On Drawing the Right Conclusions in Psycholinguistics: Some Critical Areas of Control in Experimental and Analytical Practice

Author(s): Bruce L. Derwing

Proceedings of the Ninth Annual Meeting of the Berkeley Linguistics Society (1983), pp. 64-73

Please see “How to cite” in the online sidebar for full citation information.

Please contact BLS regarding any further use of this work. BLS retains copyright for both print and screen forms of the publication. BLS may be contacted via <http://linguistics.berkeley.edu/bls/>.

The Annual Proceedings of the Berkeley Linguistics Society is published online via [eLanguage](#), the Linguistic Society of America's digital publishing platform.

On Drawing the Right Conclusions in Psycholinguistics: Some Critical Areas of Control in Experimental and Analytical Practice

Bruce L. Derwing
University of Alberta

For the past decade or so I have been involved in a research program concerned with the experimental testing of linguistic rules, particularly in phonology (and morphophonology). Two quite distinct interpretations of the notion "linguistic rule" should perhaps be distinguished at the outset. The first, oldest and most familiar one views a rule as any *regularity* which can either be extracted from or imposed upon a sample of transcribed utterances in some language. This is the notion of rule with which linguists have been, until relatively recently, at least, almost exclusively preoccupied. The second interpretation of the notion of "rule" in linguistics refers to some generalization internalized by speakers in the process of language acquisition and which plays some role in the speaker's use of the language, i.e., in the production and comprehension of at least some of his utterances. It is important to understand that a relation of very uncertain relevance exists between these two fundamentally different notions of "rule," and the reason for this seems obvious enough: the mere discovery (or invention) of a regularity by the linguist with respect to some particular linguistic corpus provides no guarantee whatever that this same regularity has been similarly identified, extracted or invented by the ordinary language learner (i.e., as Kiparsky has put it, "the fact that a generalization can be stated [is not] enough to show that it is psychologically real" (1968, p. 172).¹ Moreover, the proliferation of linguistic "schools" over the past several years, even within the basic framework of generative grammar, has made it increasingly obvious that any linguistic phenomenon can be quite correctly described (i.e., in a way that is in complete accord with the available linguistic facts) in a dizzying multitude of ways, depending on the initial theoretical assumptions of the particular fellow who happens to be doing the describing. Thus, as Y.-R. Chao put it so well nearly 50 years ago, "[D]ifferent [linguistic] systems or solutions are not simply correct or incorrect, but may be regarded only as being good or bad for various purposes" (1934, p. 363). One cannot therefore in any sense "demonstrate" knowledge of any particular formulation of a rule simply by analyzing adult speech data in terms of that rule, and "maximum generality" is no less arbitrary a decision criterion than any number of others that might be formulated and employed (cf. Derwing, 1974). In this context, therefore, the appropriate course of action seems clear enough, if linguistics is really serious about extending its efforts beyond the bounds of an arbitrary descriptive science and if linguistic theorizing is to develop on the basis of a firm empirical underpinning: we must *first* seek out means of discovering which rules/regularities have *in fact* been learned or acquired by speakers and proceed from there.

What, then, might be a suitable test for knowledge of a linguistic rule? One promising line of inquiry has been suggested by the long tradition of association in linguistics (from Humboldt to Paul to Jespersen to Bloomfield to Chomsky; see Derwing *et al.*, 1979, pp. 4-5) between the concept of rule and the phenomenon of linguistic "creativity." The suggestion is, of course, that it is precisely the knowledge of a rule that enables a speaker to deal appropriately with new or novel linguistic forms. To test for knowledge of a particular linguistic rule, therefore, we need to look at more than familiar speech forms as used under ordinary circumstances of language use, as any number of these might just as well have been learned by rote. What is required is supplementary information about how the language user manipulates *unfamiliar* forms

that are *novel* to his linguistic experience.²

An early and familiar experimental advocate of this particular route to the testing of rule knowledge was Jean Berko (1958). Her approach was to assure novelty by inventing a variety of nonsense words, which were subsequently administered to a sample of child and adult subjects in syntactic frames deemed appropriate to elicit such inflected forms as the plural and possessive of nouns, the past tense of verbs, etc. Her findings are well known: she demonstrated quite conclusively that her subjects could properly and consistently inflect forms that they had surely never been exposed to previously, and on this basis it seemed quite safe to conclude that *some* kind of rule(s) had been learned which enabled them to do this.

