rehabilitation terms might be:

> Treatment efficacy is the probability that individuals with hearing impairment in a carefully-defined diagnostic category will benefit from the application of a specific audiological rehabilitation procedure as determined by performance above a predetermined level on an outcome measure that meets stated standards for reliability and for validity of inferences that are typically drawn from it.

THE NATURE OF OUTCOME MEASURES

The Goal of Adult Audiological Rehabilitation

One of the most important concepts for efficacy research is that of the outcome measure. It is important because the selection of outcome measures, in essence, indicates what the profession considers to be the goals of audiological rehabilitation. Thus, if one's goal is to increase a client's use of conversation repair strategies, the outcome measure should quantify some aspect of the client's repair performance in conversations. In this chapter we adopt the position that the overall goal of adult audiological rehabilitation is to increase the likelihood that mutually satisfactory communication will occur between the client and people in his/her everyday environment (see Montgomery, 1993, p. 314 for elaboration). Thus outcome measures will be considered useful and accurate to the extent that they reflect the quality and extent of communication behavior in everyday communication situations. This emphasis on improvement in real world communication as the goal for adult aural rehabilitation is consistent with current statements of audiological rehabilitation goals in general (Alpiner, Kaufman, & Hanavan, 1993, p. 3).

The Nature of Everyday Communication

To understand the overall goal of adult aural rehabilitation and to appreciate the difficulties associated with designing appropriate outcome measures, one must have in mind the elements and processes that underlie everyday communication. To illustrate some of these, we have chosen the prototypical communication situation, the conversation (Levelt, 1989).

In a conversation, two or more people engage in an exchange of verbal and nonverbal symbols that indicate the content and the intention of the message being conveyed. The process is highly interactive and greatly fragmented with one interlocutor responding to the other with elaborations on the conversational topic, abrupt changes of topic, requests for clarification, verification of message reception, attempts to begin speaking or continue speaking when faced with competition, and continual changes of role and status as the intentions and purposes of the communication evolve. All of these aspects of conversing are controlled by complex implicit restrictions on behavior (pragmatic rules) that vary according to culture, social class, and situation. In addition to these rapid interactive processes, the interlocutor must also deal with messages that are obscured by audible and visible noise. Finally, the messages are filtered by his or her current emotional state and mapped onto a database of linguistic and world knowledge for interpretation.

The adult with impaired hearing thus faces a daunting task: he or she, faced with reduced auditory processing of the rapid, fragmented speech, must make additional use of channel complementarity (lipreading), linguistic and world knowledge redundancy (context), and pragmatics (topic verification and conversation repair) as well as attempt to control the situation (assertively act to reduce noise) in order to increase the probability of successful communication. Thus the charge to the client might be: complement, compensate, contextualize, control, and communicate. As rehabilitative audiologists or speech-language pathologists, we face an equally daunting task: We must devise outcome measures that indicate the client's success in actually doing these things. In the next section we discuss outcome measures and diagnostic tests that have been developed to assess each of five factors that influence the client's ability to communicate.

OUTCOME MEASURES

Outcome measures are the interface between the profession's abstract goals for rehabilitation and the client's response to specific clinical procedures. An outcome measure is derived from the constructs implicit in a particular goal and contains items (observations, ratings, questions, stimuli, etc.) that elicit information on the presence of that construct in the client's or other respondent's behavior. That is, an outcome measure is a test in the broadest sense of the word. In treatment efficacy research, an outcome measure is a test applied to a cohort of clients in such a way that a statement can be made of the probability of benefit due to a clinical procedure or procedures.

It is convenient to consider outcome measures according to which aspects of the communication breakdown they focus upon. A simple description of the major components of speech communication that contribute to the communication difficulties of adults with a hearing impairment as those components relate to rehabilitation is shown in Table 1.

Signal Presence and Quality

The quality of the signal reaching the cochlea or retina is dependent upon the amplification/vision correction that is available and upon the mechanical/optical transmission of the signal through the middle ear and refractive system of the auditory and visual systems, respectively. Moreover, the presence of the visual signal at the retina is dependent upon the client attending visually to the talker's

Table 1

Definitions of Components of Speech Communication
That May Contribute to Communication Difficulty in the Adult

1. **Signal presence and quality** – level and signal-to-noise ratio of auditory and visual speech signals reaching the client's sensory cells
2. **Sensory processing capability** – signal transduction through auditory and visual channels
3. **Speech recognition behavior** – (amplified) auditory (A) and visual (V) and combined auditory-visual (A-V) recognition of speech (including client's knowledge base as well as cognitive and linguistic correlates of speech understanding)
4. **Situation management** – (a) awareness of difficulty in various situations, ability to analyze a communication setting and identify components that contribute to difficulty/breakdown; (b) ability to change or prevent negative components in communication settings
5. **Emotional reactions** – attitudes towards communication handicaps, reactions to breakdown and general personality structure of the individual

lips. Attending is a behavioral skill and can be taught as part of the audiological rehabilitation. In our experience, consistent visual attending to the talker may account for much of the improvement seen in speechreading. The measurement of treatment efficacy obviously must start at these most elemental levels, which can be viewed as the baseline requirement for effective audiological rehabilitation. That is, the provision of appropriate amplification, vision correction, and visual attending behavior is a fundamental requirement of an effective audiological rehabilitation program and the establishment of efficacy in these areas is important. The testing involves real-ear measurements, and aided loudness discomfort testing of the hearing aid and other standard hearing aid fitting procedures as well as the provision of appropriate assistive listening devices and the assessment of corrected vision by vision professionals. These topics are considered at length elsewhere, and are not discussed here. (See Mueller, Hawkins, & Northern, 1992, for a review of real ear measurements.) Assessment of visual attending behavior on the other hand is more subjective and less easily performed. However, a series of experiments is under way at the University of Illinois that involve physically tracking eye-gaze of speechreaders (Lansing, 1992). Perhaps the techniques developed there will allow assessment of visual attending strategies and consistency.

