PERSONNEL PSYCHOLOGY
2002, 55

# THE EFFECTS OF REQUIRED ELABORATION OF ANSWERS TO BIODATA QUESTIONS

NEAL SCHMITT
Department of Psychology
Michigan State University

CHARLES KUNCE
ACT, Inc.

The impact of a request that examinees elaborate on their answers to a subset of items in a biodata instrument was evaluated. Four forms of a test in which different subsets of items are elaborated were randomly administered to 4 groups of examinees taking a pilot form of a selection instrument for a civil service position. Results indicated significantly lower scores on items for which elaborations were requested than the items for which no elaborations were requested. Lower scores were also observed for nonelaborated items when these items were embedded among those that were elaborated, and lower scores were found when the elaborated items were presented only in the first half of the test. Although the results suggest that requiring elaborated answers may reduce scores on biodata items, several practical and theoretical questions should be investigated to determine the utility of this approach as a method of reducing socially desirable responding.

The use of noncognitive measures in personnel selection is very attractive in a wide variety of personnel selection contexts. They have demonstrated validity (Hunter & Hunter, 1984; Reilly & Chao, 1982; Schmidt & Hunter, 1998) and often display very small or no minority–majority subgroup differences (Schmitt, Clause, & Pulakos, 1996; Stokes & Toth, 1996). Moreover, the interpersonal and attitudinal characteristics measured in these instruments would seem to be increasingly important with the rapid growth of service industries (Hough, 1998) and the increasing emphasis on work teams (Guzzo & Salas, 1995; Hackman, 1991). On the other hand, the correct answers to questions on a noncognitive test are often easy to guess, and uncertainty about the impact of distorted responses on selection decisions and validity continues to concern both practitioners and those interested in the constructs underlying noncognitive measures (Snell, Sydell, & Lueke, 1999). Whether this distortion is deliberate impression management or the result of self-

Correspondence and requests for reprints should be addressed to Neal Schmitt, Department of Psychology, Michigan State University, E. Lansing, MI 48824-1117; schmitt@msu.edu.

deception (Paulhus, 1984; 1991), it will have some impact on the construct(s) measured by noncognitive tests. The purpose of this paper is to present an evaluation of a new approach to the control of such distortion in a biodata instrument consisting mostly of items that would be considered noncognitive in nature.

Most of the literature on faking of noncognitive measures has involved the investigation of personality tests (e.g., Ellingson, Sackett, & Hough, 1999; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Paulhus, 1984), but some research on faking has also been conducted on biodata instruments (Lautenschlager, 1994). The interest in faking biodata instruments has increased with the change in the nature of these instruments over the last 5 decades of their use. Early biodata instruments included questions that were relatively objective and verifiable (e.g., what is the highest educational degree you have obtained?). Studies (Keating, Paterson, & Stone, 1950; Mosel & Cozan, 1952) indicated that there was some inflation of responses even on these forms, but that the correlation between applicant responses and data collected from previous employers was in the high .90s. In addition, earlier empirical keying approaches that were sometimes not as transparent have been replaced with an emphasis on construct-oriented scoring keys (Hough & Paullin, 1994; Schmitt, Jennings, & Toney, 1999). Contemporary biodata questions are now often indistinguishable from personality items in content, response format, and scoring. Personality tests typically contain items regarding values and attitudes and biodata items generally focus on past achievements or behaviors, but even this distinction is not obvious in many biodata applications today.

Researchers disagree about the meaning and effect of response distortion or faking. There is general agreement that applicant and incumbent test scores on various noncognitive measures differ substantially (e.g., Hough et al., 1990; Jackson, Wroblewski, & Ashton, 2000; Rosse, Stecher, Miller, & Levin, 1998). It also seems to be the case that using various methods to remove examinees suspected of faking or correcting the scores of these individuals has no impact on the criterion-related validity of noncognitive measures (Hough et al., 1990; Ones, Viswesvaran, & Reiss, 1996). Some researchers maintain that the ability to project a positive self-image in an employment context is actually job related, though the meta-analysis by Ones et al. (1996) indicated that measures of social desirability did not function as a predictor of job performance nor as a mediator or suppressor of predictor–criterion relationships. More recent work, however, has found that response distortion does affect who is hired, indicating that there are individual differences in the extent of distortion and that there may be some impact on the nature of the construct measured (Douglas, McDaniel, & Snell, 1996; Rosse et al., 1998).