Berko also, however, left quite a number of interesting questions unanswered by her study. What *particular* rule had been learned, for example (see Derwing 1979 for a discussion of a number of alternative hypotheses all fully consistent with Berko's findings), and how did this rule develop in the child over time? Not only did Berko deal with a very narrow age range in her child subjects, she also employed only a tiny sampling of the possible range of nonsense stem-types. (On the plurals test, her largest, for example, only about a third of the possible stem-final phonemes were represented, and no stem-final consonant clusters were included at all). Subsequent researchers also contributed relatively little towards the filling of either of these gaps. The very first experimental project I became involved in, therefore, was intended to supply answers to such questions as these. (Small portions of this research have now been reported in the literature and we now know at least some of the answers. For details, see especially Derwing & Baker 1979.)

The most ambitious study connected with this project was one that I began as long ago as 1972. In that study I constructed two sets of 100 mixed nonsense and real stems to be used with each of five different inflections and administered them to 112 subjects evenly distributed over the age range from 3 to 9 years. Ignoring ancillary IQ and other control data also collected, this eventually yielded a total of 56,000 verbal responses, and it took more than a year simply to collect them. For the remainder of this paper, I want to try to explain the main reasons why, even after all this time, the bulk of these data still lie undisturbed in a departmental computer file.

From almost the beginning, this line of research was beset by a number of major challenges, both externally and internally generated. In the former category we encounter first the "logical loophole" argument. This amounts essentially to a claim that behavioral, experimental evidence cannot in principle be used in order to falsify any particular linguistic hypothesis. The argument rests on the distinction (made most forcefully by Chomsky 1965) between language knowledge ("competence"), which is the subject of the linguist's main concern, and the psychological mechanism(s) by which this knowledge might be said to be put to use in actual linguistic behavior ("performance"). The form of the argument goes something like this: if C represents some claim about "competence" and P is a hypothesis (usually unspecified) about "performance," then B, any particular instance of elicited linguistic behavior, is a prediction based on the conjunction of C and P. It is thus only this *conjunction* that may be confirmed or disconfirmed by means of any experimental test (i.e., to put it directly, C can always be preserved in the face of negative evidence by attributing the difficulties to P). On this basis Kiparsky has argued that psycholinguistic experiments can "constitute evidence in one direction only - a positive result will confirm the psychological reality of the tested grammatical rule, but a negative one does not disconfirm it" (1975, p. 203).³

Clearly, however, this line of argument is an indictment not of

experimental psycholinguistics, but rather of the kind of linguistic theorizing that is permitted to invoke it. One simple illustration (originally suggested by D. Broadbent, I am told) should suffice to demonstrate this. Suppose we find some child who is quite adept at basic arithmetic. One possible hypothesis about the "competence" thought to underlie this skill might be to attribute the child not with something so mundane as a learned, laborious, step-by-step procedure for carrying out simple arithmetic operations, but rather with knowledge of number theory. And what if experimental results are found that seem to fly in the face of this hypothesis? Just chalk them up as "performance errors" and the well-favored theory remains inviolate. To invoke this kind of "loophole," therefore, is quite obviously to do something that does not at all enhance the scientific status of the discipline. It is simply not credible to insulate serious scientific hypotheses in this way from the possibility of empirical disproof, and what is required, clearly, is an integrated performance model whose components are all equally susceptible to the same prospect of being demonstrably wrong.

The problem has arisen historically, of course, because linguists have traditionally been concerned primarily with describing the form of utterances, especially (in more recent years) in terms of the possible relationships between one set of linguistic objects or structures and other such sets. This kind of purely descriptive work was being carried out long before any serious concerns were raised about such matters as the "psychological reality" of the formulations, much less questions about how such "purely syntactic" descriptions might be integrated into any kind of workable performance model. When the basic (generative) theoretical and descriptive framework was originally conceived, therefore, not only were such obviously critical factors as discourse context and pragmatics excluded from consideration, but even meaning in general (cf. Chomsky 1957). It comes hardly as a great surprise, therefore, that, as Bresnan has recently observed, "Despite many...expressions of hope by linguists, and despite intensive efforts by psycholinguists, it remains true that generative-transformational grammars have not yet been successfully incorporated in psychologically realistic models of language use" (1982, p. xviii). A model that is originally designed to do one thing is not likely to prove very effective for doing something radically different, and any attempts to force-fit the model to a task for which it is not well suited are bound to be strained, artificial, and, in the end, unsuccessful (see Derwing & Baker 1978 for further discussion).

