Questions of Efficacy in Clinical Aphasiology

G. Albyn Davis
University of Massachusetts, Amherst, Massachusetts

The several attempts to investigate efficacy of aphasia treatment have varied from large group studies, to single subject experimental designs, to less robust "pre-experimental" case studies. The most basic question asked by clinical aphasiologists has been "Does aphasia therapy work?" (e.g., Darley, 1979). This particular question is ambiguous, and the many reported experiments are examples of more specific versions of this basic question.

In order to establish where we are with respect to our accomplishments and in order to have a direction for planning further research, it would be helpful to have a guide to the many possibilities in efficacy questions. My purpose today is to discuss this perspective regarding the efficacy issue. Fifty-four published studies were analyzed with respect to 24 categories of questions. Some types of questions have been asked relatively often. Other types, including some of the most important ones, have rarely been addressed in group or single subject research.

EFFICACY QUESTIONS

Many publication titles reflect the form of an explicit efficacy question, as we ask "What is the effect of X on Y?" X stands for an independent variable manipulated by an experimenter, such as hyponsis and imagery training. Y stands for a dependent variable measured by the experimenter, such as a client's naming performance. Thompson, Hall, and Sison (1986) asked about "Effects of hypnosis and imagery training on naming..." Burger and Wertz (1984) examined "The effect of a token training program on auditory comprehension..." More generally, we often ask, as Basso and her colleagues did, about "Effects of speech and language treatment on recovery..." (Basso, Capitani, and Vignolo, 1979). Looking at similar issues, McReynolds at the University of Kansas has trained a few in our midst to ask about the "acquisition, generalization, and maintenance" of several dependent variables (e.g., Thompson and Kearns, 1981).

In attempting to characterize over 50 studies, it appeared to me that treatments (X) could be divided into four broad categories:

(1) We have asked about a variable or component within a treatment strategy, as when Potter and Goodman (1983) wondered what would happen when a patient hears laughter in the background while working on his gestures.
(2) We have been curious about the efficacy of unusual or unique strategies, such as MIT, HELPSS, VAT, and PACE. I have also decided to include gestural facilitation of speech, training of gesturing, and unusual items such as tokens within this category.
(3) We wonder about the value of general or basic methods that include Schuell's stimulation or the operant conditioning of basic language behaviors.
(4) Finally, we often want to compare two different strategies, as when Loverso, Prescott, and Selinger (1985) compared microcomputers and real people as clinicians.

In addition, we want to see our clinical manipulations have an effect on a patient's progress, requiring that the dependent variable (Y) consist of at least two measurements. We especially want to see progress in something other than the stimulus conditions and behaviors we observe during treatment. Clinical investigators measure either the behaviors of treatment or something else as a measure of generalization. Sometimes we measure both, with continuous probing strategies and occasional standardized tests.

Questions about treatment can be differentiated further according to the range of generalization measured in the dependent variable. This is important, because treatment can have an effect during treatment with little progress seen in something else. Appendix A contains a characterization of different "distances" between a stimulus-response condition of treatment and a dependent measure of progress. These distances increase with the number of variables in stimulus and response that distinguish between the measure and the treatment, and they may be considered to reflect a range of value that can be placed on a treatment effect. With respect to standard structured treatment, levels 0 to 3 represent increasingly dissimilar structured tasks, while levels 3 and 4 represent more functional or naturalistic instances of progress. TIME represents generalization at any of the distances to some time after the completion of treatment, often called a "maintenance" measure.

By combining the four types of independent treatment variables with the six distances of dependent variables, we can have fun imagining 24 general types of treatment questions. These possibilities are portrayed in Table 1, making "Does aphasia therapy work?" a rather dull boy. One type of question (e.g., Q-9) might concern the effect of a unique treatment strategy on a different type of verbal behavior in a different stimulus condition. Another question (e.g., Q-17) might address the effect of basic aphasia treatment on spontaneous verbal behaviors in natural circumstances. It is unlikely that questions will be asked about the effect of a specific component of treatment on behaviors that are very different from those elicited in the treatment (i.e., Q-3 through Q-6). Therefore, the matrix actually generates 20 reasonable questions.

Table 1. A matrix indicating types of efficacy questions generated by pairing categories of independent treatment variables (X) with dependent measures of behavior (Y). It is unlikely that certain questions will be studied (*).

Generalization (Probe Distances)		Categories of X: Tx Variable	Tx Unique	Tx General	Tx Comparison
DTx	During Tx				
0	Treated	Q-1	Q-7	Q-13	Q-19
1	One variable	Q-2	Q-8	Q-14	Q-20
2	Mult variables	Q-3*	Q-9	Q-15	Q-21
3	Predictors	Q-4*	Q-10	Q-16	Q-22
4	Natural	Q-5*	Q-11	Q-17	Q-23
TIME		Q-6*	Q-12	Q-18	Q-24

When we think of questions to study, we should ask ourselves whether we have a question that is worth answering, a question that other people might be wondering about, and a question that is answerable.