Sensory Processing Capability

The second component of the communication process is less clearly defined. Although this component, where the basic quality of the signal-processing in the sensory system is considered, may be conceptually clear, in practice it is difficult to isolate purely sensory contributions from the processing of the incoming signals by the brain. Presumably such non-linguistic, psychophysical tasks as gap detection, measurement of difference limen for frequency, determination of the extent of forward masking, and generation of psychophysical tuning curves in general, yield relatively pure measurements of specific sensory capabilities. However, such measurements, although they may be highly reliable and quite precise in practiced subjects, have failed to achieve acceptance as practical measures of the impaired auditory (or visual) system for rehabilitation purposes. It may well be possible to develop psychoacoustic tests (or their visual analogs) with prognostic value or to detect subtle changes in sensory capacity perhaps with physiological indices like otoacoustic emission levels, but the application to treatment efficacy in audiological rehabilitation will no doubt remain difficult.

Speech Recognition Behavior

Monosyllabic word tests. In this section traditional monosyllabic and spondaic word tests are briefly reviewed (for a thorough discussion and criticism, see Elkins, 1984), but more time is spent on newer innovative techniques that in their diverse ways seek to estimate the client's real-world performance or at least try to provide reliable measures of sentence and continuous discourse processing. Obviously, the most important component of communication breakdown due to hearing loss is the residual ability to understand speech through the auditory system and the extent to which that ability can be complemented by the visual channel. Given appropriate amplification/vision correction and visual attending behavior (Component 1) and an impaired but stable sensory system (Component 2), it is possible to assess the client's ability to recognize speech under a variety of conditions. Further, assuming no short-term change in amplification/vision correction and sensory processing, it would then be possible to consider measures of speech recognition as potential indicators of treatment efficacy.

It is well to keep in mind, however, that benefit was defined above as a change in everyday communicative effectiveness due to treatment. If improvement is largely a function of more effective client behavior and control of emotions, however, then one would expect little improvement in raw word recognition ability. Indeed, it is hard to imagine why word recognition would improve, beyond the changes attributed to better attending behavior, more willingness to guess, and general test sophistication and knowledge of materials. Alternatively, one could argue that the client has learned to recognize or differentiate subtle auditory cues through training. Furthermore, better attending and willingness to guess, at least, are reasonable explanations for improved word recognition (Van Tasell & Hawkins, 1981).

The exact reasons for better average performance upon retest following training remain to be sorted out. Several studies have shown modest improvement in monosyllabic word or consonant-vowel nonsense syllable recognition through short-term training (Black, O'Reilly, & Peck, 1963; Dodd, Plant, & Gregory, 1989; Massaro, Cohen, & Gesi, 1993; Montgomery, Walden, Schwartz, & Prosek, 1984; Rubinstein & Boothroyd, 1987; Walden, Erdman, Montgomery, Schwartz, & Prosek, 1981; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). Interestingly, the improvement seems to take place very rapidly with most occurring in the first two hours of treatment (Walden et al., 1977). This observation is consistent with the attending/guessing explanation, but does not prove it. The exact nature of the improvement in raw word recognition that occurs in intensive training and whether it is temporary or permanent are important theoretical questions that only a large scale study with several control groups could answer. Such a study can be done, but one must question its clinical value because there is no evidence that modest improvement in word recognition ability, even if attainable through treatment, is related to meaningful improvement in understanding continuous discourse (Giolas & Epstein, 1963) or to communicating effectively in the client's daily living. Walden et al. (1981) showed that improved auditory and visual discrimination of nonsense syllables did correlate with improvements in sentence understanding. This is encouraging, but still a long way from the elusive "real world."

The need for functional measures is illustrated by two successful examples from the University of South Carolina Audiological Rehabilitation program. The first is an older man with a severe hearing loss and very poor speechreading skills. He communicated primarily with his wife and frequently was unable to understand words that she said, especially proper names and new topics, despite numerous repetitions and rephrasings. This problem was largely solved by teaching her to fingerspell (and him to understand fingerspelling), so that the first few letters of a difficult word could be fingerspelled and the frustrating verbal repetitions could be avoided. The second example is an intelligent but somewhat shy man who lived by himself in a nursing home. He communicated adequately in a one-to-one situation but missed much of the conversation at the dinner table which was one of the few social situations in his day. The dinner table difficulty was resolved to a noticeable extent by his being able to admit and discuss his hearing loss with the other diners. Now they will sometimes make an effort to speak more loudly and clearly to him, but more importantly, he now feels comfortable to ask for repetitions and openly employ other communication strategies. Obviously, neither of these men's improvements in daily communication would be reflected in higher auditory or auditory-visual speech recognition scores in the clinic.

Key words in sentences. Many studies have employed sentences in testing speech recognition performance in auditory, visual, and auditory-visual modes. These studies have sometimes scored whole sentences as correct or incorrect (an inefficient practice calling for hundreds of sentences to avoid the unreliability of short "yes/no" tests, Thornton & Raffin, 1978). More commonly, the scoring involved the percentage of key words correctly received (Dancer, Davis, & O'Neil, 1987; Sims, 1982). Typically, the Central Institute for the Deaf (CID) sentences (Davis & Silverman, 1978) with an average of five key words/sentence are used in lists of 10 sentences (50 key words) each. Even 50 yes/no items is not enough to insure high interlist equivalency, however, and it may be necessary to administer two or more lists. Demorest and Bernstein (1992), for example, found that approximately 40 sentences (several hundred words) were necessary to yield a generalizability coefficient of 0.90. The need for a large number of items was also seen in the development of the Connected Speech Test (CST: version 2) by Cox, Alexander, Gilmore, and Pusakulich (1988). Cox and her colleagues developed a set of 48 paragraphs for measuring continuous discourse intelligibility and found it necessary to combine scores from four paragraphs (a total of 100 key words) to achieve adequate comparability.