Leaving behind this ultimately self-defeating mode of objection to experimental work, therefore, let us now move on to a more substantive and important line of criticism. From the standpoint of methodology, for example, the most serious and potentially most devastating of these challenges is certainly the one that bears on the question of the *validity* of the results that have been obtained using the particular experimental technique under discussion here. In 1977, for example, Kiparsky & Menn raised the issue of what they called the "strangeness effect" (p. 64). This criticism was based on research conducted by Haber (1975), which yielded a rather bewildering array of bizarre responses by adults in a Berko-type pluralization task (e.g., the plural of [blif] is [blayvz], etc.), clearly errors of a type which do not commonly occur in spontaneous speech. I replicated Haber's study with a sample of 48 adults from a variety of different educational backgrounds and it was quite true: though the vast majority of the adult subjects performed as expected, the adult responses overall were nonetheless decidedly inferior to those obtained from my older child subjects, who generally performed in almost exact conformity with the supposed "adult rule" (i.e., /-lz/ after stem-final sibilants, /-s/ after voiceless non-sibilants, and /-z/ elsewhere). The main thrust of the Kiparsky-Menn challenge was, of course, the suggestion that the experimental situation itself (involving the rapid-fire eliciting of

responses to a long list of pictures of "nonsense" creatures and actions) was producing results which did not accurately reflect the (adult) subjects' true capacities.

I had, however, noticed very little evidence of this "strangeness effect" with any of my child subjects, and there was, after all, a perfectly straightforward potential explanation for this. Surely the adults knew perfectly well that the pictures and verbal stimuli that they were being asked to respond to were all nonsense, so perhaps, for at least some of these subjects, a nonsensical response seemed called for. For the children, on the other hand, who were quite accustomed to encountering new and unfamiliar words on a regular basis, a corresponding reaction was to be much less expected. But it was important to know for sure, since it was a validity issue that was at stake here. Moreover, it was indeed hard to quarrel with Kiparsky on the point that "the more the experimental design is made to approximate normal linguistic behavior, the more reliably can the results be used for linguistic conclusions" (1975, p. 202).⁴

At my urging, therefore, a "validity check" was subsequently conducted by Rollins (1980), in which results in a Berko-type task were compared with those obtained in a naturalistic "free play" situation. Twenty subjects were involved in this study, ranging in age from 3 to 6 years, with five children in each age group, and three separate testing situations were utilized. The first and third tests were both Berko-type tasks, in which the subjects were individually presented with pictures of six real and eight imaginary animals, the latter consisting of a set of nonsense stem-types that the prior studies had shown to be representative. The second test, conducted a week after the first and one day before the third, involved the introduction of the subjects to a set of large, brightly colored stuffed animals which corresponded to the pictures seen in the first Berko test. The examiner pointed out the names of each of the animals and continued to play with the children (two at a time) until she was confident that the children knew all of the names and could pronounce them correctly. At this point the children were allowed to play with the animals on their own and as they saw fit (some version of a "race game" turned out to be the most popular kind of activity), while the examiner merely stood by and recorded the plural forms as they were spontaneously produced (Prompting occurred only rarely, as when some particular pair of animals had been ignored and the game seemed to be winding down. At this point the examiner would say something on the order of "What about these animals. Don't you want to play with them, too? They're all alone." With this approach, plural forms for all of the items were quite readily obtained.)

The results of Rollins' study are summarized in Table 1 below, which shows the number of "correct" responses by each subject in each of the three tasks. An analysis of these results showed that only the factor of age was significant, as expected; scores were not significantly different as a function of the test type (i.e., Berko vs. the "free play" situation). For children, therefore, the results were solid, and non-validity arguments based on the "strangeness effect" seemed to be quite without force in this instance.

