# Controlling Acquiescence Response Bias by Item Reversals: The Effect on Questionnaire Validity

Chester A. Schriesheim and Kenneth D. Hill

*Educational and Psychological Measurement* 1981 41: 1101
DOI: 10.1177/001316448104100420

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1981, 41

# CONTROLLING ACQUIESCENCE RESPONSE BIAS BY ITEM REVERSALS: THE EFFECT ON QUESTIONNAIRE VALIDITY

CHESTER A. SCHRIESHEIM AND KENNETH D. HILL
Graduate School of Business Administration
University of Southern California

The prevailing conventional wisdom is that it is advisable to mix positively and negatively worded items in psychological measures to counteract acquiescence response bias. However, there has been virtually no unambiguous empirical evidence to support this recommendation. Thus, an experiment was conducted to evaluate the ability of subjects to respond accurately to both positive and reversed (negative) items on a questionnaire. Items from the LBDQ—XII Initiating Structure and Consideration subscales were used to create a written description of a fictitious manager. One hundred-fifty subjects, all upper-division business undergraduates, were given the written managerial description and then asked to complete a questionnaire containing the twenty Initiating Structure and Consideration items. The managerial descriptions were in two forms (to portray high and low Initiating Structure), and the questionnaires contained items in three forms (all positively worded, all negatively worded, and mixed). The data were evaluated using a one-way analysis of variance and post hoc *t*-tests. Significant differences in response accuracy were found between the item wording conditions. It was concluded that it may not be advisable to employ reversed (negatively-worded) items to control acquiescence response bias, as such changes may actually impair response accuracy.

This research was supported by a grant from the Research Board of the Graduate School of Business Administration at the University of Southern California. Requests for reprints should be sent to Chester A. Schriesheim, Department of Organizational Behavior, Graduate School of Business Administraton, University of Southern California, Los Angeles, CA 90007.

CONVENTIONAL wisdom has suggested that psychological measures should be constructed containing an even balance of positively and negatively worded items, so as to counteract response biases such as agreement response tendencies or acquiescence. Thus, for example, Nunnally (1978) stated that "stylistic variance . . . can be mostly eliminated by ensuring that an instrument is constructed so that there is a balance of items keyed 'agree' and 'disagree' . . ." (p. 669). Many other psychometricians have made similar statements. Clearly, the general consensus in the literature has been that measures should have both positive and negative items (e.g., Scott, 1968; Anastasi, 1976).

This consensus has found its way to many specialty areas in educational and psychological testing, including the measurement of perceived leader behaviors. Thus, for example, Schriesheim and Kerr (1974) indicated that (1) subject agreement response tendencies can usually be controlled by having an adequate number of negatively worded items, and (2) based upon their review, the major existing measures of perceived leader behavior are inadequate in terms of not having sufficient negative items. Schriesheim and Kerr concluded in their review that revised scales are needed, and that these measures should have a larger number of negatively worded items to offset acquiescence response biases.

Although the recommendations of Schriesheim and Kerr are in conformity with conventional psychometric advice, the experience of the authors (which has accumulated since the 1974 review) suggests that the use of negative items may have at least some dysfunctional consequences. In their experience, negatively worded items often reduce scale reliability, and they may be eliciting response biases or measure unintended aspects of constructs under investigation. In any event, as measurement validity requires instrument reliability (Nunnally, 1978), these impressions suggest that the use of negative items may not be cost-free. Thus, the purpose of the current study was to examine the effects of item wording on the accuracy of responses to standard measurement instruments. As some vehicle is needed to explore these effects, a measure of perceived leader behavior was employed (the revised Form XII of the Leader Behavior Description Questionnaire or LBDQ—XII) (Stogdill, 1963). This measure was selected because (1) the investigators are very familiar with the instrument, (2) extensive knowledge already exists about its psychometric properties (Schriesheim and Kerr, 1974), (3) some very limited research also exists on negative wording effects on leadership descriptions (Taylor and Bowers, 1972), and (4) the description of leader behavior allows the

use of an objective referent and accuracy scores as dependent variables (Schriesheim and DeNisi, 1978). These reasons are elaborated in paragraphs that follow.