Middelweerd and Plomp (1987) used key word scoring in lists of Dutch sentences presented in auditory (A) and auditory-visual (A-V) modes but followed an adaptive threshold procedure (adjusting presentation level a few dB after a set number of words) to obtain a speech reception threshold in dB SPL. This allowed the derivation of A-V minus A to represent the contribution of speechreading in dB. This measure of speechreading is theoretically motivated (Plomp, 1986) and it avoids some of the problems of percent correct scores, but it also has the possibility of low reliability associated with a difference score (see below) so its utility is not yet established.

Estimating intelligibility of connected discourse. Connected discourse would seem to represent everyday communication better than unrelated sentences. The problem is how to derive quantitative measures from connected discourse that represent intelligibility or understanding. One can score key words as in Cox et al. (1988), but this has the problems of yes/no scoring. Asking questions about a passage afterwards puts a load on recent memory and necessitates yes/no scoring also. Obtaining and scoring a verbal summary is equally unreliable. Several investigators have tried to solve this problem by asking the client to estimate the intelligibility of the discourse (Hygge, Ronnberg, Larsby, & Arlinger, 1992; Speaks, Parker, Harris, & Kuhl, 1972). Hawkins, Montgomery, Mueller, and Sedge (1988) generated performance-intensity (P-I) functions in a reverberant setting by varying background noise and asking for "percent words understood" in A and A-V conditions involving a female talker reading a long passage. Again, the effects of hearing loss and speechreading were available in dB difference scores, which allows at least the possibility of relating the performance to models of hearing impairment (Plomp, 1978, 1986).

Tracking task. Communication typically involves continual interaction between talker and listener – an element that is missing from other outcome measures discussed in this section. The tracking task was developed by De Filippo (De Filippo, 1988; De Filippo & Scott, 1978) as a way to evaluate communication of connected discourse in an interactive manner. The measure of communication in the tracking task is the number of words of discourse per minute that a talker can convey to a listener (as verified by repetition). The process is highly interactive with the talker signaling errors, repeating, rephrasing, and generally falling back on a hierarchy of repair strategies depending on the specific responses and errors made by the listener. Talkers differ in their skill, of course, and some talker-listener familiarization occurs, but the tracking task in the hands of the experienced investigator is a valuable technique for the ongoing comparison of communication channels, such as tactile-only versus tactile and speechreading (De Filippo, 1984) or for evaluating cochlear implants (Punch, Robbins, Myers, Pope, & Miyamoto, 1987). Critiques (Tye-Murray & Tyler, 1988) and modifications of the tracking task (Danz & Binnie, 1983; Gagné, Dinon, & Parsons, 1991) have been made (De Filippo, 1988).

Situation Management

One of the most important developments in adult audiological rehabilitation has been the renewed interest in conversation repair strategies (Erber, 1988, 1993). One element in difficult listening situations that is often under the client's control is the way in which the client attempts to gain clarification when misunderstanding occurs. These attempts take many forms, such as saying "what?," asking for synonyms, requesting a written response, etc. Several studies have addressed treatment efficacy issues in conversation repair. In these studies the effectiveness of using repair strategies was evaluated (Gagné & Wyllie, 1989; Tye-Murray, Purdy, Woodworth, & Tyler, 1990) and the clients' ability to learn various strategies was assessed (Tye-Murray, 1991). All studies used sentences or words on videotape or videodisc and evaluated a variety of repair strategies. In general, strategies increased speech recognition and clients were able to learn and use improved strategies in a clinical-research situation. These results are encouraging and represent some of the first attempts to change conversational behavior in a controlled manner.

Emotional Reactions

Self-assessment inventories. One of the most recent developments in establishing treatment efficacy in Audiological Rehabilitation is the availability of psychometrically-sound self-assessment instruments for adults with a hearing impairment. In none of the preceding outcome measures are the clients asked how they felt about a situation or how they react to frustration or how well they communicate in a particular environment. Given the importance of evaluating performance and reactions of clients in typical communication situations, and given the extreme difficulty of actually observing them in these situations, it is possible that the next best thing is to ask clients about how they feel and how they perform. (See Baer, 1990, for an impassioned argument that asking clients is the best way to assess treatment efficacy.)

Several inventories are currently in use, including the Hearing Handicap Scale (High, Fairbanks, & Glorig, 1964), the Hearing Measurement Scale (Noble & Atherley, 1970), the Hearing Handicap Inventory for the Elderly (HHIE: Ventry & Weinstein, 1982), the Hearing Performance Inventory (Giolas, Owens, Lamb, & Schubert, 1979), the Communication Profile for the Hearing Impaired (CPHI: Demorest & Erdman, 1986, 1987, 1988, 1989a, 1989b). In these inventories the client is typically asked to read a statement about a communication situation and respond in a prescribed way with a rating of how he or she feels or functions in that situation. The questionnaire consists of a large number of such items and subtests or scales may be formed by combining the responses to items that sample similar situations or feelings. It is encouraging to see the inventories being used in efficacy research (Abrams, Hnath-Chisolm, Guerreiro, & Ritterman, 1992; Kricos, Holmes, & Doyle, 1993; Smaldino & Smaldino, 1988).

