Factors Defined by Negatively Keyed Items: The Result of Careless Respondents?

Neal Schmitt
Michigan State University

Daniel M. Stults
Quaker Oats Company

A frequently occurring phenomenon in factor and cluster analysis of personality or attitude scale items is that all or nearly all questionnaire items that are negatively keyed will define a single factor. Although substantive interpretations of these negative factors are usually attempted, this study demonstrates that the negative factor could be produced by a relatively small portion of the respondents who fail to attend to the negative–positive wording of the items. Data were generated using three different correlation matrices, which demonstrated that regardless of data source, when only 10% of the respondents are careless in this fashion, a clearly definable negative factor is generated. Recommendations for instrument development and data editing are presented.

Most textbooks or publications listing recommendations concerning attitude scale construction include the caveat that questionnaire items include both negatively and positively worded item stems (e.g., Anastasi, 1980; Adkins-Wood, 1961; Thorndike, 1971; Wiggins, 1973). However, there is a relatively large body of literature on response styles, which indicates that these wording changes may make significant differences in the factor structure of scales and the item validities (Bentler, Jackson, & Messick, 1971). Bentler et al. argued convincingly for two different types of acquiescence response styles. Agreement acquiescence results when a person responds positively to all statements in a personality instrument or attitude scale. Acceptance acquiescence occurs when a person considers all personality characteristics or attitude statements as descriptive of him/herself or some object but disagrees with all statements that deny such characteristics. The objective of this article is not to resurrect the debate over types of response styles or even their existence (Block, 1971; Rorer, 1965), but rather to demonstrate that a small portion of respondents who are careless in reading the items may be responsible for the appearance of a factor consisting solely of negatively keyed items. These negatively keyed items may be either polar opposites (happy-sad) or a negation of some trait or descriptor (happy-not happy).

At the outset, it is very important to define what is meant by "careless" in this article. The careless respondent, who is the subject of this article, is not responding randomly. He/she is simply reading a few of the items in a measuring instrument, inferring what it is the items are asking of the respondent, and then responding in like manner to the remainder of the items in the instrument. This means that any item that is phrased inconsistently with the rest of the items in the instrument will elicit a response that is inconsistent with responses to the rest of the item pool and inconsistent with the responder's real position on the construct being measured. For example, a student responding to a teacher evaluation instrument with a 5-point Likert-type

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 4, December 1985, pp. 367–373

0146-6216/85/040367-07$1.60

scale may decide the faculty person is above average as a teacher and may intend to mark 4s on the scale. Instead of reading the items in the evaluation instrument, the respondent simply marks 4 to all items, including those that express a negative opinion about the faculty member. In analyzing these responses, all negatively keyed items are recoded. The result of this carelessness on the part of the respondent is not random; it is systematic. All negatively keyed items will be positively correlated with each other and negatively correlated with the remaining positively worded items. This type of responding is consistent with Bentler et al.'s (1971) notion of agreement acquiescence.

All paper-and-pencil instruments are subject to this problem. In addition to student evaluations of faculty, the same type of error is possible in performance evaluations of faculty, in performance evaluations used in other contexts, manipulation checks in social-experimental research, personality measures, attitude scales, interest inventories, and survey research. This article illustrates what *can* happen in factor analytic research of data which include a relatively small portion of respondents who are careless as defined above. This is a matter of convenience only; a similar possible problem exists with any questionnaire or self-report measure, whether or not the measure is factor analyzed. Further, it is important to note that this study was *not* demonstrating that people *have* responded this way in any previous research, but it demonstrates that this type of careless responding is one feasible explanation of the appearance of this factor. Identification of people who *do* respond in this fashion is more problematic, though some possibilities are suggested in the discussion section below.

Very frequently, authors reporting factor or cluster analyses of responses to an attitude scale or personality inventory find that a majority of the negatively keyed items (usually a minority of the items in most measures) load on, or define, a single factor. For example, Schmitt and Coyle (1976) factor analyzed a 74-item questionnaire concerning the reactions of college student applicants to placement interviewers. The second of six applicant reaction factors identified in their study was defined by negative descriptors such as the following: irritable, defensive, used inappropriate words, lost train of thought, explained in unnecessary detail, self-conscious, and so forth. Of 13 items defining this factor, only 1 was positive. Further, only 6 negative items were loaded most highly on other factors.