## *Background*

There has been considerable psychological research over the last thirty years dealing with concepts broadly classified under the general heading of "response styles." Specifically, acquiescence, response sets, and various other response biases have been examined as they relate to different types of item wordings. Most of this research has utilized measures of the California F-scale (Adorno, Frenkel-Brunswik, Levinson, and Sanford, 1950), the Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway and McKinley, 1967), and the Personality Research Form (PRF) (Jackson, 1967). (One may consult, for example, Bentler, Jackson, and Messick [1972], Clayton and Jackson [1961], Morf and Jackson [1972], Rorer [1965].)

Although this research has been valuable in questioning the interpretation of subject responses to these measures, it has not resolved the issue of response style relevance. The argument has been heated over whether or not response styles exist, and, if so, whether they impact upon research results in a meaningful way. Rorer (1965), for example, concluded that "the inference that response styles are an important variable in personality inventories is not warranted on the basis of the evidence now available" (p. 150). Jackson and Messick (1965) responded to Rorer with extensive criticism of both his data and his conclusions.

One trouble with these arguments is that they have been based on inferences drawn from the California F-scale, MMPI, and PRF measures, and there is no objective referent or comparison which can be used to assess the validity of responses or the influence of subject biases. Consequently, alternative explanations for the findings of this type of research abound. What would eliminate this difficulty is research which uses subject descriptions of someone other than themselves as the referent. Once such an external referent has been established, manipulations of the referent and the measurement instrument can then be performed by standard experimental procedures. The objective nature of the referent would provide a basis for assessing validity (in terms of description accuracy) and differences in subject scores would not have the numerous alternate explanations which exist without such a referent.

In devising an objective referent for experimental purposes, one would find it desirable to employ a stimulus which is familiar to subjects. Thus, since managers who supervise subordinates ("leaders") are probably familiar to most people, this stimulus was chosen as a base to construct an objective description referent. In addition, leader behavior descriptions are easily manipulated in a "script" format to create portraits of specific leaders, and questions about the leaders can then be asked which have an objective referent. By varying the wording of the questions (e.g., from positive to negative), the impact of item wording on the ability of subjects to answer the questions accurately could be directly assessed. This method would reduce the alternate explanations which exist without such an objective referent. In fact, an earlier study on "implicit theories" (Schriesheim and DeNisi, 1978) showed the feasibility of manipulating objective portrayals of leader behavior by a "script" design. Basically, Schriesheim and DeNisi (1978) employed the LBDQ—XII (Stogdill, 1963) to create scripts portraying leaders at various levels of Initiating Structure (task-oriented, directive) and Consideration (person-oriented, supportive) behavior. These scripts were then assessed by subjects using the LBDQ—XII itself. This procedure can obviously be modified for the current study, by using scripts and manipulating the LBDQ—XII questionnaire items themselves (creating positive and negative item forms) to determine the effects of item reversals on the accuracy of reported leadership behaviors.

In addition to the reasons just cited for employing leadership (and the LBDQ—XII) as a context for re-examining the effect of item wording (reversals) on response accuracy, there is one final justification for this approach. Taylor and Bowers (1972) conducted some very limited field research which suggests that leadership questionnaires may be affected by item wording in a manner similar to that found in some of the personality instruments that have been used in the response style research summarized in the previous paragraphs. Taylor and Bowers (1972) used a total of fifteen subjects and administered an instrument with eight positively and eight negatively worded items; they discovered that ". . . respondents found it more difficult to respond . . . where some questions were negative and some positive . . ." (p. 20). Taylor and Bowers also observed ". . . that a negatively worded item more often produces a higher mean response than does the positively worded counterpart . . .," and that ". . . the wording of the negative items spuriously changed the impact of the questions: Negative items . . . may well, by their connotation, change the position of the items on a true scale

of goodness and badness, such that respondents over-compensate." (p. 24). Thus, Taylor and Bowers uncovered evidence suggestive of possible adverse item reversal effects. It should be noted, however, that their study can only be considered suggestive of the need for further research (because of the sample size and referent problems). The purpose of the current research was therefore to explore the item reversal issue further and its implications for the validity of questionnaires.