If self-assessment inventories become established as good outcome measures for determining treatment efficacy, it will be due in part to the ways in which they are constructed. Ideally, constructs are developed and a content domain is delineated and sampled. Alternatively, a group of experienced clinicians sits down and develops a large pool of items that reflect the content domains of interest, such as reactions to communication frustration or hearing aid use. Then a long process of item refinement and pilot evaluations takes place which results eventually in a test that has good psychometric properties. The process of test development is very technical and tedious and involves large numbers of clients. (See Demorest & Walden, 1984, for a good introduction to the mechanics of test construction, and Erdman [Chapter 4, present monograph] for a current review.)

All the inventories listed above seem to have acceptable test properties, making them valuable contributions to the treatment efficacy armamentarium. Interestingly, the inventories generally do not correlate highly with the audiogram or speech recognition scores, although a wide range of correlations has been reported (see Demorest & Erdman, 1989b, for more detail). For example, in the Hearing Performance Inventory (Demorest & Walden, 1984) correlations between hearing loss and scale scores range from near zero to 0.39. This is not unexpected because, in our opinion, adjustment and performance in everyday communication depends largely on psychological and behavioral factors and may not relate strongly to details of the hearing loss. That is, low correlations are what one would expect if clients with similar audiograms reacted differently to difficult listening situations and the scales were good indicators of those differences. Of course, low correlations between hearing loss and inventory scores could also reflect error in the measurement of either the loss or the inventory's constructs, so a less optimistic interpretation is also possible.

Although the inventories may represent our best estimates to date of the clients' performance, reactions, and feelings, in day-to-day communication, the inventories are far from perfect. Their content does not yet tap many aspects of interactive communication, they are based on no accepted model of communication breakdown, they are not designed to assess change *per se* and may do it ineffectively (without a model of impaired communication, it is difficult to know). Also, the use of inventories assumes literacy and motivation on the part of the client.

STATISTICAL ISSUES IN MEASURING CHANGE

Psychometric Problems

Problems involving measurement of change stem from the difficulty in modelling the components that contribute to a person's score on a test. The first and most widely used model comes from classical test theory. Classical test theory arose in the first half of this century as a response to the needs for tests, such as intelligence tests and army induction procedures. (See Allen & Yen, 1979, for a good introduction to the topic.) It is likely that every major psychological test involving large numbers of yes/no or multiple-choice items used classical test theory in its development. Classical test theory is thus a powerful and (conceptually, at least) simple model of human test behavior. In classical test theory a subject's score on a test is conceived as

$$X = X_T + e_x, \text{ and} \quad (1)$$

$$Y = X_T + G_T + e_y, \quad (2)$$

where X and Y are pre- and post-test scores, respectively, X_T is the unobservable true score, G_T is the gain (or loss) due to training, and e_x and e_y are normally distributed, uncorrelated error with means of zero. That is, each subject has some different, unknown error associated with his or her score each time the test is taken. Across a group of subjects these errors are assumed to be normally distributed and to be uncorrelated with other components of the model.

The gain, or change, due to the intervening treatment is thus clearly modelled as the G_T term. Classical test theory has difficulty, however, when it comes to actually measuring that gain because the raw scores are the only numbers that can be manipulated. The change in a subject's score due to the treatment is conceptualized as a difference score, D, obtained by subtracting the pre-training score from the post-training score:

$$D = Y - X = X_T + G_T + e_y - X_T - e_x, \quad (3)$$

$$D = G_T + (e_y - e_x). \quad (4)$$

Here, one can already see a hint of the difficulties to come when one considers what happens in a group of subjects. For example, there are now two sources of error (e_y and $-e_x$) in the difference score for each subject. Now e_x and e_y are normally distributed around zero across the subjects, so they will each be positive approximately half the time and negative half the time (if one could go down the list of subjects and actually see the errors separately). This means that the quantity ($e_y - e_x$) will be a larger positive or negative number in approximately 50% of the subjects' difference scores (where the signs of e_y and $-e_x$ are the same) and a smaller (absolute value) number when the signs are different. Thus the error variance of the difference score in the group is larger and the reliability of D will be lower than would be expected based on the test-retest reliabilities of X and Y.

Next consider the correlation between the pre- and post-test scores, r_{xy}. First, the correlation will tend to be high because the true pretest component (X_T) is in both scores. Further, one would expect that in audiological rehabilitation there would be a fairly high correlation between the two tests because our clients differ considerably in most of the skills that our tests are likely to be testing and are fairly consistent. That is, a client who scores 75 initially might receive an 83 on the posttest, while a person who received 35 initially could score 41 with training. Across a group of subjects this would result in a high pretest-posttest correlation, perhaps 0.80. However, it can be shown that if the pretest-posttest correlation (r_{xy}) is high, then the reliability of D (r_{DD}) again goes down. For example, theoretically if we hold the reliability of X and Y constant at +0.80 and somehow vary r_{xy} from 0.00 to 0.80, r_{DD} goes down from 0.80 to 0.00. The tendency for the reliability of the difference score to be low is a problem in efficacy research to the extent that one is concerned with individual differences in change across time. It is entirely possible to have useful measures of group change even though the individual difference scores are not reproducible upon replication. In general the seriousness of the problems with the raw difference score depends to a large extent on the magnitude of the error relative to the effect size of the treatment and on the focus of the research (group vs. individual change).

Finally, there is some uncertainty in interpreting results when the amount of change in two groups is being compared if the mean initial scores differ. Analysis of covariance, if used carefully, may reduce the problem to a manageable extent. These and other issues are discussed in Harris (1963), especially chapters by Bereiter (1963) and Lord (1963), and in Embretson (1991). Other examples drawn from rehabilitative audiology may be found in Montgomery and Demorest (1988).