METHOD

Fifty-four studies were selected for review. These studies addressed an efficacy question and employed measures of progress. Thirty-two were single-subject studies, and about 50% of these were robust experiments in that they followed design rules for maximizing internal validity (e.g., Kazdin, 1982). Twenty-two were group studies with a mixture of small, medium, and large groups. Twenty-seven studies appeared in Clinical Aphasiology Conference Proceedings. Thirty-seven had been available to Bloodstein (1984) when he prepared the recent second edition of his introduction to speech pathology, in which he informs undergraduates of "the rather small amount of research" on the results of aphasia treatment.

The main concern of this study was to determine the frequency with which the 24 possible efficacy questions were asked in the 54 studies. This was facilitated with the use of a data-base management program called Infostar published by MicroPro, Inc. Attributes of each study were filed on forms created for storing these attributes. I was able to retrieve all studies conforming to particular independent and dependent variables. For example, I could also retrieve all studies that were multiple baseline designs across behaviors with Broca's aphasic subjects.

Classifying the independent and dependent variables of these studies was not always straightforward. For example, Loverso's comparison between micro-computers and clinicians was treated as a comparison because of his use of a multiple treatments design, but his study could be thought of as a study of an attribute of a single treatment, with the attribute being +/- human. Deciding on the distance between a measurement probe and a treatment condition was exasperating when the differences between them were not clearly described in the publication. Sometimes I had to make an educated guess. Another difficulty occurred when trained and untrained items were collapsed into a single probe. With single scores from such probes, effects could not be separated. Furthermore, the survey does not address believability of the results according to rules of scientific evidence. It is a survey of questions that have been asked rather than a survey of whether these questions were answered.

RESULTS

The results are summarized in Table 2, and specific frequencies are reported in Appendix B. As noted in the Appendix, subtotals often surpass the total number of studies, because many studies addressed more than one question.

Table 1 indicates that most studies, 39 of 54, involved unique and basic treatments. Also, most of the questions in these studies dealt with their effects on progress in structured conditions of the type that have been used in standard assessment and treatment of language. The multiple variable distance from treatment (i.e., level 2) was measured most often, with 19 instances of pre- and post-test summary scores and the rest being continuous probes. The numbers in Appendix B can be somewhat difficult to interpret, especially regarding level 2. For example, unique treatments were examined in 15 studies, but some of these involved more than one version of the level 2 distance.

Table 2. The relative frequency with which the different general questions have been studied. Several = 10 or more. Some = 6 to 8 times. Few = 1 to 4 times. 0 = none found among the 54 reviewed.

Y Generalization (Probe Distances)		Categories of X: Tx Variable N = 6	Tx Unique N = 18	Tx General N = 21	Tx Compared N = 9
DTx	During Tx				
0	Treated	few	some	some	few
1	One variable	few	several	several	few
2	Mult variables	few	several	several	some
3	Predictors	0	few	few	few
4	Natural	0	few	few	few
TIME		few	few/some	some	few

Relatively little research has focused on specific treatment variables and comparison of treatment strategies. Relatively little research has examined whether treatment affects predictors of functional language or functional language behavior in natural conditions.

GENERAL DISCUSSION

Some of you may be surprised with the small number of investigations pertaining to specific components of treatment. However, it is important to keep in mind that a vast number of "efficacy" studies pertain to questions in which the dependent variable is restricted to a particular moment in time. Such questions arise daily as we wonder if one cue is more stimulable than another cue or if a slowed pace improves comprehension performance. One example is "The effects of cuing on picture naming..." (Pease and Goodglass, 1978). Such studies illustrate basic research designed to extend our understanding of aphasia. Darley (1982) cited these studies as evidence for the value of certain stimulus variations in "maximizing input and output" for patients. Brookshire (1975) asked a similar question, as his title reads "Effects of prompting on spontaneous naming of pictures..." However, he examined progress as his dependent variable; and such studies have been infrequent.