## *Method*

### *Sample*

The subjects in this experiment consisted of 150 upper division undergraduates enrolled in business administration courses at the University of Southern California. The experiment was conducted on a voluntary basis, and all responses were kept anonymous. Assignment to the experimental conditions was random, both within and across classes, and an equal number of subjects was assigned to each condition. Upon completion of the experiment, all subjects were debriefed and dismissed.

### *Procedure*

Each subject was given a one-page description of the behaviors displayed by a fictitious supervisor (each subject received only one description). The subjects were asked to read their particular "script" very carefully, and when they had completed reading the script, they were instructed to turn it face down on their desks (they were not allowed to consult it further). The subjects were then given questionnaires and asked to describe the behavior of the supervisor in the script by using this instrument. Two script conditions and the three questionnaire conditions were randomly distributed, and each subject received only one script and one questionnaire condition. The subjects were instructed to copy the "form number" from the top of the script onto their completed questionnaire for condition identification purposes. All materials were then collected.

The measures used to record the subjects' responses to the scripts consisted of items from the Consideration (*C*) and Initiating Structure (*IS*) subscales of the LBDQ—XII (for the reasons previously discussed). Extensive development went into each item to prepare a negative item form (positive form if the item was originally negative), based upon substantial input from a panel of 20 faculty and

PhD students in the Organizational Behavior Department at the Graduate School of Business Administration of the University of Southern California. Table 1 presents the positive and negative item forms employed.

The three variations of the IS and C subscale items in the LBDQ—XII used to record the subjects' responses were either (1) all positively worded, (2) all negatively worded, or (3) evenly-mixed positive and negative items. The negatively worded items on the mixed instrument were those designated by numbers 7, 8, 11, 12, 13, 15, 16, 17, 18, and 20 (all items are numbered in accord with Table 1).

The manipulation of the level of Initiating Stucture in the scripts was accomplished by varying adverbs describing each behavior. Although the LBDQ—XII Consideration (*C*) and Initiating Structure (*IS*) scales employ ten items (each), it seemed unrealistic to provide complete information by including all ten items in the scripts. Thus, the two scripts contained descriptions of the leader on eight items for both C and IS subscales (the omitted ones being items 9, 12, 13, and 20). Both scripts had a common amount of extraneous

TABLE 1
*Positive and Negative Item Forms*

*Initiating Structure Items:*
1. He (lets) (does not let) group members know what is expected of them.
3. He (encourages) (discourages) the use of uniform procedures.
5. He (tries out) (does not try out) his ideas in the group.
7. He (makes) (does not make) his attitudes clear to the group.
9. He (decides) (does not decide) what shall be done and how it will be done.
11. He (assigns) (does not assign) group members to particular tasks.
13. He (makes sure) (does not make sure) that his part in the group is understood by group members.
15. He (schedules) (does not schedule) the work to be done.
17. He (maintains) (does not maintain) definite standards of performance.
19. He (asks) (does not ask) that group members follow standard rules and regulations.

*Consideration Items:*
2. He is (friendly and approachable) (unfriendly and unapproachable).
4. He does little things (to make it pleasant) (which make it unpleasant) to be a member of the group.
6. He (puts) (does not put) suggestions made by the group into operation.
8. He (treats) (does not treat) all group members as his equals.
10. He (gives) (does not give) advance notice of changes.
12. He (mixes with others) (keeps to himself).
14. He (looks out for) (is unconcerned with) the personal welfare of group members.
16. He is (willing) (unwilling) to make changes.
18. He is (willing) (unwilling) to explain his actions.
20. He (consults the group before acting) (acts without consulting the group).

filler material designed to provide neutral background information about an "average" supervisor (to make the manipulations less obvious and to prevent the subjects from uncovering the purpose of the scripts before they had been told to turn the scripts over and complete the LBDQ—XII).