In this paper, we first briefly review previous attempts to assess and control response distortion of noncognitive instruments with a special emphasis on biodata. We then describe and evaluate a relatively new attempt to control for the inflation that may result from conscious/self-deceptive distortions. This procedure is based on the finding that response distortion on application blanks is more likely on questions that are not objective and cannot be verified (Becker & Colquitt, 1992) and the finding in the broader social cognition literature that people tend to overstate their abilities unless they believe their actual abilities will be verified (Fiske & Taylor, 1991). We acknowledge at the outset, however, that in this study, we cannot verify that response differences under various conditions are/are not a function of distortion or the result of test-taking motivation, the time available, or some other factor. Nor do we have evidence of the impact of our manipulation on test validity. However, given the key role that potential response distortion plays in decisions about test use and interpretation, any demonstration of significant differences in a predicted direction should reveal a potentially important intervention.

*Attempts to Assess and Control Faking*

In a thorough review, Lautenschlager (1994) concluded (a) faking biodata items was possible and did occur even when relatively objective verifiable items were used; (b) faking may be less likely or possible when items were empirically, as opposed to rationally, keyed; and (c) attempts to control faking have been largely unsuccessful. Lautenschlager (1994) pointed to several differences across studies of the faking of biodata instruments that make it difficult to provide unqualified conclusions about the extent or control of faking. For example, early studies (Cascio, 1975; Goldstein, 1971; Keating et al., 1950; Mosel & Cozan, 1952) examined the extent of faking on a few items with verifiable answers such as might appear on an application blank. Of these studies, those that reported correlations between responses obtained when faking was suspected and conditions in which faking was unlikely reported a high degree of correspondence. When comparing absolute agreement of responses in these two conditions, there was evidence of inflation of responses in a direction favorable to the applicant (Mosel & Cozan, 1952).

This finding points to another inconsistency across studies: the operational definition of accuracy. In addition to the two measures cited above (correlations between responses and verified information and absolute agreement between these sources of information), the most frequently used index has been a mean difference between groups of persons whose responses are suspect (i.e., they are applicants for desirable

options in the second form as well, but in addition, they were required to provide elaborations of their answers to 17 of the first 46 items. These 17 items were scattered throughout the first half of this form. The third form also required answers to the same 91 items with no elaboration of items in the first half of the test, but 18 items scattered throughout the second half required elaboration. The fourth form required answers to the 91 items and elaboration of the full set of 35 items required of respondents to Forms 2 and 3. Of the 311 total participants, 77, 79, 77, and 78 responded to Forms 1, 2, 3, and 4 respectively. Several different scores were computed based on the items in these forms, including total scores on all 91 items, scores on each half for the elaborated items, and scores on each half for the nonelaborated items. These five scores allowed us to compare responses to the same and different items from examinees who were and were not asked to elaborate their responses.

*Data Analyses*

Basic descriptive analyses were conducted to determine scale reliability, means, standard deviations, and intercorrelations. The three primary hypotheses were addressed using a 2 (elaboration required vs. no required elaboration) by 2 (test half) analysis of variance. The five scores described above were the dependent variables in these analyses of variance. Finally, the correlations between the biodata scores and the scores on the essay, verbal ability, and job knowledge tests were computed to determine whether scores on elaborated and nonelaborated items were related to these alternate measures.

## *Results*

In Table 1, we present the means and standard deviations for each of the four conditions for all five biodata scores. Coefficient alpha reliabilities and intercorrelations for these five indices and gender and education are presented in Table 2. Gender was not related to any of the test measures, but education was correlated significantly ($p < .05$) with all tests ranging from .20 with the total of the biodata items to .41 with job knowledge. Race differences between Caucasians and African Americans were .25 standard deviations on the biodata measure and between .86 and .94 on the other three measures. Differences between Caucasians and the other two sizable subgroups (Asian and Hispanic Americans) were approximately half that large. Caucasian means were the highest in all cases.

As can be seen in Table 2, all four subsets of items as well as the total scores on the biodata instrument exhibited acceptable levels of reliabil-

fake. The instrument used, Assessment of Background and Life Experiences (ABLE), is similar to that used in the research described in this paper in that Likert-type response scales and questions about background and experiences were used to assess what were thought to be job-relevant aspects of personality. Across 11 scales, the standardized mean difference (*d*) between honest and faked conditions ranged from .31 to .86. Corrections for this difference using their measure of social desirability reduced these mean differences to a range of −.23 to .15, or effectively zero. Although this result would suggest that the social desirability correction was effective, results of an analysis of decisions regarding who would be selected under honest and fake response sets with and without the correction suggested that there was no difference in the proportion of time that correct decisions about the "best" people were made. Their conclusion was that social desirability scores were not equivalent to scores obtained when respondents were honest.