Another frustration encountered in classical test theory is the fact that estimates of the subject's performance level are biased by how difficult or easy the test items are for the subject. As Embretson (1991) states:

> In classical test theory, the relative difference in ability between two persons or the same person on two occasions, depends on whether the test items are easy, moderate or difficult. For example, two abilities that are measured by total score may seem quite different on a set of moderately difficult items, but be virtually indistinguishable on very easy or very hard items. (p. 190)

(Also see Fischer [1976] and Embretson [1991, p. 195] for good examples of this effect.)

In general it is tempting to conclude that the classical test theory model is not ideal for the measurement of change in human behavior. As a result, several newer models have been proposed which conceptualize the components of the raw score very differently. (Collins & Horn [1991] provide an excellent review of current thinking.) Many of these models stem from item response theory (IRT: Lord, 1980). In general the models and computational procedures are complicated and data intensive, but offer promise of several benefits: (a) avoiding the low reliability of the difference score, (b) recasting the unidimensional classical test theory model into multidimensional form, (c) controlling for initial ability differences, and (d) avoiding the bias due to item difficulty.

Whether these promises will be realized and what experimental effort will be required are not apparent at present. All the models seem to be in continual development (Collins & Horn, 1991) so it is far too early to tell if the IRT models will be useful in evaluating treatment efficacy. They have not been applied to measurement problems in communication disorders and have only seen limited use in psychology in general.

Comparison of Improvement Scores

If one ignores the theoretical issues raised above, comparison of improvement scores is still difficult, even in the simple case of a percent-correct score on a speech recognition test. Consider, for example, a woman with a hearing impairment who acquires good visual attending and guessing strategies for speechreading and improves from an initial 30% to 45% on an auditory-visual integration test of speech recognition. She may be considered more successful than a young man who improves from 80% to 90% on the same test, even though the man's performance level was considerably higher and he may function well in most situations.

However, comparing her improvement (15%) with his (10%) is highly questionable because it assumes that all items are equal in their importance to the communication skill being tested (not that they are equal in difficulty, of course, or people would tend to get them all right or all wrong). Serious difficulties arise at the ends of the scale. Comparing improvements directly implies that improvement from 0% to 10% is equal in its implied impact on everyday communication to a change from 50% to 60% or from 90% to 100%. One could argue that changes in both the very high and very low scores are not as important as similar amounts of change in the middle range because the extremes are not sensitive to improvement in the client's ability to understand running speech (Hawkins et al., 1988). (A 10% recognition score still predicts near zero running speech intelligibility and 90% already yields 100% running speech intelligibility.) This is a problem with the *interpretation* of the test because the test may not reflect everyday speech recognition very well in the high or low performance ranges. From a different point of view, one could argue that changes in the extreme low and high scores are not comparable to similar changes in the middle of the performance range because of the "s" shape of the typical performance-intensity (P-I) function which undoubtedly underlies our test. That is, in the P-I function, large increments in the intensity of the speech are necessary to increase performance from 0%-10% and from 90%-100%, whereas much smaller increments are needed for 10% improvement in the middle of the function.

Another method for comparing improvement in percent correct scores has been used (Walden et al., 1981), which can be called relative improvement. Consider the young man who went from 80% to 90% on our test, pre- to post-treatment. One could argue that this client improved a much larger percentage of the *remaining possible improvement.* That is, if the pre-treatment score is 80%, only 20% change is possible and even though the subject only improved

10%, that percent was one-half the possible improvement [10/(100 − 80)] × 100 = 50% relative improvement. The woman, on the other hand, showed only [15/(100 − 30)] × 100 = 21.4% relative improvement. Even this relative improvement approach has some drawbacks. First of all, it assumes that the score of 100% is attainable and represents some realistic end point of whatever communication skill is being tested. For example, if even normal hearing subjects fail to score above 90% correct, then it is not valid to consider 100% as the basis for the remaining possible improvement. Also, while calculating relative improvement, it is possible to obtain decrements in performance following therapy as well as increases. Suppose our client dropped from 80% to 70%. Has he shown a 50% loss in relative improvement [10/(100 − 80) × 100] or a 12.5% relative loss (compared to the maximum *loss* possible of 80%) [10/(80 − 0)] × 100]?

Alternatively, one can calculate the improvement as a percentage of the initial score. The woman improved from 30% to 45% for a gain, relative to initial performance, of 50% [(45 − 30)/30] × 100 = 50%. This approach favors the low initial scorers just as the relative improvement favors those with high initial scores.

Another transformation involves taking the log of the percent-correct score. Subtracting the log of the pre-training score from the log of the post-training score then yields the log of the relative-improvement score (post/pre). Finally, a *Z* transform of the portion of correct items yields a metric in which equal changes in *Z*-units correspond to equal changes along an underlying trait continuum, where the trait is normally distributed in the population.

The existence of several methods to calculate and compare improvement in scores derived from many-item tests illustrates the fact that, unfortunately, there is no one clear best way to ascribe improvement to differences in many-item test scores of communication performance such as monosyllabic speech recognition. We need a model for the process of (impaired) speech recognition to guide us in our selection. Some promising models of speech intelligibility exist, such as those of Plomp (1978, 1986) and Boothroyd (Boothroyd & Nittrouer, 1988), but they do not as yet directly address the issue of generating or comparing improvement scores. Furthermore, it must be pointed out that all of the above manipulations of percent correct scores assume that there is a true zero point (and 100% point) on the scale. Until more research is done to establish the validity of the inferences drawn from percent correct scores of speech recognition throughout the entire performance range, such manipulations and comparisons are of unknown and potentially dubious value.

Finally, to put this all in perspective, it is important to keep in mind that the psychometric methods discussed above are of limited value to audiological rehabilitationists in the search for treatment efficacy until we have developed measures that allow valid inferences to be drawn of functional communication in the individual with impaired hearing. That is, the finest possible test of speech recognition would tell us nothing about how well the client is able to admit his or her loss and apply repair strategies. The final sections of this chapter address the issue of drawing valid inferences from test scores and present an experimental outcome measure that attempts to overcome some of the difficulties of assessment of improvement in impaired communication in functional settings.

