

CAUSATION OR CORRELATION:
A REPLY TO OLLER AND PERKINS*

John A. Upshur, William Acton,
Bradford Arthur, and Alexander Z. Guiora¹
University of Michigan

For a number of years the University of Michigan Personality and Language Behavior Research Group has been actively investigating relations between second language behavior and various cognitive and non-cognitive dimensions of personality. We welcome therefore the publication of a critical examination of a method frequently employed in studying language proficiency and affective variables: cross-sectional, two-variable, correlation studies (Oller and Perkins 1978 and their references). We especially want to record agreement with a number of points made by Oller and Perkins.

First, simple correlations between two measures cannot be taken as proof of a causal relationship between underlying psychological variables which the measures are designed to test. If scores yielded by two measures X and Y are correlated, one does not know whether variance in the underlying variable x produces variance in the underlying variable y or whether y produces x , or, for that matter, whether z is the cause of both, or indeed whether the correlation is better explained by some other, more complex, causal relationship. We do not intend to argue that one or another of the possible hypotheses about causal relations between language proficiency and attitude is the correct one. There are two reasons: as stated, correlational data alone are by their nature incapable of proving causal relationships; and theories are in general indeterminate on empirical grounds. Choices among competing theories or hypotheses are, therefore, more likely made on the basis of elegance, parsimony and other theoretic-methodological grounds.

Second, establishing relationships among psychological vari-

*Editor's note: this response was received just as this issue was going to press. John Oller and Kyle Perkins will be invited to respond to this reply if they wish, in the next issue.

¹The authors are members of the University of Michigan Personality and Language Behavior Research Group.

ables depends upon the validity of the instruments which are used to measure them. Approval motivation, self-flattery and response set may influence a set of measures to such an extent that significant correlations are found among the measures, correlations which are independent of any relationships which might or might not exist among the underlying psychological variables. The validity of professed attitude measures is especially suspect because of the commingling of affective and cognitive operations in the measurement task. The Micro Momentary Expressions Test is another example of a suspect measure, and its authors actively discourage its use as a measure of individual differences in empathy.²

Having established agreement that correlation is not equivalent to causation and correlations among invalid measures yield no important theoretical conclusions, this reply to Oller and Perkins has two further aims. We intend to examine the strength of Oller and Perkins' arguments against the validity of self-report measures, and second, to evaluate whether any of the studies cited by Oller and Perkins provide plausible evidence for a causal relation between variance in affective variables and variance in second language attainment (as distinct from correlations between measures), a relation independent of verbal intelligence, approval motivation, self-flattery and response set.

Oller and Perkins employ two types of argument to cast doubt on the validity of so called self-report measures of affective variables. The first involves analogic reasoning from claims about the validity of self-report measures of a cognitive variable, language

²The transposition which Guiora et al. (1967) refer to is not from button pushing to language learning, but from psychotherapy to language learning. It is hypothesized that strong parallels exist between trying to understand a patient in therapy and trying to understand a speaker of a language which one knows only slightly; that the first is conducive to successful therapy and the second for successful language learning. Success in therapeutic intervention is even less easy to measure than is success in second language attainment. If personality factors can be shown to contribute to successful language learning, these findings become suggestive of traits important to therapists.

The misunderstanding by Oller and Perkins has no bearing on questions of the MME test's validity. However, because of the limited reliability or questionable validity of cognitively mediated tasks the Personality and Language Behavior Research Group has in recent years pursued two different lines of investigation: experimental studies (e.g., the effects of drugs on second language behavior, *viz.* Guiora et al., 1972), and development of more direct measures of affect (e.g., Galvanic skin response, inhibition of response) in second language behavior.

proficiency. The second type of argument involves offering plausible explanations for correlations obtained with self-report measures, explanations which differ from those which would indicate construct validity of the measures. Arguments of both types, as employed by Oller and Perkins are flawed enough to seriously weaken them. These flaws will be taken up in turn.