The level of behavior displayed by the leader was manipulated by stating that the leader either "always" behaved in the manner described by the LBDQ—XII items (the "high" manipulation), or that the leader "never" behaved in the manner described by the LBDQ—XII items (the "low" manipulation). The *nonmanipulated* Consideration dimension, which was always present, was described at a medium level of behavior. Since the LBDQ—XII uses the response categories, "always," "often," "occasionally," "seldom," and "never" (scored 5, 4, 3, 2 and 1, respectively), the medium level for Consideration was presented by describing the leader using the adverbs "often" and "seldom" three times (each) and the adverb "occasionally" twice—yielding an average level of behavior at the midpoint ("3") of the LBDQ—XII response scale.

The previous information is summarized as follows: on each of the two scripts used, the C behavior dimension was held constant (at a medium level of behavior), whereas the IS dimension was presented at either a high or low level of behavior. A sample script indicating how IS was manipulated is shown as follows:

[Sample Script—High Condition]

Robert Brown is a department supervisor. His department is one of many in a large corporation located in a medium-sized city in the Midwest. He is five feet ten inches tall, in good health and physical condition, thirty years old, married, with two children, and has worked for his employer for six years. His performance is about average in comparison with that of others at his job level, being neither particularly good nor bad. In his role as supervisor, Bob is occasionally friendly and approachable towards each subordinate, and he often looks out for their personal welfare. He *always* maintains definite standards of performance for subordinates and *always* schedules their work. Bob often treats the unit members as his equals, but seldom does little things which make it pleasant and enjoyable to work for him. He *always* makes his attitudes clear to the group and *always* tries out his ideas with his subordinates. Bob sometimes works late at the office, but he tries not to ignore his family and personal life. Bob often gives his subordinates advance notice of changes but seldom puts suggestions made by unit members into operation. He *always* lets group members know what is expected of them, and he *always* assigns them to particular tasks. Bob seldom explains his actions to his subordinates, and is occasionally willing to make changes which unit members suggest. He *always*

encourages the use of uniform procedures, and he *always* asks that group members follow standard rules and regulations. On the whole, Bob tries to balance off his work and personal lives.

The example just presented is the script designed to portray a leader who is high in IS (the manipulated dimension) and medium in C (the nonmanipulated dimension). The italicized adverbs (italicized in the passage only for illustration) are those words where substitutions were made to change the manipulated level of IS from high to low (by substituting "never").

It should be noted that IS was not manipulated to test the effects of IS level, but to create enough experimental variance to obtain reliable dependent variable measures. The design, therefore, was basically one-way with three treatments (not a two-by-three design).

### *Manipulation Check*

In order to verify that the manipulation of IS was in fact perceived, a *t*-test was conducted on the high and low IS respondent scores. After reflecting (reverse-scoring) all negative items, the investigators found a significant difference in the predicted direction ($t = 21.44$; $p < .001$) between the low ($\bar{X} = 21.43$; $SD = 6.78$; $n = 75$) and high ($\bar{X} = 42.73$; $SD = 5.30$; $n = 75$) groups—an outcome indicating that the manipulation was perceived as intended. After reflection, the C subscale scores were also significantly different ($t = 3.23$; $p < .001$) for the low ($\bar{X} = 29.56$; $SD = 4.62$; $n = 75$) and high ($\bar{X} = 31.93$; $SD = 4.37$; $n = 75$) IS conditions. However, this effect was both irrelevant and trivial. (The absolute size of the difference between mean scores on the C subscale was 2.37, whereas the corresponding difference was 21.30 for the IS contrast.) Therefore, no further exploration was made.

### *Method of Analysis*

First, coefficient alpha internal consistency reliabilities were computed within each treatment, to ensure that reliable results were obtained (all negative items having been reflected for these analyses). Following the reliability check, accuracy scores were computed (these being described in the next paragraph) and one-way analyses of variance (ANOVA) were conducted for the three treatments to evaluate the effects of the item wording conditions. Finally, post hoc *t*-tests were employed to evaluate directional differences for the significant ANOVA results.