Or so, at least. I might thus far have led you to believe. In fact (as perhaps some of you may have already noticed), I have engaged in a sort of statistical sleight of hand of a kind not uncommonly employed in psycholinguistic research generally, and especially in developmental work. Specifically, I have employed the logical fallacy of using *between-subject* data to draw *within-subject* conclusions. Suppose, for example, that two subjects were to take the same ten-item True/False test and that the first scored correctly on items #1-6, the second on items #5-10. Both subjects would score 60% overall and might on that basis - following quite standard contemporary analytical practice - be grouped together as

"equivalent" performers, despite the fact that that they scored the same on only 20% of the items, viz., items #5-6!) It is important to notice that a version of this same perverse logic has already been presented to you in the form of Table 1. Look again at the results for Subject #1 on the three tests: 6/14 correct on Berko-1, 7/14 on the Free Play task, and 6/14 correct on Berko-2. Is this really "equivalent" performance on the three tests? Perhaps, but *only* if the correct responses were generally obtained on the *same items*. If it were the case instead that this subject performed correctly on items #1-6 on the two Berko tests, say, but on items #8-14 in the free play situation, he would be performing about as *differently* in the two situations as was theoretically possible to do, and just the *opposite* conclusion would have to be drawn to the one I drew in the immediately preceding paragraph. Fortunately for my argument, when we look at Rollins' results on an item-by-item basis, we actually do find the pattern I earlier tried to convince you (on the basis of quite insufficient evidence) was in fact the case, namely, that 6 of the 7 correct responses in the free play task were obtained on the same 6 items that were performed correctly on both Berko tests; in fact, even the *incorrect* responses by this subject proved to involve errors of precisely the same kind across the board. So this subject in fact performed identically on 13 of the 14 items on all three tests, and quite similar patterns held for the other subjects, as well. I may then stick to my original conclusions, after all, though not on the basis of the data as originally presented. By the same token, however, Berko's "% correct" figures don't really tell us very much, when we get right down to it, and she gives us no further details about how consistently her subjects performed across her items. Hence the replications.

There still remains a rather disturbing fly in the ointment, however. What, for example, is meant by the term "correct" response in the discussion above? For Berko, who used an "adult standard" (1958, p. 158), this meant "responding as some adult did." But we already know that adults perform quite erratically on this type of task (and Berko, too, found some measure of variation in her adults' responses). With this approach, what is "correct" will therefore vary as a function of the particular adults tested, yielding in the end a "standard" which is far too unstable to be considered a reasonable standard at all. In my own early work, therefore, I adopted an admittedly arbitrary standard along the following lines: for the nonsense (and real, regular) items, "correct" was taken to mean "in accord with the general rule" (as described a few paragraphs above), while real, irregular forms were evaluated according to the preferences indicated in some standard dictionary (see Derwing & Baker 1977 for details). This fixes the standard, at least, but is it a *good* standard? There are some compelling reasons to suspect not. If we seek to understand the course of rule-learning in the child's development and the various factors that bear upon it, presumably we are not so much interested in comparing the child's performance at each stage with the adult's rule (even if we knew for certain what the form of that rule was), but we are rather more concerned with the question of how the child interprets the data for *himself*. That is, since learning a rule presumably involves making a generalization that cuts across an entire class of forms, it is critical that we discover which forms the child groups together (i.e., which ones he treats *in the same way*) at each stage, whether these groupings conform to the (presumed) "terminal state" or not. In short, it is the *within-child* patterns of responses, once again, that are of crucial relevance here.