A similar pattern of factor loadings is seen in a study of perceived support for innovation in secondary schools. Siegel and Kaemmerer (1978) evaluated a pool of 525 statements thought to be descriptive of innovative and traditional organizations. Their final three-factor solution included a factor they titled Tolerance of Differences, which included a predominance of negatively keyed items such as "This place seems to be more concerned with the status quo than with change" and "The best way to get along in this organization is to think the way the rest of the group does."

Another example of this phenomenon appears in industrial/organizational and work stress research and involves a measure of role conflict and ambiguity (Rizzo, House, & Lirtzman, 1970). Recently, Tracy and Johnson (1981) have pointed out that all eight items of the role conflict scale are worded to represent stressful or conflict-laden characteristics of a work role. The six role ambiguity items are all worded to represent nonstressful or unambiguous characteristics of the role. The intended meaning of the scales (conflict vs. ambiguity) is totally confounded with the difference in wording indicating either stress (which was labeled role conflict) or comfort (labeled role ambiguity).

A similar effect has been noted in early research on the *F*-scale (Adorno, Frankel-Brunswik, Levinson, & Sanford, 1950). Items consist of relatively strongly worded opinions, most of which express a critical attitude about human nature. When investigators began questioning whether *F*-scale scores reflected an authoritarian personality or a response style, a reflected *F*-scale was constructed. Correlations between the original *F*-scale and this reflected scale were only .20 (Chapman & Campbell, 1957; Messick & Jackson, 1958). Jackson and Messick (1961, 1962) found similar factor analytic results for the MMPI, though subsequent item-reversal studies (the original items are reversed) of the MMPI indicated high correlations between the original

measures and the reversal measures (Lichtenstein & Bryan, 1965; Rorer & Goldberg, 1965a, 1965b).

In summary, the result of factor analyses on scales with negatively keyed items frequently leads to the identification of a factor defined wholly or mostly by those negatively keyed items. The literature cited alone indicates that this finding is relatively widespread in the sense that it occurs in a variety of research areas. Examples included studies of interview impressions, personality scales, and role ambiguity.

The objective of the present study was to show how a "negative factor" can be produced by a relatively small number of careless respondents who do not notice that some items are the opposite in meaning to the majority of the items. In a series of simulations, the proportion of "careless" respondents and the proportion of negatively keyed items were varied for data generated from three different correlation matrices reflecting different levels of item intercorrelation.

Method

Data Generation

To simplify comparisons, three 30-item correlation matrices were selected to serve as the sources of the data which were generated and analyzed. These matrices were chosen so as to represent different levels of item intercorrelation and different substantive content.[1]

The first matrix (ASSMT) represented the intercorrelations of ratings on 15 skill dimensions by two raters in an assessment center (see Schmitt, 1977, for a description of the rating dimensions). The average item intercorrelation across the 30 items was .36; the range of item intercorrelations was from .00 to .82. Principal components analysis yielded seven factors with an eigenvalue greater than 1.0. Eigenvalues for these seven factors were 10.78, 3.42, 2.14, 1.85, 1.51, 1.35, and 1.0. Although use of the eigenvalue criterion would have suggested seven factors, the scree criterion (Cattell, 1966) suggested three factors, as did content considerations in earlier component analyses (Schmitt, 1977).

The second matrix consisted of intercorrelations of responses to 30 items in the Central Life Interest (CLI) measure developed and researched by Dubin and his colleagues (Dubin, 1956; Dubin & Champoux, 1974; Dubin & Goldman, 1972). These 30 items are meant to measure a single factor, but they are dichotomously scored, hence item intercorrelations are relatively low. In this sample, average item intercorrelations were .13; the range of intercorrelations was from −.13 to .39. Eigenvalues for the 11 factors whose eigenvalues were greater than 1.0 were 3.91, 1.86, 1.63, 1.47, 1.32, 1.28, 1.23, 1.11, 1.05, 1.02, and 1.01. Use of the scree criterion, plus the fact that these are items designed to measure a single concept, would have suggested a single factor. Because of the relatively low level of item intercorrelation, many "small" factors were obtained.