The method used to analyze the data was to compare the accuracy scores of the subjects between the three item conditions. Because

both scripts portrayed the manipulated IS behaviors at the extremes of the 5-point LBDQ—XII response scales (i.e., as "never" or "always" displayed), accuracy scores were computed as the absolute value of the difference between the behavior levels portrayed in the scripts and the IS scores provided by the subjects. For example, the supervisor whose behavior is described in the sample script presented earlier should have been given the highest possible score on IS by the subjects. Thus, on each IS item, a score of "5" would be considered perfect accuracy; for each IS item, then, the difference between 5 and the score provided by the subjects constituted the accuracy score ($5 - X$). Similarly, the supervisor portrayed as "low" on IS should have received an IS score of "1" on each item, and the accuracy score for subjects in this condition was computed by subtracting 1 from each IS item score ($X - 1$). It should be noted that the magnitude of accuracy scores as thus computed was *inversely* related to their degree of accuracy (i.e., the smaller the score, the more accurate the description); all negatively worded items were reflected (reverse-scored) for the accuracy score computations, and different item combinations were summed to form dependent variables for the ANOVA and the *t*-test analyses. (These combinations are described in the next paragraph.)

## *Results*

As noted previously, the results of the experiment were first analyzed for reliability. Table 2 presents the reliability results for each sample relative to each of the three experimental conditions, as well as for the total sample. Table 2 also breaks the results down to reflect the manner in which the items were presented on the questionnaires. There were five items held positive on the mixed items instrument and five held negative. These are reported separately. (For further elucidation one may consult footnotes "b" and "c" to Table 2.) As shown in Table 2, the reliabilities for IS scores were quite high for all the items, with the only exception being the five negative items in the mixed item condition (.70)—a reliability coefficient that is still acceptable, however, according to Nunnally (1978).[1]

Table 2 also shows the mean accuracy scores for each sample

[1] The reliabilities for Consideration, on the other hand, were quite low for each of the item groups and treatments (ranging from .02 to .48, with a mean of .28 and a median of .21). This outcome was as expected, because of a lack of variance in Consideration descriptions (since this dimension was held constant at the midrange of the response scale in both scripts). (An item analysis confirmed low item variability as the source of the low reliabilities.)

**TABLE 2**
*Coefficient Alpha Reliabilities, Treatment Accuracy Scores, and One-Way ANOVA Results*

| | Experimental Treatment Group | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Positive ($N$ = 50) | | | Mixed ($N$ = 50) | | | Negative ($N$ = 50 | | | Total Sample ($N$ = 150) | | | Summary ANOVA[a] | |
| Group of Items | $\alpha$ | $\bar{X}$ | SD | $\alpha$ | $\bar{X}$ | SD | $\alpha$ | $\bar{X}$ | SD | $\alpha$ | $\bar{X}$ | SD | MS | F |
| All 10 IS Items | .96 | 7.10 | 4.73 | .89 | 10.50 | 6.11 | .91 | 10.44 | 7.59 | .93 | 9.35 | 6.42 | 189.33 | 4.84** |
| 5 Positive IS Items[b] | .91 | 3.54 | 2.32 | .91 | 3.90 | 2.82 | .80 | 5.18 | 3.97 | .88 | 4.21 | 3.17 | 37.15 | 3.83* |
| 5 Negative IS Items[c] | .93 | 3.56 | 2.79 | .70 | 6.60 | 4.26 | .88 | 5.26 | 4.50 | .86 | 5.14 | 4.09 | 116.06 | 7.54*** |

[a] The analyses of variance ($df$ = 2, 147) were conducted across the three treatment groups (positive, mixed, and negative items) within each of the three item groups (all 10 IS items, 5 positive items, and 5 negative items); the total sample results shown were not involved in these analyses.

[b] These are the five items which appeared in positive form for both the all positive and mixed item treatment groups (they were in negative form for the all negative treatment group).

[c] These are the five items which appeared in negative form for both the mixed and all negative item treatment groups (they were in positive form for the all positive treatment group).

* $p < .05$.

** $p < .01$.

*** $p < .001$.

corresponding to each of the three item treatments and for the total sample, with the items broken down into three groups: all ten IS items, the five positively worded IS items on the mixed instrument, and the five negatively worded IS items on the mixed instrument. As can be seen from Table 2, the mean accuracy scores were higher for all three item groups for the negative wording treatment as compared to the positive wording treatment (indicating less accuracy). Similarly, for all three item groups, the mean accuracy scores were higher for the mixed item wording treatment than for the positive wording treatment. Finally, for two of the item groups (all ten IS items and the five negative items), the mean accuracy scores were higher for the mixed item wording treatment than for the negative treatment.