THE PROBLEM SOLVING APPROACH

In our group rehabilitation program at The University of South Carolina, we have adopted a problem solving approach (PSA) to audiological rehabilitation that leads to outcome assessment. In PSA a great deal of time is spent with the client in group and individual therapy listing and defining the specific situations where communication breakdowns occur in his or her daily life. Communication breakdown is carefully defined for the client and refers not only to his or her difficulty in understanding speech but to the specific reasons for the difficulty and to the many ways in which communication and especially conversations can be less than optimal. For example, the client may not use the pragmatics of the conversation to his or her advantage and may miss out on effective turn-taking or change the topic inappropriately. The client may not be aware of the new topic of a conversation ("the recent hurricane") and be unable to use that knowledge to fill in information that is missed. The client may overreact to the occurrence of minor communication difficulty and withdraw mentally (stop trying), get angry, or adopt other maladaptive behaviors which then become the primary source of communication breakdown. Thus both the dynamics of the situation and the sources of difficulty are analyzed in detail.

In practice, in the PSA the client keeps a daily log and the specific problems or difficult communication situations encountered are carefully listed and the reasons behind the breakdown in each case are analyzed. Thus the client engages in structured and systematic self-observation or monitoring. The observations are reported at least weekly and are analyzed and clarified as needed by the group leader. Spouses, relatives, and caregivers contribute to the list, as well. The list is then "attacked" with the typical audiological rehabilitation methods such as assertive listening, spouse training, speechreading, assistive listening devices, repair strategies, self-assessment inventories, and counseling (see Montgomery, 1993, or Erber, 1988, 1993, for more details), and at the end of a specified time, usually coinciding with the end of our 6-week adult group, an assessment is made of the overall percentage of problem occurrences per week that have been resolved. The overall percentage is based on the success in each difficult situation, weighted by the frequency of the situation. This figure may be broken down into types of problems if desired (e.g., 54% of telephone breakdowns were avoided, 100% of television listening difficulty was eliminated, 20% of 3-party conversations were improved). The numbers thus represent, *by definition,* a measure of the extent to which the client's communication difficulty has been resolved. The PSA would appear to allow more valid inferences to be drawn from it than an improvement score on a speech recognition test and may compare favorably with the self-assessment inventories, but the possibility of valid inferences is achieved at the cost of a great deal of clinician time and client effort. Its potential strength, we believe, lies in its totally individualized nature. It is tailored to the individual's performance in the everyday communication environment and reflects his or her specific needs, values, and goals.

The PSA has many other benefits from a strictly clinical point of view. It provides the client with lifelong skills for analyzing the reasons for communication breakdown in new situations encountered after therapy is terminated, it readily accommodates spouses and caregivers, and it adapts well to either group or individual audiological rehabilitation. Most importantly, the PSA empowers the client to take responsibility for his or her rehabilitation. The PSA is thus clearly within the domain of the rehabilitation model as opposed to the medical model (see Erdman, 1993, p. 385 for further discussion of these models). Note that the PSA is in its infancy. It yields essentially a zero-based difference score and thus has some of the weaknesses mentioned earlier concerning that type of score as well as other more obvious difficulties of accurate reporting. The PSA is mentioned here only as an example of a functional outcome measure, reflecting reported improvement in satisfaction in difficult everyday situations – a suggested direction for future development of outcome measures in adult audiological rehabilitation.

Our original definition of the goal of audiological rehabilitation for the adult emphasized improving communication ability in the everyday life of clients who have a hearing impairment. PSA clearly reflects that definition. To the extent that clients are capable of observing situations (and themselves in situations) and are aware of communication breakdown and difficulty, PSA is a measure of the success or failure they experience.

RECOMMENDATIONS

In this chapter we have considered the assessment of treatment efficacy in the rehabilitation of the adult with impaired hearing. Several problems in determining the efficacy of our clinical procedures were considered, including the lack of a model of functional communication, the lack of outcome measures that reflect that model, and the lack of agreement on specific diagnostic categories. Statistical difficulties in the generation and interpretation of outcome measures were also discussed.

The problems that hinder or discourage efficacy and efficiency research in audiological rehabilitation, however, are indicators of the directions in which that research should take. First, the major premise of this chapter is that efficacy and efficiency measurements in audiological rehabilitation should be based on outcome measures that reflect communication performance in everyday listening and speaking situations. If one accepts that premise, then the primary focus of research should be the development of a model of interactive verbal communication where one (or more) of the interlocutors has reduced auditory processing capacity. (See Alpiner et al., 1993, p. 6 for an example of a model with an interactive element.) This model would lead, in turn, to the development of outcome measures that reflect the parameters of the model and allow the measurement of various aspects of interactive communication. Second, there is a need for refined diagnostic categories that go beyond the audiogram to include the effects of general factors such as mental state, age, and mobility, and specific factors like assertiveness, pragmatic competence, and reactons to communication breakdown. Finally, if progress is made along these fronts, then efforts should be directed toward large-scale studies of the efficacy of existing and new clinical procedures using the outcome measures and diagnostic groups that are being developed. Thus, if we as a profession can reach a consensus on the goals of audiological rehabilitation for the adult, and can agree on a model of successful verbal communication, there exists great potential for establishing more firmly the efficacy of our clinical practices.

REFERENCES

Abrams, H.B., Hnath-Chisolm, T., Guerreiro, S.M., & Ritterman, S.I. (1992). The effects of intervention strategy on self-perception of hearing handicap. *Ear and Hearing, 13,* 371-376.