The third matrix (SEMSQ) of intercorrelations was generated by responses to the 10 items of the Rosenberg self-esteem measure (Rosenberg, 1965) and the 20 items of the Minnesota Satisfaction Questionnaire (Weiss, Dawis, England, & Lofquist, 1967), which is usually divided into intrinsic and extrinsic satisfaction subscales. Average item intercorrelations were .28; the range of intercorrelations was from −.01 to .71. The eigenvalues of the seven factors with eigenvalues greater than 1.0 were 8.53, 3.25, 1.65, 1.25, 1.18, 1.09 and 1.04. Again, the scree criterion as well as content considerations might have suggested a three-factor solution.

The content of the items that served as the basis of this study was not particularly important, but the matrices were chosen because they represented actual, but relatively diverse, matrices in terms of item intercorrelation.

Data for the study were generated in the following manner for each of the three initial correlation matrices:

1. The complete factor loading matrix (Number of factors equals 30) was computed for each matrix.

[1]The three correlation matrices are available from the first author.

2. The signs of the factor loadings for 4, 8, or 12 randomly selected items were changed to represent unreflected negatively keyed items.
3. Each of these factor loading matrices was then used as input to the Ohio State Correlated Score Generation Method (Wherry, Naylor, Wherry, & Fallis, 1965), and the "responses" of 400 people were generated. Each "positively" keyed item was given a mean of 5; negatively worded items were given means of 3. All items had standard deviations of 1.2. Decimals were truncated and response values greater than 7 were recoded to 7; those less than 1 were recoded to 1. The result was a set of 400 responses to 30 items, each with a 1 to 7 response scale and intercorrelations that were representative of the original correlation matrices.

Data Analysis

All negatively keyed items were recoded for all 400 cases for each of three basic sets of data as they normally would be, and principal components analyses were conducted. The factor loadings matrices for these analyses should be reflective of what would be obtained if substantively meaningful interpretations were made by all respondents to all items (0% careless results). The eigenvalue criterion was used to determine how many factors to rotate as would be fairly typical in exploratory factor analyses. Varimax rotation of these factors was used in all analyses. Next, factor analyses were conducted for the same set of data when a randomly selected subset of the cases was left *unrecoded*. Data matrices based on four different proportions of unrecoded cases (5%, 10%, 15%, and 20%) were analyzed to determine how many careless respondents can create a factor loading matrix in which there is a factor identified primarily or wholly by negatively keyed items.

Dependent Variable

As evidence that these manipulations were creating a factor identified solely by unrecoded items, the number of negatively keyed items that appeared on each factor and the number that appeared on the same factor for each condition were counted. In all cases, the factor loading that was highest determined the placement of a variable on a factor.

Results

The results of the counts of the factor loadings of negatively keyed items are presented in Table 1. As can readily be seen in Table 1 for each data set (ASSMT, CLI, SEMSQ), a clearly identifiable "negative" factor appears when only 10% of the respondents answer as if they failed to notice that some portion of the items were worded inconsistently with the majority of the items. Some clustering of negative items is already present when only 5% of the respondents are "careless," but probably not enough that investigators would recognize the problem. The number of "negatively" keyed items in the item pool did not seem to have much effect on the identification of a negative factor, though with an increase in the number of such items, the negative factor became more prominent in the solution. Originally it was expected that there would be differences across matrices in how easily a negative factor is created by careless responding, but this did not seem to take place. Results were consistent across the three matrices studied.

Space considerations preclude reproduction of all factor matrices for the conditions which were generated, but the results for all three correlation matrices were highly similar.[2] With no "careless" respondents, the negatively keyed items were scattered across all factors (using highest factor loading as a means of defining factors) as would be expected if the respondents were sensitive to the content of the items. This pattern continues to be true when 5% of the cases were not coded appropriately. With 10% "careless" respondents, however, the first factor is typically defined by the "negative" items. In the event of 15% and 20% "careless" respondents, all negatively worded items were found to load on the first factor. As the percentage of "careless" respondents increased from

[2]These factor analytic results are available from the first author.