Examination of the magnitudes of the differences in accuracy indicates that, overall, the mixed and all negative item treatments resulted in about fifty percent more inaccuracy than that obtained in the all positive treatment (i.e., 10.50/7.10 = 1.48; 10.44/7.10 = 1.47; as is evident in Table 2). Furthermore, even the greatest accuracy (that obtained for the all positive wording treatment) still resulted in approximately 24% inaccuracy (7.10/30.00 = 0.24; 30.00 representing the average of the actual behaviors portrayed in the high [50.00] and low [10.00] scripts). Thus, these results would appear to have practical significance (statistical significance being discussed subsequently), as the differences in accuracy (24 versus approximately 36%) appeared to be large enough to warrant concern among instrument users.

The one way analysis-of-variance results are shown on the right-hand side of Table 2 for each of the three item groups. As shown, the obtained significant results for each item group indicated a clear effect of item wording on accuracy.

The results of the post hoc *t*-tests comparing mean accuracy scores by treatment group are presented in Table 3. This table shows a noteworthy pattern. When one compares the accuracy of the three item treatments (all positively worded, mixed positively and negatively worded, and all negatively worded items), the following can be seen. For all ten items combined, both the mixed and all negative wording treatments were significantly less accurate than the all positive wording treatment. For the five items which were presented in positive form in the all positive and mixed treatments (and in negative form in the all negative treatment), significantly less accuracy was obtained in the negative wording treatment group as compared with that in the positive wording treatment. Finally, for the five items which were negative in the mixed and in the all

negative wording treatments (and positive in the all positive wording treatment), significantly greater accuracy was obtained for the all positive treatment in comparison with both the mixed and all negative wording treatments. Thus, the all positive wording treatment generally yielded significantly greater accuracy than did either the mixed or all negative wording treatments. There was no significant difference in accuracy between the mixed and all negative wording treatments.[2]

### *Discussion and Conclusion*

In examining the effects of item reversals by the use of an objective description referent, this study has shed some interesting light on an issue that has here-to-fore been the focus of arguments based upon ambiguous data and results. Overall, the findings suggest, first, that the use of all positively worded items resulted in more accurate descriptions than the use of either mixed or all negatively worded items, and that these differences were both practically and statistically significant (as revealed in the first row of entries in Tables 2 and 3). This outcome was directly in contrast to the conventional psychometric recommendations summarized earlier. Second, the decrement in accuracy seemed to be generally a function of the negatively worded items themselves, not of their exerting a strong contextual effect on the positive items.

As shown in Tables 2 and 3, there was no statistically significant difference ($t = -0.70$) in mean accuracy between the all positive (3.54) and mixed (3.90) wording treatments for the five positively worded items. However, the difference ($t = -4.22$) between the positive (3.56) and mixed (6.60) wording treatments was significant for the five negatively stated items (an outcome indicating an effect of item negativeness). Likewise, the difference in means between the mixed (3.90) and all negative (5.18) wording treatments for the five positively worded items ($t = -1.86$) was very close to significance ($p < .06$) (a result suggesting a negativeness effect), whereas the difference for the five negatively stated items ($t = 1.53$) was not close to significance ($p > .10$). These findings, coupled with the clear-cut results shown in the second column of Table 3, seemed to

[2] The same exact patterns as discussed previously (except even stronger) were obtained in this experiment for Consideration for all three treatments and all three item groups. Although these patterns are not reported because of the unreliability of the data (as indicated in footnote 1), they do increase confidence in the general findings of this investigation.