A third problem, finally, associated with the analysis of developmental data is the real cruncher: how are subjects to be partitioned into homogeneous "strategy groups" in the first place? This problem has plagued language acquisition studies from the beginning and falls under the general rubric of the "stage" controversy. How does one define a

discrete "stage" and how many of the beasts are there? How different (similar) does the performance of two subjects have to be in order for us to say that they are operating at a different (the same) level? And how can the investigator ever hope to identify the critical patterns or strategy differences (similarities) among his subjects unless he first has some reliable basis for grouping these subjects for purposes of comparison? It was largely this problem, therefore, that was responsible for the long delay in doing anything with the massive corpus of data described early on in this paper, for there was simply no way to deal with it all unless some non-arbitrary means could be found to tell us which subjects belonged with which. Berko herself avoided the problem by simply pooling all of her data and trying to draw some general conclusions about "children" (despite her finding that there were significant differences between her pre-school and first-grade subjects). The standard approach of course, has been to group subjects according to age categories, though it is well known that this is a generally quite unreliable procedure. Age is invariably correlated significantly with performance in developmental work, of course, but the magnitude of the correlation is often far too low to yield much clear evidence of the operative developmental trends (see Innes 1974 for a good illustration of this.) And we have no reason at all to believe, of course, that developmental stages click off in precise synchrony with the motion of the earth around the sun. The currently popular adoption of the "MLU" standard, finally, while it indicates some improvement over the age criterion, at least, is certainly also far from satisfactory. (Does "Stage III" really begin at MLU=2.75 and end at MLU=3.50, as suggested by Brown [1973, p. 56]? The arbitrariness is apparent. See also the extended critique of this measure by Crystal 1974.)

To make a long story short, a significant breakthrough may at long last be at hand. Thanks largely to the inventive genius of my psychologist colleague, William J. Baker, we now have available at Alberta an analytical technique that seems to circumvent all three classes of difficulties just outlined above. The procedure focuses on within-subject response patterns (specifically, on co-occurrences of responses among test items) and it utilizes an original development of the statistical technique known as "hierarchical cluster analysis" to establish both the grouping of the subjects and (within bounds) the number of groups involved. (It can also be utilized to effect an analogous grouping of the test items, once the subject-groups are first identified.) This work has already been described in some detail elsewhere (see Baker & Derwing 1982ab) and time does not permit even a cursory explication here. Suffice it to say for the moment that after a careful evaluation of the technique as applied to familiar data, we now feel justified in extending it to the data from the large inflections study already alluded to, and Figure 1 below represents the first analysis to come out of that effort. It shows the partitioning of the 112 subjects into five distinct strategy groups, based on their performance in a Berko-type pluralization task involving 59 common nonsense stems. Both an age trend and a correlated systematic "honing" of the relevant stem-classes (clusters and all) as subjects progress up the scale provide strong presumptive evidence that a "stage" interpretation is appropriate for these groups. The results so far, in fact, could hardly be more encouraging. (But that, too, must be the topic of another paper.)

Footnotes

¹Contrast Valian, who blandly asserts that the grammar (whose? which one?) is psychologically real "by definition" (1976, p. 64!)

²Kiparsky, by invoking the competence/performance distinction to be discussed further below, makes the claim that, for some rules, at least,

"not even partial productivity would be a necessary criterion for psychological reality" (1975, p. 198). If so, the problem still remains to find some (other) test to determine whether or not any particular formulation of such rules is in fact part of a speaker's knowledge. Claims about knowledge are empirical, psychological claims and for them to be of any scientific interest whatever, they have to be tested (and hence testable). So if Berko's test won't do, then other tests have to be devised.

³In my original abstract I also promised to discuss here an experiment which could beat this logic on its own terms, namely, by testing two rules of the same putative type in precisely the same experimental situation. Since the same performance factors must presumably be operative for both rules in such a case, if one of the rules is borne out in the experiment and the other is not, the particular "loophole" described cannot be invoked to save it. Such an experiment has, in fact, been performed with results as indicated, but the work was conducted by someone else and the data, thanks to the combined efforts of the flu virus and the postal service, were not forthcoming in time for this meeting. I am loath to get into so potentially controversial a study without having the full facts at hand, so I have decided simply to expand somewhat the other portions of my paper and to leave this particular experiment for some other occasion.

⁴Syntacticians, too, might usefully take note of this particular admonition, given the prevailing mode of argumentation built mainly on a base of "grammaticality" judgments - largely self-generated - with respect to sentences totally isolated from any kind of discourse or human interactive contexts. A good, consistent set of such judgments from linguistically untrained subjects has yet to be seen.