The range of generalization effects that have been determined for our treatments has been relatively "close to the vest." When collapsing the data from levels 0, 1, and 2a, we see that 60% of dependent variables have reflected conditions that were identical to or similar to the conditions of treatment. When we focus on the functional levels of 3 and 4, we see that only 12% of efficacy questions dealt with effects of treatment on predictors or actual observations of naturalistic use of language. There are only four instances of explicit observation of natural communicative performance. Two of these involved an informant rating developed for the VA Cooperative studies (Wertz, Collins, Weiss, et al., 1981; Collins, Wertz, et al., 1983). Another was the nifty "novel social dyad" used by Thompson and Byrne (1984), and the other was simply a dialogue of conversation presented by Helm-Estabrooks, et al. (1981). Usually we ask a family member about how the patient is doing at home. Therefore, we simply do not know very much about the effects of our treatment at levels 3 and 4, unless it can be established that our standard language assessments are valid predictors of naturalistic language ability.

Finally, single subject studies of various strengths outnumber small and large group studies, and the single subject studies are more widely distributed throughout the question matrix. Three expensive large group studies, those of Basso, Wertz, and Shewan, dealt with questions in only three cells of the matrix, namely, with questions 15, 21, and 23 (Basso, et al., 1979; Shewan and Kertesz, 1984; Wertz, et al., 1981). Collins, et al. (1983) recently extended the VA Cooperative study to question 24. Single case experiments have demonstrated their potential for covering a wide range of questions. The development of multiple treatment designs should push this research into the comparison of strategies, which has been neglected somewhat in single case investigation.

In conclusion, I would like to make a few recommendations for clinical practice and experimentation. First, we should not collapse probes; we should keep trained and untrained items distinct so that distances of generalization can be recognized. Second, we should try more 3 and 4 level probes, such as the restaurant probe presented by Warner and Thompson at this conference. Third, we need more component analyses and treatment package comparisons, especially when there are claims that two different methods can achieve the same goal.

REFERENCES

Basso, A., Capitani, E. and Vignolo, L.A. Influence of rehabilitation on language skills in aphasic patients: A controlled study. _Archives of Neurology_, 36, 190-196, 1979.

Bloodstein, O. _Speech Pathology: An Introduction_ (Second Edition). Boston: Houghton Mifflin, 1984.

Brookshire, R.H. Effects of prompting on spontaneous naming of pictures by aphasic subjects. _Human Communication_ , Autumn, 63-71, 1975.

Burger, L.H. and Wertz, R.T. The effect of a token training program on auditory comprehension in a case of Wernicke's aphasia. In R.H. Brookshire (Ed.), _Clinical Aphasiology: Conference Proceedings, 1984_. Minneapolis, MN: BRK Publishers, 173-180, 1984.

Collins, M.J., Wertz, R.T., et al. Coping with success: The maintenance of therapeutic effect in aphasia. In R.H. Brookshire (Ed.), _Clinical Aphasiology: Conference Proceedings, 1983_. Minneapolis, MN: BRK Publishers, 156-163, 1983.

Darley, F.L. Treat or neglect? _Asha_, 21, 628-631, 1979.

Darley, F.L. _Aphasia_. Philadelphia: W.B. Saunders, 1982.

Helm-Estabrooks, N., Fitzpatrick, P.M. and Barresi, B. Response of an agrammatic patient to a syntax stimulation program for aphasia. _Journal of Speech and Hearing Disorders_, 46, 422-427, 1981.

Kazdin, A.E. Single-case Research Designs: Methods for Clinical and Applied Settings. New York: Oxford University Press, 1982.

Loverso, F.L., Prescott, T.E., Selinger, M., Wheeler, K.M. and Smith, R.D. The application of microcomputers for the treatment of aphasic adults. In R.H. Brookshire (Ed.), Clinical Aphasiology: Conference Proceedings, 1985. Minneapolis, MN: BRK Publishers, 189-195, 1985.

Pease, D.M. and Goodglass, H. The effects of cuing on picture naming in aphasia. Cortex, 14, 178-189, 1978.

Potter, R.E. and Goodman, N.J. The implementation of laughter as a therapy facilitator with adult aphasics. Journal of Communication Disorders, 16, 41-48, 1983.

Shewan, C.M. and Kertesz, A. Effects of speech and language treatment on recovery from aphasia. Brain and Language, 23, 272-299, 1984.

Thompson, C.K. and Byrne, M.E. Across setting generalization of social conventions in aphasia: An experimental analysis of "loose training." In R.H. Brookshire (Ed.), Clinical Aphasiology: Conference Proceedings, 1984. Minneapolis, MN: BRK Publishers, 132-144, 1984.

Thompson, C.K., Hall, H.R. and Sison, C.E. Effects of hypnosis and imagery training on naming behavior in aphasia. Brain and Language, 28, 141-153, 1986.

Thompson, C.K. and Kearns, K.P. An experimental analysis of acquisition, generalization, and maintenance of naming behavior in a patient with anomia. In R.H. Brookshire (Ed.), Clinical Aphasiology: Conference Proceedings, 1981. Minneapolis, MN: BRK Publishers, 35-41, 1981.