Another attempt to control faking involves warning responders about the consequences or possible detection of faking. Dwight and Donovan (1998) in a meta-analysis reported a standardized mean difference of .21 between warning and no warning conditions. Because the usual difference between honest and fake good conditions is usually about .5 (Ones, Viswesvaran, & Korbin, 1995), warnings removed less than half of the impact of faking. The warning literature is relevant to the study reported in this paper because all respondents did receive a warning that indicated that their responses were subject to verification and that "deliberate attempts to falsify information may be grounds for not employing you or dismissing you after you begin work." This warning appeared in the pilot booklets used in our study because it was also being used in the operational exam booklets administered later, but the meaning of such a warning to pilot examinees who are not actually seeking employment is questionable. The impact of our attempt to assess and minimize the possibility of response distortions must be measured in the context of this warning and this candidate group.

In sum, the research on faking indicates a sizable difference between individuals who are thought or told to respond honestly and those that are suspected or told to respond in a manner that would enhance their prospects of being chosen for some desirable outcome. The use of corrections based on scores on special social desirability scales does not seem to be effective. Warnings about the need to be honest in one's responses have some impact, but likely remove less than half the impact of social desirability. The use of option-keyed responses (Kluger et al., 1991) may remove much of the impact of social desirability, but there are practical reasons why they often cannot be used.

In the study described in this paper, we explored the impact of requiring respondents to a biodata instrument to elaborate on their responses to a subset of the items. These elaborations involved requiring the candidates to more fully describe the manner in which their answers were true, or to actually relate incidents that supported their answers. A form of this technique has been used previously in accomplishment records (Hough, 1984). Accomplishment records require the respondent to indicate previous experiences that are relevant to a particular job-related skill. They are also required to indicate their role or responsibility in these incidents and to provide references that can verify the extent of their contribution to the project or experience they describe. Obviously, a full accomplishment record would require considerable writing on the part of applicants who are responding to a biodata form of several dozen or a hundred or more items, which is often typical. There is evidence that such demands discourage some job applicants, and perhaps differentially so, across various examinee subgroups (Ryan, Ployhart, Greguras, & Schmit, 1998; Ryan, Sacco, McFarland, & Kriska, 2000). Our approach was to require a very abbreviated form of applicant elaboration to a subset of the items in a 91-item instrument. Typical of the elaborations required are the examples in the Appendix. Our use of elaborated items was based on the assumption that the required elaboration of examinee answers would stimulate more appropriate responses either because elaboration forces the applicant to remember more accurately (i.e., avoid self-deception) or to avoid attempts to manage impressions in ways that will maximize a favorable hiring outcome. Specifically, our hypotheses are as follows:

*Hypothesis 1*: The inclusion of elaboration requirements on some items will produce lower scores on the elaborated items.

*Hypothesis 2*: The inclusions of elaborated items will also produce lower scores on other nonelaborated items in the same form of the instrument.

*Hypothesis 3*: The effect of elaboration requirements will carry over to portions of the test forms on which the elaboration is no longer required.

In addition to these hypotheses, we also correlated scores on elaborated and nonelaborated items with tests of verbal ability and job knowledge. If scores on elaborated items are more highly associated with verbal ability or job knowledge than are scores on nonelaborated items, it might mean that this requirement would unfairly advantage those individuals with greater levels of verbal and/or writing skills and greater levels of knowledge of the job for which they are being considered. Some might

consider this pattern of correlations evidence of validity, but it is important to note that if our purpose is to measure personality constructs, these "inflated" correlations may constitute contaminants.

## *Method*

### *Sample*

The sample consisted of 311 examinees taking a pilot form of a selection instrument for a federal civil service job. Fifty-eight percent were female; 11% were Asian American; 9% were African American; 14% were of Hispanic descent; 63% were Caucasian, and 3% were of other ethnic backgrounds. Their educational level ranged from high school graduate with no college (3%) to doctorate (2%). The modal educational category was some college, but no degree was indicated by 39%, followed by bachelor's degree (22%), some graduate study (11%), and master's or law degree (11%). These examinees responded to the measures used in this study in seven different testing locations throughout the United States. Every fourth candidate at each site was given one of the four forms of the biodata instrument described below. An anonymous reviewer suggested that because different forms were administered to candidates at each site, some individuals would notice that other examinees were writing more or less than they were and be suspicious. This is possible, but not likely, because only one fourth of the examinees were asked to do no elaborations, because all were physically separated by relatively large spaces to avoid cheating problems, and because other parts of the examination required writing (see below).

Examinees were recruited by ads in the local media and promised $100 to complete the testing. However, in their admission letter, they were told that "those who score high overall with accurate and complete responses on the [biodata examination] will receive a $10 bonus (a total of $110)." This statement of incentive was intended to increase the degree of motivation of the pilot examinees on the biodata instrument. After the test, all participants were paid $110 regardless of actual performance. Examinees were also warned that their responses were "subject to verification" and that "deliberate attempts to falsify information may be grounds for not employing you or for dismissing you after you have begun work." Although it was clear to all examinees that these were pilot examinations, some "applicants" were potentially interested in employment by the agency funding the research. The agency was an attractive employer who received several thousands of applicants for a few hundred positions annually.