Oller and Perkins argue by analogy that the validity of self-ratings of affective variables are not apt to exceed the validity of self-ratings of language proficiency. Next they report a series of correlations between self-reports and tests of proficiency and report an upper-bound for validity of self-ratings at .69. Even if one is prepared to accept the analogy, the upper-bound is not supported. The claim about validity of self-rating of affect is a claim about the correlation between a rating and an underlying psychological variable. In this case the value which the psychological variable takes is a true criterion. The reports about validity of proficiency self-ratings on the other hand are reports about correlations between two measures *neither* of which is perfectly reliable or valid. Thus the correlation which Oller and Perkins propose as an upper-bound is not a proper upper-bound for a correlation between a self-rating and a true criterion at all; it is attenuated by unreliability and invalidity of the proficiency test which is used as a criterion measure. For example, if the validity of Murakami's dictation score as a test of speaking ability were .69 (and this would be extremely high for a correlation between a test and a true criterion), then self-ratings of speaking ability could correlate 1.00 with a true criterion of speaking ability, and by Oller and Perkins' analogic reasoning the proper upper-bound for validity of self-ratings of effect becomes not .69 or 50% of total variance, but 1.00 or 100% of total variance.

No claims are made here that self-reports are perfectly valid. What is claimed, however, is that Oller and Perkins have not provided a reasonable upper-bound for validity of self-ratings of affect. Because one cannot provide such an upper-bound for self-ratings of proficiency short of 1.00 no upper-bound for any other self-ratings short of 1.00 can be analogically inferred.

Oller and Perkins' second line of attack on the validity of affective measures involves showing that three traits—other than those which a self-report instrument is designed to measure—may influence subjects' reports of self-ratings. These are the approval motive, self-flattery and response set. As Oller and Perkins rightly point out these three traits may lead to spurious correlations among measures which claim to test different constructs if the

traits produce consistent, non-random variance across measures. That is, the approval motive, self-flattery and response set must (1) be differentially present in different subjects and (2) should affect total scores or ratings on the two measures to a comparable degree. The second condition can be met if either the traits are not independent and their total effect on two measures is comparable, or the traits are independent and their individual effects are comparable. These are rather exacting conditions to satisfy, but are not out of the question. The degree to which they are satisfied will determine the degree to which two attitude measures are spuriously correlated. Unfortunately Oller and Perkins did not point out this logical requirement of their argument, the premises (which may not be so readily accepted) that approval motive, self-flattery and response set are variable across subjects and comparably present in two measures for the measures to be spuriously correlated.

It appears that Oller and Perkins may have overlooked this logical requirement. For example, they mention "institutionalizing" the approval motive in attitude measures. The most obvious effect of institutionalizing, it seems, is to do away with individual differences in approval motivation. The effect of this is, of course, to reduce extraneous non-random variance in the measures and, therefore, to reduce spurious correlations among measures which might otherwise be affected by subjects' differential attempts to look good. In other words, approval motivation would have an *increased main effect* upon scores but would exert *decreased within-subjects influence* upon scores. That is, average scores would be increased but variance would be decreased.

Oller and Perkins cite two studies (with data summarized in Table 1) to show the influence of self-flattery.³ The data showing a high relationship between self-ratings on various dimensions of personality and valuations of those dimensions are given only one interpretation by the authors although there is available a simple alternative which is at least as plausible. Oller and Perkins suggest that subjects tend to rate themselves highly on characteristics which they value. The alternative is that subjects tend to value highly those characteristics which they possess and realize that they possess. Both interpretations are instances of self-flattery. The

³Although Oller and Perkins erroneously average coefficients of correlation, a proper computation of the correlation among all rating-valuation correlations yields a marginally higher value. This is not however, important to their argument.

first, however, challenges the validity of self-ratings of personal characteristics while the second does not. What Oller and Perkins present as a clear argument with supporting data turns out, therefore, to be ambiguous at best.