TABLE 3
*Post Hoc Comparisons of Mean Treatment Group Accuracy Scores*

| Groups of Items | Comparison (t-test) Values for Experimental Treatment Groups | | |
|---|---|---|---|
| | Positive vs. Mixed | Positive vs. Negative | Mixed vs. Negative |
| All 10 IS Items | −3.11** | −2.64** | 0.04 |
| 5 Positive IS Items[a] | −0.70 | −2.52* | −1.86 |
| 5 Negative IS Items[b] | −4.22*** | −2.27* | 1.53 |

[a] These are the five items which appeared in positive form for both the all positive and mixed item treatment groups (they were in negative form for the all negative treatment group).

[b] These are the five items which appeared in negative form for both the mixed and all negative item treatment groups (they were in positive form for the all positive treatment group).

* $p < .05$.

** $p < .01$.

*** $p < .001$.

indicate that it was the negatively worded items which were the source of inaccuracy in the descriptions, and not a decrement in accuracy for positively stated items when they were mixed with negatively worded items. However, the weak trend for the five negatively stated items under the mixed versus all negative wording treatments ($t = 1.53$; $p < .15$) suggests that inaccuracy was slightly compounded for negatively written items when they were mixed with positively stated items (and that this circumstance was why the total set of ten mixed items was not more accurate than the group of all negatively worded items). The somewhat low reliability of these five negatively worded items for the mixed wording treatment group (.70; as shown in Table 2) also suggests a small detrimental context effect for negatively stated items when they were mixed with positively worded items.

The results of this study suggested a rather important conclusion for measurement instrument design: that the inclusion of negatively worded items can result in less accurate responses and therefore impair the validity of obtained results. Thus, although the inclusion of negatively stated items may theoretically control or offset agreement response tendencies, their actual effect is to reduce response validity. This situation suggests that current recommendations (e.g., Nunnally, 1978) concerning the desirability of including both positive and negative items on a questionnaire may be premature and apparently warrant much further investigation. Clearly, one study is not adequate to dismiss conventional wisdom, especially since acquiescence effects may be more troublesome on other types of instruments (such as measures of non-leadership variables). Howev-

er, the current study should suggest caution in the use of negative (reflected) items, at least until further investigations are conducted.

## REFERENCES

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., and Sanford, R. N. *The authoritarian personality*. New York: Harper, 1950.

Anastasi, A. *Psychological testing* (4th Ed.). New York: MacMillan, 1976.

Bentler, P. M., Jackson, D. N., and Messick, S. A rose by any other name. *Psychological Bulletin*, 1972, 77, 109–113.

Clayton, M. B. and Jackson, D. N. Equivalence range, acquiescence and over generalization. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 371–382.

Hathaway, S. R. and McKinley, J. C. *MMPI manual* (Rev. Ed.). New York: Psychological Corporation, 1967.

Jackson, D. N. *Personality Research Form Manual*. Goshen, N. Y.: Research Psychologists Press, 1967.

Jackson, D. N. and Messick, S. Acquiescence: The nonvanishing variance component. *American Psychologist*, 1965, 20, 498.

Morf, M. E. and Jackson, D. N. An analysis of two response styles: True responding and item endorsement. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1972, 32, 329–354.

Nunnally, J. C. *Psychometric theory* (2nd Ed.). New York: McGraw-Hill, 1978.

Rorer, L. G. The great response style myth. *Psychological Bulletin*, 1965, 63, 129–156.

Schriesheim, C. A. and DeNisi, A. S. The impact of implicit theories on the validity of questionnaires. Paper presented at the American Psychological Association Convention (Division 14, Organizational and Industrial Psychology), Toronto, Canada, August, 1978.

Schriesheim, C. A. and Kerr, S. Psychometric properties of the Ohio State leadership scales. *Psychological Bulletin*, 1974, 81, 756–765.

Scott, W. A. Attitude measurement. In G. Lindzey (Ed.), *The handbook of social psychology*. Vol. 2 (2nd Ed.). Reading, MA: Addison-Wesley, 1968.

Stogdill, R. M. *Manual for the leader behavior description questonnaire—Form XII*. Columbus: Bureau of Business Research, Ohio State University, 1963.

Taylor, J. C. and Bowers, D. G. *Survey of organizations: A machine-scored, standardized questionnaire instrument*. Ann Arbor: Institute for Social Research, University of Michigan, 1972.