*Measures*

The measures included four versions of the biodata instrument, a job knowledge test, an essay examination, and a verbal ability test, and were administered in that order to all participants in the study. The essay examination required the examinee to respond to one of eight questions assigned randomly. The questions related in a general way to the content of the jobs for which these examinations were designed, but they required no specialized job knowledge. Most essay topics required the candidate to take and defend a position on an issue in current economic, social, or political events. Individuals were allowed 50 minutes to complete their essays. Each essay was scored on a 1–4 point basis by two raters. Any disagreements on the rating of an essay required reconciliation between the judges, who were forced to agree on a rating. The score assigned to the essay was the sum of the two raters' judgments, so the available total scores were 2, 4, 6, and 8 only. Because independent ratings of the essays were not recorded, interrater agreement/reliability indices could not be computed.

The verbal ability test consisted of 125 items that attempted to measure basic grammar and usage, punctuation, writing style, organization, and sentence structure. The test consisted of text with underlined segments containing potential errors in grammar, punctuation, clarity, and so forth. For each underlined segment, four alternative ways to express the underlined portion were presented. Each choice represented an alternate method of expression or interpretation. A total of 105 minutes was allowed for completion of this portion.

The job knowledge test consisted of 165 items derived directly from knowledge statements that were judged by job experts to be required at entry to these jobs. All were multiple choice items with four options. The time limit on this section was 105 minutes.

The central focus of our study was four forms of a biodata instrument designed to measure a variety of noncognitive skills judged to be important to successful job performance in the target jobs, including interactions with others, initiative or persistence, adaptability, conflict resolution, leadership, stress tolerance, oral communication skills, and planning or prioritizing. Each form consisted of 91 items; each candidate received one of the four forms on a random basis. The items themselves were identical in all forms. The forms differed in terms of the number and placement of follow-up questions wherein examinees were required to provide an elaboration of their item responses (see examples in Appendix). In Form 1, no elaboration was required; all responses were made on five option scales that represented continua rather than discrete options. Examinees responded to the same 91 items with the same

options in the second form as well, but in addition, they were required to provide elaborations of their answers to 17 of the first 46 items. These 17 items were scattered throughout the first half of this form. The third form also required answers to the same 91 items with no elaboration of items in the first half of the test, but 18 items scattered throughout the second half required elaboration. The fourth form required answers to the 91 items and elaboration of the full set of 35 items required of respondents to Forms 2 and 3. Of the 311 total participants, 77, 79, 77, and 78 responded to Forms 1, 2, 3, and 4 respectively. Several different scores were computed based on the items in these forms, including total scores on all 91 items, scores on each half for the elaborated items, and scores on each half for the nonelaborated items. These five scores allowed us to compare responses to the same and different items from examinees who were and were not asked to elaborate their responses.

*Data Analyses*

Basic descriptive analyses were conducted to determine scale reliability, means, standard deviations, and intercorrelations. The three primary hypotheses were addressed using a 2 (elaboration required vs. no required elaboration) by 2 (test half) analysis of variance. The five scores described above were the dependent variables in these analyses of variance. Finally, the correlations between the biodata scores and the scores on the essay, verbal ability, and job knowledge tests were computed to determine whether scores on elaborated and nonelaborated items were related to these alternate measures.

## *Results*

In Table 1, we present the means and standard deviations for each of the four conditions for all five biodata scores. Coefficient alpha reliabilities and intercorrelations for these five indices and gender and education are presented in Table 2. Gender was not related to any of the test measures, but education was correlated significantly ($p < .05$) with all tests ranging from .20 with the total of the biodata items to .41 with job knowledge. Race differences between Caucasians and African Americans were .25 standard deviations on the biodata measure and between .86 and .94 on the other three measures. Differences between Caucasians and the other two sizable subgroups (Asian and Hispanic Americans) were approximately half that large. Caucasian means were the highest in all cases.