Wertz, R.T., Collins, M.J., Weiss, D., et al. Veterans Administration cooperative study on aphasia: A comparison of individual and group treatment. Journal of Speech and Hearing Research, 24, 580-594, 1981.

APPENDIX A

Definitions of Y, varying as to levels of "distance" between a treatment and a measurement device. Corresponding to concepts of stimulus and response generalization, a distance would be determined with respect to differences in stimulus and response variables. These variables include "intra-language" variables (e.g., lexicon, syntax, length) and "extra-language"/contextual variables (e.g., person, interaction style, setting).

DTx: During Treatment. When responses are scored during conditions of treatment, as in determining whether a patient meets a task criterion.

0: Probe of Trained Behavior. No generalization measured. Generally considered to be an acquisition probe, when stimuli and behavior of treatment are assessed without special cues and feedback. May be items interspersed among treatment items or a post-session probe.

1: Probe Untrained Behavior/one variable difference. Structured assessment of stimulus/response condition differing by only one variable from the basic condition of the relevant treatment. Usually the same task but different items. Takes into consideration that a simple change in stimulus item (e.g., a picture) brings a change in response with it. Changing items of a naming task is considered to be a "one variable" difference.

2: Multiple Variable Difference. Measurement condition differs by more than one variable from the relevant treatment task. This category is highly differentiated, especially with respect to possible "distances" from the treatment. Some probes are very close to 1. Others might be considered to be closer to 3, depending on strength of evidence (and conviction) that standard aphasia tests are predictive of natural performance.

2a: Probe differs from treatment along two or three variables.

2b: Probe is very different from treatment. May represent focus on certain subtests or specific sections of an aphasia test.

2c: General summary scores from aphasia/language tests, such as the PICA Overall or WAB AQ/CQ. These represent a "mixed bag" with respect to "distance" from treatment.

3: Predictors of natural/function performance. Structured probes, and administered in standard clinical setting, usually by a client's clinician. Include FCP and CADL, and inventions such as role-playing tasks or PACE.

4: Natural Performance. Measures of linguistic/communicative behaviors in natural conditions. Stimulus variables include interaction style (e.g., conversation), person (e.g., spouse), situation (e.g., ordering from menu), and setting (e.g., retaurant).

APPENDIX B

Frequencies with which specific question categories have been asked in 54 studies of aphasia treatment efficacy. Subtotals often surpass totals, because many studies involve multiple dependent variables.

	Y Distances	Categories of X Tx Variable N=6	Tx Unique N=18	Tx General N=21	Tx Comp N=9	
0	Treated	4	6	6	3	19
1	One variable	3	10	11	2	26
2	Mult variable	2	15	12	8	37
2a	Few variable	2	5	3	2	12
b	More variable	0	8	4	0	12
c	Test sums	0	7	6	6	19
3	Predictors	0	3	2	2	7
4	Natural	0	1	1	2	4
	TIME	1	5	6	4	16

DISCUSSION

Q: Would you explain more about TIME as a dependent variable? Why doesn't it fit in any of the levels?

A: Time is a qualitatively different form of generalization. It is intended to reflect maintenance of progress seen at any of the levels. Maintenance at some time after completion of treatment can be seen with respect to any of the stimulus and response generalization distances.

Q: I would like to see you doing more of this research in clinics.

A: I agree. My intention with this paper was to show the need for more research, most of which has been done in VA Medical Centers and university clinics. I think we should be measuring generalization in rehabilitation centers and other clinics. I have been trying on a limited basis so far to get clinicians to help me out with such research. I have found some reluctance, but I have not yet tried all the possibilities. Clinicians have to _want_ to participate in this research.

Q: Audrey Holland observed natural communication by following patients around, and, as I remember, she showed that the CADL correlated quite highly with that validation measure. Also, the Boston, FCP, and the PICA correlated highly with that measure. Do you think that's telling us anything about our standardized tests and the validity of using them as outcome measures?

A: I may be wrong, but I think she checked the CADL and the PICA in correlation with the measure in which she followed people around in natural settings, and she found that they did not correlate very highly. I think the correlations were in the .60s. What was surprising was a high correlation between the CADL and the PICA.

Q: I think you are right that they correlated in the .60s, but it was significant at a .001 level. All of those measures were.

A: Well, there's an issue as to whether or not statistical significance of a correlation would be clinically significant. Holland's method might be a way of developing clinical measures at level 4, but that particular technique has not been used up to this time as measure of progress in an efficacy study.

C: You're right that it was not used at that time as a measure of progress in an efficacy study, but it was a look at whether a person's use of language or ability to communicate out of our sight had any relationship to the way we tend to measure it.