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Alpiner, J., Kaufman, K., & Hanavan, P. (1993). Overview of rehabilitative audiology. In J. Alpiner & P. McCarthy (Eds.), *Rehabilitative audiology: Children and adults* (2nd ed.) (pp. 3-16). Baltimore: Williams & Wilkins.

ASHA. (1992, November). *Models for the evaluation of treatment efficacy.* Conference, San Antonio, TX.

Baer, D.M. (1990). The critical issue in treatment efficacy is knowing why treatment was applied: A student's response to Roger Ingham. In L.B. Olswang, C.K. Thompson, S.F. Warren, & N.J. Minghetti (Eds)., *Treatment efficacy research in communication disorders* (pp. 31-40). Rockville, MD: American Speech-Language-Hearing Foundation.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison, WI: The University of Wisconsin Press.

Black, J.W., O'Reilly, P.P., & Peck, L. (1963). Self-administered training in lipreading. *Journal of Speech and Hearing Disorders, 28,* 183-186.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America, 84,* 101-114.

Collins, L.M., & Horn, J.L. (1991). *Best methods for the analysis of change.* Washington, DC: American Psychological Association.

Cox, R.M., Alexander, G.C., Gilmore, C., & Pusakulich, K.M. (1988). Use of the Connected Speech Test (CST) with hearing-impaired listeners. *Ear and Hearing, 9,* 198-207.

Dancer, J.E., Davis, P.N., & O'Neil, J. (1987). Word recognition versus sentence comprehension on a speech reading task. *American Annals of the Deaf, 132,* 43-45.

Danz, A.D., & Binnie, C.A. (1983). Quantification of the effects of training the auditory-visual reception of connected speech. *Ear and Hearing, 4,* 146-151.

Davis, H., & Silverman, S.R. (1978). *Hearing and deafness* (4th ed.). New York: Holt, Rinehart and Winston.

De Filippo, C.L. (1984). Laboratory projects in tactile aids to lipreading. *Ear and Hearing, 5,* 211-227.

De Filippo, C.L. (1988). Tracking for speechreading training. In C.L. De Filippo & D.G. Sims (Eds.), *New reflections on speechreading* [Monograph]. *The Volta Review, 90,* 215-237.

De Filippo, C.L., & Scott, B.L. (1978). A method for training and evaluating the reception of ongoing speech. *Journal of the Acoustical Society of America, 63,* 1186-1192.

Demorest, M.E., & Bernstein, L. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research, 35,* 876-892.

Demorest, M.E., & Erdman, S.A. (1986). Scale composition and item analysis of the Communication Profile for the Hearing Impaired. *Journal of Speech and Hearing Research, 29,* 515-535.

Demorest, M.E., & Erdman, S.A. (1987). Development of the Communication Profile for the Hearing Impaired. *Journal of Speech and Hearing Disorders, 52,* 129-143.

Demorest, M.E., & Erdman, S.A. (1988). Retest stability of the Communication Profile for the Hearing Impaired. *Ear and Hearing, 9,* 237-242.

Demorest, M.E., & Erdman, S.A. (1989a). Factor structure of the Communication Profile for the Hearing Impaired. *Journal of Speech and Hearing Disorders, 54,* 541-548.

Demorest, M.E., & Erdman, S.A. (1989b). Relationships among behavioral, environmental, and affective communication variables: A canonical analysis of the CPHI. *Journal of Speech and Hearing Disorders, 54,* 180-188.

Demorest, M.E., & Walden, B.E. (1984). Psychometric principles in the selection, interpretation, and evaluation of communication self-assessment inventories. *Journal of Speech and Hearing Disorders, 49,* 226-240.

Dodd, B., Plant, G., & Gregory, M. (1989). Teaching lip-reading: The efficacy of lessons on video. *British Journal of Audiology, 23,* 229-238.

Elkins, E. (Ed.). (1984). *Speech recognition by the hearing impaired.* Rockville, MD: American Speech-Language-Hearing Association.

Embretson, S.E. (1991). Implications of a multidimensional latent trait model for measuring change. In L.M. Collins & J.L. Horn (Eds.), *Best methods for the analysis of change* (pp. 184-197). Washington, DC: American Psychological Association.

Erber, N. (1988). *Communication therapy for hearing-impaired adults.* Abbotsford, Victoria, Australia: Clavis Publishing.

Erber, N. (1993). *Communication and adult hearing loss.* Melbourne: Clavis Publishing.

Erdman, S. (1993). Counseling hearing-impaired adults. In J. Alpiner & P. McCarthy (Eds.), *Rehabilitative audiology: Children and adults* (2nd ed.) (pp. 374-416). Baltimore: Williams and Wilkins.

Erdman, S.A. (1994). From research focus to research tool. In J.-P. Gagné & N. Tye-Murray (Eds.), *Research in audiological rehabilitation: Current trends and future directions* [Monograph]. *Journal of the Academy of Rehabilitative Audiology, 27,* 67-90.

Fischer, G.H. (1976). Some probabilistic models for measuring change. In D.N.M. De Gruijter & L.J.T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97-110). New York: Wiley.

Frattali, C. (1992). Measuring functional communication: A research project to establish the reliability and validity of a functional communication measure for adults. *National Institute on Disability and Rehabilitation Research grant to the American Speech-Language-Hearing Association.* Announced in *Asha, 34*(8), 16.

Gagné, J.-P., Dinon, D., & Parsons, J. (1991). An evaluation of CAST: A Computer-Aided Speechreading Training program. *Journal of Speech and Hearing Research, 34,* 213-221.

Gagné, J.-P., & Wyllie, K.A. (1989). Relative effectiveness of three repair strategies on the visual-identification of misperceived words. *Ear and Hearing, 10,* 368-374.