As can be seen in Table 2, all four subsets of items as well as the total scores on the biodata instrument exhibited acceptable levels of reliabil-

TABLE 1

*Means and Standard Deviations of the Dependent Variables[a] for the Four Test Forms*

| Test form | | Total | Elab items-1 | Elab items-2 | Non-elab items-1 | Non-elab items-2 |
|---|---|---|---|---|---|---|
| No elabs | *M* | 322.68 | 59.25 | 58.80 | 107.99 | 96.49 |
| | *SD* | 45.46 | 12.41 | 10.73 | 13.70 | 12.56 |
| Elab half 1 | *M* | 304.00 | 50.80 | 55.86 | 104.08 | 93.01 |
| | *SD* | 40.91 | 12.54 | 10.16 | 11.98 | 11.72 |
| Elab half 2 | *M* | 314.91 | 59.33 | 52.43 | 108.23 | 94.07 |
| | *SD* | 32.81 | 10.14 | 9.72 | 9.97 | 9.49 |
| Elab half 1 and 2 | *M* | 295.93 | 50.50 | 48.55 | 105.63 | 92.08 |
| | *SD* | 36.03 | 12.84 | 10.92 | 10.95 | 12.00 |

*Notes*: [a]Total is the score on all 91 biodata items; elab items-1 and elab items-2 refer to the scores on items that were elaborated on in the first and second halves of some forms; non-elab items-1 and non-elab items-2 are scores on nonelaborated items in the two halves of the test.

TABLE 2

*Correlations between Scores on Various Test Item Composites, Essay, Verbal Ability, Job Knowledge Tests, Gender, and Education[a]*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Elab items-1 (1) | .85[c] | | | | | | | | | |
| Elab items-2 (2) | .70[b] | .81 | | | | | | | | |
| Non-elab items-1 (3) | .60 | .60 | .83 | | | | | | | |
| Non-elab items-2 (4) | .60 | .68 | .81 | .81 | | | | | | |
| Total (5) | .85 | .86 | .87 | .89 | .94 | | | | | |
| Job knowledge (6) | .32 | .22 | .20 | .18 | .27 | .91 | | | | |
| Verbal (7) | .32 | .22 | .21 | .20 | .28 | .72 | .98 | | | |
| Essay (8) | .16 | .12 | .14 | .12 | .15 | .39 | .51 | – | | |
| Gender (9) | .07 | .02 | .00 | −02 | .02 | −.16 | .09 | −.04 | – | |
| Education (10) | .22 | .15 | .15 | .13 | .20 | .41 | .30 | .30 | −.09 | – |

[a]Total is the score on all 91 biodata items; elab items-1 and elab-items-2 refer to the scores on items that were elaborated on in the first and second halves of some forms; non-elab items-1 and non-elab items-2 are scores on nonelaborated items in the two halves of the test.

[b]All correlations above .11 are statistically significant, $p < .05$.

[c]Diagonal contains coefficient alpha reliabilities for all tests except Essay and the gender and education variables for which no reliability data are available. Gender was coded 1 for females, 2 for males. Education codes ranged from 1 (high school) to 7 (doctorate).

ity. The means in Table 1 indicate that when there was no elaboration of items, the total score and all four subtest scores were the highest. When elaboration of items was required, scores on the total test and the elaborated item sets were the lowest. If we examine the second column of numbers in Table 1, we can see that elaboration on the 17 items in the first half of the test produced a mean score between eight and nine points lower (equal to a standardized mean difference, *d*, of approximately .7)

than when examinees responded to the same items in their nonelaborated form. When we examine scores on the 18 elaborated items in the second half of the test, we see a 10 point difference in means (58.80 vs. 48.55, $d = .8$) for the form in which no elaboration occurred versus the form for which elaboration of items occurred in both halves. For the case in which the first half items were elaborated, but there was no elaboration of the targeted items on the second half, there was some "carry over" on these items in that the means were 58.80 versus 55.86 for the targeted items. When elaboration occurred in the first half, but not the second half, there was also evidence of some carry-over impact on the nonelaborated items in the first half of the test (107.99 vs. 104.08) as well as the nonelaborated items on the second half of the test (means were 96.49 vs. 93.01). The transfer of the impact of requiring elaboration on some items to responses to items that were not elaborated was about 1/3 to 1/2 as large as the impact on the scores for elaborated items themselves. Because the number of nonelaborated items was larger than the number of elaborated items, the magnitude of the impact was comparatively smaller.

More formal tests of our three hypotheses and assessments of the magnitude of effects are presented in Table 3. Table 3 is a presentation of the results of the two (elaboration of items in the first half of the test vs. no elaboration in the first half) by two (elaboration of items in the second half of the test vs. no elaboration in the second half) analyses of variance for the five outcome variables. These outcomes are obviously correlated (see Table 2), so these analyses of variance are partially redundant. For the total test score, the effect of elaboration of items in the first half is statistically significant ($p < .05$). This is also true for each of the other four outcomes as well.

These results are supportive of Hypotheses 1, 2, and 3. Not only are scores on the elaborated items lower, but so are scores on nonelaborated items on both halves of the test. The effect sizes differ, however, as was observed above. For the total test, $d = .46$ when elaboration of items is required in the first half of the test. Considering scores on just those items on which elaborated responses were required, the effect sizes were .68 (the first half) and .62 (the second half). Carry-over effects are much smaller; standardized mean differences between scores on nonelaborated items on forms in which elaboration occurred in the first half of the test were .28 and .23. Similarly, the effect of elaboration in the first half on elaborated items in the second half produced a standardized mean difference of .31.