Table 1
Number of Factors on Which Negatively Keyed Items Appear and the Largest Number of Negatively Keyed Items on a Single Factor

Matrix and Percent of Careless Respondents	Number of Negatively Keyed Items					
	Four		Eight		Twelve	
	N_F[a]	N_{NS}[b]	N_F	N_{NS}	N_F	N_{NS}
SEMSQ Matrix						
0%	4	1	6	3	7	3
5%	4	1	5	3	5	4
10%	1	4	2	7	2	9
15%	1	4	1	8	2	10
20%	1	4	1	8	1	12
ASSMT Matrix						
0%	2	2	3	4	5	3
5%	2	2	3	4	4	4
10%	1	4	1	8	3	9
15%	1	4	1	8	1	12
20%	1	4	1	8	1	12
CLI Matrix						
0%	4	1	7	2	6	3
5%	3	2	3	6	4	5
10%	1	4	2	7	1	12
15%	1	4	1	8	1	12
20%	1	4	1	8	1	12

[a] N_F is the number of different factors on which the negatively keyed items were most highly loaded.

[b] N_{NS} is the number of negatively keyed items which loaded highest on a single factor. This count was always done on the factor defined by the largest number of negative items.

10% to 15% to 20%, the size of the factor loadings for the negative items increased. Although there were slight variations in this pattern, the results across matrices and number of negatively keyed items were remarkably similar.

Conclusions and Recommendations

The results of these analyses have a clear implication for researchers who factor or cluster analyze data in which the wording of items is varied. Such researchers should be highly suspicious of factors loaded primarily with negatively keyed items. Likewise, consumers of this research should question substantive interpretations of such negative factors. The results of this study indicate that, with only 10% of the respondents ignoring the wording of items, a negative factor will appear regardless of the substantive meaning of the items.

What can a researcher do if he/she is concerned about this problem or when he/she recognizes that it is a potential problem in the analysis of item responses? First, questionnaire instructions may include a warning to potential respondents that some questions will be negatively keyed and that they should attend to all items.

Second, researchers should be especially concerned with overall questionnaire length or with a

lengthy set of items that employ the same response format. The temptation to include similar items to increase the internal consistency of a set of items measuring a single construct must be balanced by a concern that respondents will become fatigued or bored when they answer many like-sounding items. This precaution is consistent with research by Trott and Jackson (1967), who found that an acquiescence factor was strongly associated with the speed of presentation of personality items. When items were presented under speeded conditions, the largest factor obtained indicated an almost complete separation between true- and false-keyed scales. With less demand for speed, the acquiescence factor was sixth largest and not as clearly defined. Moreover, the content factors were more easily determined.

Third, researchers should be especially cautious concerning negative factors when responses to questionnaires are "involuntary" or when there is some reason to sabotage the research effort. This is certainly possible when the respondents are college sophomores, but it is equally likely in data collection efforts carried out with varying degrees of organizational sponsorship. All three of these recommendations are qualitative and speculative. Based on available evidence, the exact influence each of these factors has in producing careless responses of the type described in this article cannot be indicated. In this context, it may be useful to experiment with the wording of directions and the length of questionnaires/instruments as well as the serial position of any negatively keyed items. Further, the context in which data are collected could be varied in an effort to assess the effect of context on the presence or absence of "negative" factors.

Fourth, data should be edited in a way in which unusual response patterns may be detected. For example, each respondent's data should be examined to find unusual responses. If negative and positive items are recoded so as to be consistent, then a respondent whose primary responses on a 7-point scale are 5 and 6 would be suspicious if negatively worded item responses were 2 and 3. Responses from these individuals would be best deleted prior to any further analyses. A more systematic analysis of these "careless" respondents is possible with use of item response theory (IRT). Latent trait analyses (Lord, 1980; Wright & Stone, 1979) allow the determination of which item responses made by an individual are not well predicted by the IRT model. As a consequence, it is possible to detect unusual responses at the individual level. These unusual responses would be a deviation from those predicted by the IRT model. Since sample size and number of item requirements for IRT analyses are large, however, latent trait parameters may not be obtainable for many instruments.

Finally, editors and reviewers of papers reporting factor analyses in which a negative factor appears should demand that authors consider the possibility that some portion of their respondents were careless and that appropriate editing of data take place.