The data presented in Table 1 may also have some bearing on Oller and Perkins' caveat that self-ratings may be highly contaminated by response set. This warning is well taken, and response set may at times be an important source of variance in self-ratings. That it need not be so, however, is illustrated by the Table 3 data. Of the fifty-six correlations between self-ratings on one scale and valuations of a different scale, forty-nine failed to reach significance. Since self-ratings and valuations of identical scales correlated substantially, if response set were very important then self-ratings and valuations of different scales should correlate significantly. But the number of correlations significant at $p < .05$ is only 7-1/2% greater than the number expected on the basis of chance alone.

Although Oller and Perkins have failed to establish a proper upper-bound for the validity of self-report measures of attitude, and although their arguments in favor of spurious correlations among measures of proficiency and measures of affect seem to be flawed and therefore overstated, none the less the possibility still exists that obtained correlations do not reflect relations among underlying psychological states. Therefore the correlations obtained in the studies which Oller and Perkins cite must be interpreted with caution. The only study cited which does not rely predominantly upon correlational data is Savignon's (1972) investigation of three "methods" of teaching French. Savignon employed a three group, quasi-experimental design in which the three groups differed with respect to the "laboratory component" of their first year French classes. One group, E_1 , interacted in French with native speakers of French; a second group, E_2 , was provided with materials (films, etc.) designed to promote appreciation of French culture; the third group participated in a standard audio lab. The results of a series of language tests and attitude and orientation measures were analyzed by analysis of variance.

Most of Savignon's findings about changes in reported attitudes may be explained away as Oller and Perkins have done.⁵ She did, however, find a significant group x time interaction for integrative orientation. Group E_1 , the learners whose laboratory

⁵There is a serious problem with the "sorting out" explanation, however. The median correlation between verbal intelligence and five positive attitude measures administered at the end of the course to all groups is .01.

sessions required a degree of social integration with native French speakers, showed a decrease in integrative orientation over the first half of the course. The other groups showed slight but insignificant increases. This decrease is opposite to what one would expect of reasonably intelligent students on the basis of approval motivation. The distribution of response set should not differ from group to group and so should not contribute to the interaction; if it has any effect at all it should be to promote uniformity of scores over time—an effect opposite to that found for group E₁. Self-flattery can account for this decrease in integrative orientation under the twin conditions that students (1) are able to perceive their limited abilities to integrate into a contrived French speaking society, and (2) express values to fit their perceptions of abilities rather than report abilities which fit the values they hold. It seems that in this case, therefore, one must admit a relation between a change in values and experience in language learning—or else one must conclude that the significant interaction is a rare, chance occurrence.

In conclusion, it seems that there is only limited evidence for a relation between affective states and second language attainment and that further and more convincing evidence is unlikely to be provided by simple correlational studies. Further, it seems that causal relationships among affect and proficiency will not be found to exist importantly in linear finite models such as those which are implied by Oller and Perkins' figures 1-3. Hypotheses derived from feedback models may be more fruitful areas for investigation. We find it more plausible that both cognitive and noncognitive aspects of personality affect language attainment and also that language attainment affects personality.

REFERENCES

- Guiora, A. Z., Beit-Hallahmi, Robert C. L. Brannon, C. Y. Dull, T. Scovel. 1972. The effects of experimentally induced changes in ego states on pronunciation ability in a second language: an exploratory study. *Comprehensive Psychiatry* 13:5.
- Guiora, A. Z., Harlan L. Lane, L. A. Bosworth. 1967. An exploration of some personality variables in authentic pronunciation of a second language. In H. Lane and E. Zale (eds.). *Studies in Language and Language Behavior*. Ann Arbor: University of Michigan.
- Savignon, Sandra J. 1972. *Communicative Competence: An Experiment in Foreign Language Teaching*. Philadelphia: The Center for Curriculum Development, Inc.