The impact of elaboration in the second half of the test is less clear. The main effect of elaboration for the 18 items in the second half of the test produced a statistically significant effect only for the items for

TABLE 3

*Analyses of Variance of the Effects of Elaboration on the Different Scores*[a]

| | | Total | | Elab items-1 | | Elab items-2 | | Non-elab items-1 | | Non-elab items-2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Effect | *df* | *F* | *d*[b] | *F* | *d* | *F* | *d* | *F* | *d* | *F* | *d* |
| Elab–half 1 | 1 | 15.99* | .68 | 8.22* | .31 | 5.89* | .28 | 4.24* | .23 | | |
| Elab–half 2 | 1 | 2.83 | | .01 | | 32.95* | .62 | .45 | | 1.60 | |
| Elab 1 × elab 2 | 1 | .00 | | .02 | | .16 | | .24 | | .32 | |
| Error | 283 | (1587.86)[b] | | (144.73) | | (107.90) | | (137.56) | | (132.14) | |

[a]Total is the score on all 91 biodata items; elab items-1 and elab items-2 refer to the scores on items that were elaborated on in the first and second halves of some forms; non-elab items-1 and non-elab items-2 are scores on nonelaborated items in the two halves of the test.

[b]*d* refers to the standardized mean difference between the groups relevant to a given *F* test. No *d*s are presented for nonsignificant effects. Values in parentheses are *MS* error values.

*$p < .05$.

which elaboration was required. As noted above, the size of the effect (.62) is similar to that observed for the elaborated items in the first half of the test. Elaboration of these items has no impact on the first half scores for either elaborated or nonelaborated items and no effect would be expected. However, the effect of required elaboration in the second half on nonelaborated items on the second half of the test is not statistically significant and the effect on total test score is marginally significant ($p < .10$). So it appears that without elaboration on the first half of the test, the effects of elaboration on the second half occurs only on the elaborated items themselves. Taken as a whole, the results provide strong support for the first hypothesis, but somewhat less support for Hypotheses 2 and 3. There were definite carry-over effects for elaboration requirements placed in the first half of the test, but little evidence for carry over when items were elaborated only in the second half of the test. It may be that the elaboration manipulation is simply not powerful enough after examinees have answered close to 50 nonelaborated items.

Table 2 contains information relevant to questions about the degree to which abilities or knowledge on other job relevant measures might be related to the impact of required elaboration. Correlations between an essay exam, a verbal ability test, and a job knowledge test and the biodata measures are contained in the last three rows of this table. As can be seen, there are significant, but relatively low, correlations between each of these other tests and the biodata scores. As we speculated might be the case, correlations between these other tests and scores on elaborated items are higher than the correlations between these tests and the nonelaborated items, but none of these differences are statistically significant. If the more able examinees as measured by the job knowledge and English exams get better scores on the elaborated items as we speculated might be the case, it is not obvious in these correlational data. Moreover, an examination of the same correlations for the condition in which there was no elaboration and the condition in which there was elaboration in both halves produced the same pattern of correlations. It appears that the impact of elaboration on the relationship of biodata with other components of the test was minimal or nonexistent.

### *Discussion*

The results of this study demonstrate that the requirement that examinees elaborate on their responses to biodata test items reduces their scores on those elaborated items about .6 standard deviation units. This change is approximately equal to the difference between the "honest" responses" and "fake good" responses in previous research (e.g., Ellingson et al., 1999; Ones et al., 1996). It also seems that the effect of requir-

ing elaboration reduces scores on nonelaborated items in the same test, though this transfer effect is much smaller. If this research is replicated in other situations and for other measures, it would seem that requiring examinees to elaborate on their answers to noncognitive test items may be an alternative way to reduce socially desirable responding. However, we have no data that indicate the differences in conditions are the result of differences in socially desirable responding or faking good at this point. As mentioned in the introduction, the differences could be due to other factors such as fatigue or lack of motivation. Fatigue was not a likely explanation because the biodata measure was the first measure taken by examinees and examinees appeared highly motivated by the monetary incentive.

The interpretation of these results should be moderated by the fact that this is a pilot examinee group whose motivation for taking the test is money rather than the prospect of employment. However, in an attempt to more closely approximate the motivation level of actual candidates, these examinees were offered a bonus for "accurate and complete" responses to the instrument. Examinees were also told that those who score high overall with accurate and complete answers would be given an additional $10. This incentive might have produced different motivational sets: (a) to score high; (b) to give accurate responses; or (c) to give complete responses. These different possible motivational sets may have produced different elaboration behaviors, but we have no way of ascertaining what those differences might have been. In addition, examinees were warned that responses were subject to verification. The effects of the elaborations evaluated in this study may be somewhat more notable because they occur in the presence of this warning.