It bears repetition that all this study demonstrates is what *could* happen if respondents failed to notice negatively keyed items. Research directed to a determination that respondents actually *do* respond this way should be conducted. Data editing to identify such respondents would be necessary.

It should be pointed out that this study analyzed randomly generated data based on only three correlation matrices. Other factors may influence the appearance and prominence of a negative factor. However, the consistency with which the negative factor presents itself, even when the proportion of "careless" examinees is relatively small, indicates that this is a likely explanation of the occurrence of at least some of the reports of negative factors in the published literature. Given the relative frequency with which a negative factor is reported in the literature and the ease with which such a factor is produced, researchers should be especially wary when their factor analyses produce factors that are loaded primarily by negative items. Further, users of questionnaires should also take steps to minimize the problem in the construction of their instruments and the directions which accompany those instruments.

References

Adkins-Wood, D. (1961). *Test construction*. Columbus OH: Merrill.

Adorno, T. W., Frankel-Brunswik, E., Levinson, D. J., & Sanford, R. W. (1950). *The authoritarian personality*. New York: Harper.

Anastasi, A. (1980). *Psychological testing*. New York: MacMillan.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76,* 186–204.

Block, J. (1971). On further conjectures regarding acquiescence. *Psychological Bulletin, 76,* 205–210.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276.

Chapman, L. J., & Campbell, D. T. (1957). Response set in the *F*-scale. *Journal of Abnormal and Social Psychology, 54,* 129–132.

Dubin, R. (1956). Industrial workers' worlds: A study of the "central life interests" of industrial workers. *Social Problems, 3,* 131–142.

Dubin, R., & Champoux, J. E. (1974). Workers' central life interests and job performance. *Sociology of Work and Occupations, 1,* 313–326.

Dubin, R., & Goldman, D. R. (1972). Central life interests of American middle managers and specialists. *Journal of Vocational Behavior, 2,* 133–141.

Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement, 21,* 771–790.

Jackson, D. N., & Messick, S. (1962). Response styles on the MMPI: A comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology, 65,* 285–299.

Lichtenstein, E., & Bryan, S. H. (1965). Acquiescence and the MMPI: An item-reversal approach. *Journal of Abnormal Psychology, 70,* 290–294.

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

Messick, S., & Jackson, D. N. (1958). The measurement of authoritarian attitudes. *Educational and Psychological Measurement, 18,* 241–253.

Rizzo, J. R., House, R. J., & Lirtzman, S. I. (1970). Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly, 15,* 150–163.

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

Rorer, L. G., & Goldberg, L. R. (1965a). Acquiescence and the vanishing variance component. *Journal of Applied Psychology, 49,* 422–430.

Rorer, L. G., & Goldberg, L. R. (1965b). Acquiescence in the MMPI? *Educational and Psychological Measurement, 25,* 801–817.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton NJ: Princeton University Press.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. *Journal of Applied Psychology, 62,* 171–176.

Schmitt, N., & Coyle, B. W. (1976). Applicant decisions in the employment interview. *Journal of Applied Psychology, 61,* 184–192.

Siegel, S. M., & Kaemmerer, W. F. (1978). Measuring the perceived support for innovation in organizations. *Journal of Applied Psychology, 63,* 553–562.

Thorndike, R. L. (1971). *Educational measurement*. Washington DC: American Council on Education.

Tracy, L., & Johnson, T. W. (1981). What do the role conflict and role ambiguity scales measure? *Journal of Applied Psychology, 66,* 464–469.

Trott, D. J., & Jackson, D. N. (1967). An experimental analysis of acquiescence. *Journal of Experimental Research in Personality, 2,* 278–288.

Weiss, R. V., Dawis, G., England, G. W., & Lofquist, L. W. (1967). *Minnesota Studies in Vocation Rehabilitation: Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis: University of Minnesota.

Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr., & Fallis, R. F. (1965). Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika, 30,* 303–313.

Wiggins, J. S. (1973). *Personality and prediction*. Reading MA: Addison-Wesley.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Author's Address

Send requests for reprints or further information to Neal Schmitt, Department of Psychology, Michigan State University, East Lansing MI 48824-1117, U.S.A.