Giolas, T.G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6,* 349-358.

Giolas, T.G., Owens, E., Lamb, S.H., & Schubert, E.D. (1979). Hearing Performance Inventory. *Journal of Speech and Hearing Disorders, 44,* 169-195.

Harris, C.W. (1963). *Problems in measuring change.* Madison, WI: The University of Wisconsin Press.

Hawkins, D.B., Montgomery, A.A., Mueller, H.G., & Sedge, R.K. (1988). Assessment of speech intelligibility by hearing-impaired listeners. In B. Berglund, U. Berglund, J. Karlsson, & T.

Lindvall (Eds.), *Noise as a public health problem: Vol. 2. Hearing, communication, sleep and nonauditory physiological effects* (pp. 241-246). Stockholm: Swedish Council for Building Research.

High, W.S., Fairbanks, G., & Glorig, A. (1964). Scale for self-assessment of hearing handicap. *Journal of Speech and Hearing Disorders, 29,* 215-230.

Hygge, S., Ronnberg, J., Larsby, B., & Arlinger, S. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research, 35,* 208-215.

Kricos, P., Holmes, A., & Doyle, D. (1993). Efficacy of a communication training program for hearing-impaired elderly adults. *Journal of the Academy of Rehabilitative Audiology, 25,* 69-80.

Lansing, C.R. (1992). Patterns of eye gaze during speechreading: Initial results [Abstract]. *Asha, 34*(10), 235.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge: MIT Press.

Lord, F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.), *Problems in measuring change* (pp. 21-38). Madison, WI: The University of Wisconsin Press.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D., Cohen, M., & Gesi, A. (1993). Long-term training, transfer, and retention in learning to lipread. *Perception and Psychophysics, 53,* 549-562.

Middelweerd, M.J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82,* 2145-2147.

Montgomery, A.A. (1993). Management of the hearing-impaired adult. In J. Alpiner & P. McCarthy (Eds.), *Rehabilitative audiology: Children and adults* (2nd ed.) (pp. 311-330). Baltimore: Williams and Wilkins.

Montgomery, A.A., & Demorest, M.E. (1988). Issues and developments in the evaluation of speechreading. In C.L. De Filippo & D.G. Sims (Eds.), *New reflections on speechreading* [Monograph]. *The Volta Review, 90*(5), 193-214.

Montgomery, A.A., Walden, B.E., Schwartz, D.M., & Prosek, R.A. (1984). Training auditory-visual speech reception in adults with moderate sensorineural hearing loss. *Ear and Hearing, 5,* 30-36.

Mueller, H.G., Hawkins, D.B., & Northern, J. (1992). *Probe microphone measurements: Hearing aid selection and assessment.* San Diego, CA: Singular Publishing Group.

Noble, W.G., & Atherley, G.R.C. (1970). The Hearing Measure Scale: A questionnaire for the assessment of auditory disability. *Journal of Auditory Research, 10,* 229-250.

Office of Technology Assessment. (1978). *Assessing the efficacy and safety of medical technologies.* Washington, DC: Congress of the United States, Office of Technology Assessment, OTA-11-75.

Olswang, L.B., Thompson, C.K., Warren, S.F., & Minghetti, N.J. (1990). *Treatment efficacy research in communication disorders.* Rockville, MD: American Speech-Language-Hearing Foundation.

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *Journal of the Acoustical Society of America, 63,* 533-549.

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech and Hearing Research, 29,* 146-154.

Punch, J.L., Robbins, A.M., Myers, W.A., Pope, M.L., & Miyamoto, R.T. (1987). Relationships among selected measures of single-channel cochlear implant performance. *Ear and Hearing, 8,* 37-43.

Rubinstein, A., & Boothroyd, A. (1987). Effect of two approaches to auditory training on speech recognition by hearing-impaired adults. *Journal of Speech and Hearing Research, 30,* 153-160.

Sims, D.G. (1982). Hearing and speechreading evaluation for the deaf adult. In D.G. Sims, G.G. Walter, & R.L. Whitehead (Eds.), *Deafness and communication assessment and training.* Baltimore: Williams and Wilkins.

Smaldino, S.E., & Smaldino, J.J. (1988). The influence of aural rehabilitation and cognitive style disclosure on the perception of hearing handicap. *Journal of the Academy of Rehabilitative Au-*

diology, 21, 54-67.

Speaks, C., Parker, B., Harris, B., & Kuhl, P. (1972). Intelligibility of connected discourse. *Journal of Speech and Hearing Research, 15,* 590-602.

Thornton, A.R., & Raffin, M.J.M. (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research, 21,* 507-518.

Tye-Murray, N. (1991). Repair strategy usage by hearing-impaired adults and changes following communication therapy. *Journal of Speech and Hearing Research, 34,* 921-928.

Tye-Murray, N., Purdy, S., Woodworth, G., & Tyler, R.S. (1990). The effect of repair strategies on the visual identification of sentences. *Journal of Speech and Hearing Disorders, 55,* 621-627.

Tye-Murray, N., & Tyler, R.S. (1988). A critique of continuous discourse tracking as a test procedure. *Journal of Speech and Hearing Disorders, 53,* 226-231.

Van Tasell, D., & Hawkins, D. (1981). Effects of guessing strategy on speechreading test scores. *American Annals of the Deaf, 126,* 840-844.

Ventry, I.M., & Weinstein, B.E. (1982). The Hearing Handicap Inventory for the Elderly: A new tool. *Ear and Hearing, 3,* 129-134.

Walden, B.E., Erdman, S.A., Montgomery, A.A., Schwartz, D.M., & Prosek, R.A. (1981). Some effects of training on speech recognition in hearing-impaired adults. *Journal of Speech and Hearing Research, 24,* 207-216.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130-145.