This research should be replicated and extended in several different ways to determine the potential of this method to control for socially desirable responding. First, attempts to generalize this research to other test items and test types should be made. Verification questions for some items may be impossible to write or request in a credible manner. This would be true of many attitude or value statements such as appear on some personality tests. Verification requests would seem credible on most biodata items. One study might be to ascertain the impact of requiring elaborations on items that vary in the ability of an organization or another individual to verify statements made in response to elaboration requests.

A second question that should receive attention would be to assess further the impact of elaboration requests on items for which elaboration is not requested. How many or what proportion of items should require elaboration to discourage socially desirable responses? This study suggests that such requests placed at the beginning of a test are more

effective than elaboration requests on items that appear later in the test. Questions of placement of items, as well as the type of items for which elaboration is requested, should be further delineated.

On a more theoretical level, it would be interesting and perhaps helpful to know more about the construct validity of this manipulation. Is it the result of a suppression of the tendency to manage impressions or is it the result of helping the person remember better and to self-evaluate more realistically (Paulhus, 1984; 1991)? Correlating the scores on elaborated and nonelaborated versions of a noncognitive test with measures of social desirability and integrity test scores may provide useful information about the psychological meaning of response changes. Likewise, comparing the impact of elaboration requests when the outcome of one's response has significantly different consequences for high and low scorers may be informative.

A very important practical question that should be addressed is the degree to which elaborated and nonelaborated responses display the same degree of criterion-related validity. Our results indicate that correlations with other ability tests and job knowledge are similar. Ones et al.'s (1996) meta-analysis also indicates that other attempts to correct for social desirability do not change the validity of noncognitive measures. This question also needs to be addressed in the case of the proposed solution described in this study. In one attempt to discover the nature of the items on which the elaboration manipulation either did or did not produce different scores, we computed the mean difference between elaborated and nonelaborated versions of the same item. We then rank ordered these items in terms of this difference. Items on which elaboration produced the least difference included the following: times you attended a conference, seminar, or workshop as a means of gaining new skills, number of languages other than your native language you have attempted to learn, number of times you have made travel arrangements for yourself or a group, number of times you have helped a coworker or team member get acclimated to a group, the number of jobs or previous positions in which customer service skills have been required, and the number of leadership positions held in the last 5 years. By contrast, those items on which the difference between elaborated and nonelaborated items was the greatest included the following: the number of times served as a spokesperson for a group, number of projects in which you were required to meet multiple deadlines, the number of times you were required to assume authority to improve a work group's efficiency, and the number of times you were able to solve a dispute with a store or business. The first set of items may be more concrete in the action required, allowing for fewer alternate interpretations of one's role in the activity or perhaps less interpersonal in their nature. Perhaps the first

group of items are task oriented in nature and the latter are more social or interpersonal.

It is also possible that respondents will need to be convinced that a potential employer or user of the biodata responses is prepared to do the follow-up work necessary to verify the responses of the participants to realize the benefits of such elaboration. Answers to these questions will provide evidence as to whether the apparently positive impact of elaboration requests observed in this study will provide a more generally useful approach to the problem of socially desirable responding on noncognitive tests.

## REFERENCES

Becker TE, Colquitt AL. (1992). Potential versus actual faking of a biodata form: An analysis along several dimensions of item type. PERSONNEL PSYCHOLOGY, *45*, 389–406.

Cascio WF. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology, 60*, 767–769.

Crowne DP, Marlowe D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354.

Douglas EF, McDaniel MA, Snell AF. (1996, August). *The validity of non-cognitive measures decays when applicants fake*. Paper presented at the annual meeting of the Academy of Management, Cincinnati, OH.

Dwight SA, Donovan JJ. (1998, April). *Warning: Proceed with caution when warning applicants not to dissimulate*. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Ellingson JE, Sackett PR, Hough LM. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155–166.

Fiske ST, Taylor SE. (1991). *Social cognition*. New York: McGraw-Hill.

Goldstein IL. (1971). The application blank: How honest are the responses? *Journal of Applied Psychology, 55*, 491–492.

Guzzo RA, Salas E (Eds.). (1995). *Team effectiveness and decision making in organizations*. San Francisco: Jossey-Bass.

Hackman JR (Ed.). (1991). *Groups that work (and those that don't)*. San Francisco: Jossey-Bass.

Hough LM. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69*, 135–146.

Hough LM. (1998). The millennium for personality psychology: New horizons or good ole daze. *Applied Psychology: An International Review, 47*, 233–261.