References

- Baker, W.J. & Derwing, B.L. 1982a. Response coincidence analysis as evidence for language acquisition strategies. *Applied Psycholinguistics* 3.193-221.
- Baker, W.J. & Derwing, B.L. 1982b. Toward an empirical definition of 'stage' in language development. In C.E. Johnson & C.L. Thew (Eds.), *Proceedings of the Second International Congress for the Study of Child Language*. Lanham, MD: University Press of America.
- Berko, J. 1958. The child's learning of English morphology. *Word* 14.150-177.
- Bresnan, J. (Ed.), 1982. *The mental representation of grammatical relations*. Cambridge, Mass.: MIT Press.
- Brown, R.W. 1973. *A first language*. Cambridge, Mass.: Harvard University Press.
- Chao, Y.-R. 1934. The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology, Academia Sinica*, 4.363-397.
- Crystal, D. 1974. Review of *A first language* by R.W. Brown. *Journal of Child Language*, 1.289-307.
- Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.
- Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.
- Derwing, B.L. 1974. Review of *Linguistic speculations* by F.W. Householder. *Language Sciences* 30.25-32.
- Derwing, B.L. 1979. English pluralization: a testing ground for rule evaluation. In G.D. Prideaux, Derwing, B.L. & Baker, W.J. (Eds.), *Experimental linguistics*. Ghent: E. Story-Scientia.
- Derwing, B.L. & Baker, W.J. 1977. The psychological basis for morphological rules. In J. Macnamara (Ed.), *Language learning and thought*. New York: Academic Press.

- Derwing, B.L. & Baker, W.J. 1978. On the re-integration of linguistics and psychology. In R.N. Campbell & P.T. Smith (Eds.), *Recent advances in the psychology of language*. Part B: *Formal and experimental approaches*. New York: Plenum Press.
- Derwing, B.L. & Baker, W.J. 1979. Rule learning and the English inflections (with special emphasis on the plural). In G.D. Prideaux, Derwing, B.L. & Baker, W.J. (Eds.), *Experimental linguistics*. Ghent: E. Story-Scientia.
- Derwing, B.L., Prideaux, G.D. & Baker, W.J. 1979. Experimental linguistics in historical perspective. In G.D. Prideaux, Derwing, B.L. & Baker, W.J. (Eds.), *Experimental linguistics*. Ghent: E. Story-Scientia.
- Haber, L. 1975. The muzzy theory. *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*. Chicago: CLS.
- Innes, S.J. 1974. Developmental aspects of plural formation in English. Unpublished M.Sc. thesis, University of Alberta.
- Kiparsky, P. 1968. Linguistic universals and language change. In E. Bach & R.T. Harms (Eds.), *Universals in linguistic theory*. New York: Holt, Rinehart & Winston.
- Kiparsky, P. 1975. What are phonological theories about? In D. Cohen & J.R. Wirth, *Testing linguistic hypotheses*. Washington, D.C.: Hemisphere.
- Kiparsky, P. & Menn, L. 1977. On the acquisition of phonology. In J. Macnamara (Ed.), *Language learning and thought*. New York: Academic Press.
- Rollins, W.C. Laboratory vs. 'free' testing situations in language acquisition research. Unpublished B.A. Honors thesis, University of Alberta, 1980.
- Valian, V. 1976. The relationship between competence and performance. *CUNYForum* 1.64-101.

TABLE 1. Proportions of "Correct" Responses on Three Experimental Tasks (Rollins 1980)

Subject	Berko-1	Free Play	Berko-2
1	6/14	7/14	6/14
2	11/14	9/14	11/14
3	8/14	9/14	8/14
4	2/14	2/14	2/14
5	2/14	2/14	2/14
6	10/14	11/14	12/14
7	9/14	10/14	11/14
8	12/14	12/14	13/14
9	11/14	11/14	11/14
10	5/14	7/14	8/14
11	14/14	14/14	14/14
12	8/14	9/14	8/14
13	11/14	12/14	12/14
14	11/14	11/14	11/14
15	12/14	14/14	14/14
16	12/14	13/14	13/14
17	12/14	12/14	12/14
18	11/14	12/14	12/14
19	10/14	10/14	10/14
20	12/14	14/14	14/14

FIGURE 1. Subject Groups on the Plurals Task (n=112)