Hough LM, Eaton NK, Dunnette MD, Kamp JD, McCloy RA. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581–595.

Hough LM, Paullin C. (1994). Construct-oriented scale construction: The rational approach. In Mumford MD, Owens WA (Eds.), *The biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 109–145). Palo Alto, CA: Consulting Psychologists Press.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–95.

Jackson DN, Wroblewski VR, Ashton MC. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388.

Keating E, Paterson DG, Stone CH. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34*, 6–11.

Klein SP, Owens WA. (1965). Faking of a scored life history blank as a function of criterion objectivity. *Journal of Applied Psychology, 49*, 452–454.

Kluger AN, Reilly RR, Russell CJ. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76*, 889–896.

Lautenschlager GJ. (1994). Accuracy and faking of background data. In Stokes GA, Mumford MD, Owens WA (Eds.), *Biodata handbook* (pp. 391–419). Palo Alto, CA: Consulting Psychologists Press.

Mosel JM, Cozan LW. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 36*, 365–369.

Mumford MD, Owens WA. (1987). Methodology review: Principles, procedures, and findings in the applications of background data measures. *Applied Psychological Measurement, 11*, 1–31.

Ones DS, Viswesvaran C, Korbin WP. (1995, April). *Meta-analysis of fakability estimates: Between-subjects versus within-subjects designs*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Ones DS, Viswesvaran C, Reiss AD. (1996). Role of social desirability in personality testing for personnel selection: A red herring. *Journal of Applied Psychology, 81*, 660–679.

Paulhus DL. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

Paulhus DL. (1991). Measurement and control of response bias. In Robinson JP, Shaver PL, Wrightsman L, Andrews FM (Eds.), *Measures of personality and social-psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Reilly RR, Chao GT. (1982). Validity and fairness of some alternate selection procedures. PERSONNEL PSYCHOLOGY, *35*, 1–62.

Rosse JG, Stecher MD, Miller JL, Levin RA. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Ryan AM, Ployhart RE, Greguras GJ, Schmit MJ. (1998). Test preparation programs in selection contexts: Self-selection and program effectiveness. PERSONNEL PSYCHOLOGY, *51*, 599–622.

Ryan AM, Sacco JM, McFarland LA, Kriska SD. (2000). Applicant self-selection: Correlates of withdrawal form a multiple hurdle process. *Journal of Applied Psychology, 85*, 163–179.

Schmidt FL, Hunter JE. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmitt N, Clause C, Pulakos ED. (1996). Subgroup differences in ability as assessed by different methods. In Cooper CL, Robertson I (Eds.), *International review of industrial and organizational psychology* (pp. 115–140). New York: Wiley.

Schmitt N, Jennings D, Toney R. (1999). Can we develop measures of hypothetical constructs? *Human Resource Management Review, 9*, 169–184.

Schrader AD, Osburn HG. (1977). Biodata faking: Effects of induced subtlety and position specificity. PERSONNEL PSYCHOLOGY, *30*, 395–404.

Snell AF, Sydell EJ, Lueke SB. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resources Management Review, 9*, 219–242.

Stokes GS, Toth CS. (1996). Background data for personnel selection. In Barrett RS (Ed.), *Fair employment strategies in human resource management* (pp. 171–179). Westport, CT: Quorum.

Thornton GC, Gierasch PF. (1980). Fakability of an empirically derived selection instrument. *Journal of Personality Assessment, 44*, 48–51.

Wiggins JS. (1959). Interrelationships among MMPI measures of simulation under standard and social desirability instructions. *Journal of Consulting Psychology, 23*, 419–427.

## APPENDIX

### *Examples of Elaborated Biodata Items*

A. How many work groups have you led in the past 5 years?

1. 0
2. 1
3. 2
4. 3
5. 4 or more

If you answered 2 to 5 above, briefly describe the work groups and projects you led.

_____

_____

_____

_____

_____

B. How often have you rearranged files (business, computer, personal) to make them more efficient in the last year?

1. Very frequently
2. Often
3. Sometimes
4. Rarely
5. Never

If you answered 1 to 3 above, list the approximate dates and how much time you spent on this task each time.

_____

_____

_____

_____

_____

C. In how many of your previous jobs have you had to interact extensively (an hour or more per day) with clients or customers?

1. 0
2. 1
3. 2
4. 3

5. 4 or more

If you answered 2, 3, 4, or 5, describe the nature of this contact(s) briefly for each job. Describe no more than four.

_____
_____
_____
_____
_____

D. How many software packages have you used to analyze data?

1. 0
2. 1
3. 2
4. 3
5. 4 or more

If you answered 2, 3, 4, or 5, indicate the software programs and the nature of the data analysis briefly. Include no more than 4.

_____
_____
_____
_____
_____