Journal of Applied Psychology
2003, Vol. 88, No. 6, 979–988

0021-9010/03/$12.00 DOI: 10.1037/0021-9010.88.6.979

Impact of Elaboration on Socially Desirable Responding and the Validity of Biodata Measures

Neal Schmitt, Fred L. Oswald, Brian H. Kim,
Michael A. Gillespie, and Lauren J. Ramsay
Michigan State University

Tae-Yong Yoo
Kwangwoon University

The current study investigated the impact of requiring respondents to elaborate on their answers to a biodata measure on mean scores, the validity of the biodata item composites, subgroup mean differences, and correlations with social desirability. Results of this study indicate that elaborated responses result in scores that are much lower than nonelaborated responses to the same items by an independent sample. Despite the lower mean score on elaborated items, it does not appear that elaboration affects the size of the correlation between social desirability and responses to biodata items or that it affects criterion-related validity or subgroup mean differences in a practically significant way.

In the last decade, industrial–organizational psychologists have been increasingly interested and involved in expanding the performance domain considered in personnel selection research and in studies of work performance in general (Borman & Motowidlo, 1997; Campbell, Gasser, & Oswald, 1996; Campbell, McCloy, Oppler, & Sager, 1993). Similar considerations seem to be motivating research on the performance of students in academic settings (Taber & Hackman, 1976; Willingham, 1985). Organizational researchers are more frequently considering aspects of what has become known as the contextual domain, which includes social responsibility and helping behavior. Relatedly, academic researchers interested in predicting academic success are beginning to consider such performance domains as student leadership, multicultural awareness, and civic responsibility. In both cases, noncognitive constructs and measures are more likely to be useful predictors of these aspects of performance.

Whereas some researchers and theoreticians have advocated expanding the performance domain, others have demonstrated the utility of noncognitive predictors of performance (Barrick & Mount, 1991; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Mumford & Stokes, 1992). In addition to having obvious relevance to the noncognitive constructs being considered in this broadened performance domain, noncognitive predictors tend to produce smaller subgroup race differences in mean scores than do more traditional cognitive ability predictors that relate validly to task performance as well as to contextual performance (Schmitt, Clause, & Pulakos, 1996). If contextual performance is relevant to an organization, weighting it more highly in the composite used to validate predictors tends to result in a different set of valid predictors, smaller subgroup differences, and less adverse impact on protected groups (Bobko, Roth, & Potosky, 1999; Hattrup, Rock, & Scalia, 1997; Murphy & Shiarella, 1997; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997).

One continuing problem with the use of some of these noncognitive measures has been that the correct, job-related, or preferred answer to the questions is usually obvious. Furthermore, when used in high-stakes testing situations, the motivation to “fake good” or give distorted answers to these questions can be significant. Research has supported the hypothesis that mean differences between applicants and incumbents on noncognitive measures can often be substantial (see Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Jackson, Wroblewski, & Ashton, 2000; Rosse, Stecher, Miller, & Levin, 1998) and that decisions based on such tests may vary as a function of response distortion (Douglas, McDaniel, & Snell, 1996). Several methods designed to measure and control for response distortion have been studied. The purpose of this article is to describe a recent effort to replicate Schmitt and Kunce’s (2002) results obtained when respondents were required to elaborate on their answers to biodata questions, to provide further data on the relationship of elaborated answers to measures of social desirability, and to compare the validity of elaborated and nonelaborated answers.

Efforts to Reduce Response Distortion

Most of the early research on faking or response distortion was done using personality measures. In reviewing this literature, Paulhus (1984, 1991) has claimed that response distortion can be both deliberate (i.e., impression management) or the result of self-deception, and he constructed an instrument that measures both of these dimensions. Typically, researchers measure the respondent’s impression management or self-deception and use those measurements to control for social desirability statistically through partial correlation analysis. Controlling for social desirability in predictors in this fashion should produce partial correlations that are higher than the corresponding bivariate correlations (i.e., there is a classic suppressor effect). This attempt to control

Neal Schmitt, Fred L. Oswald, Brian H. Kim, Michael A. Gillespie, and Lauren J. Ramsay, Department of Psychology, Michigan State University; Tae-Yong Yoo, Department of Psychology, Kwangwoon University, Seoul, South Korea.

We acknowledge the financial support of the College Board, who sponsored this research.

Correspondence concerning this article should be addressed to Neal Schmitt, Department of Psychology, Michigan State University, East Lansing, Michigan 48824-1117. E-mail: schmitt@msu.edu

for the effects of faking has the most history (Crowne & Marlowe, 1960; Wiggins, 1959), but Ones, Viswesvaran, and Reiss (1996) found no evidence to support corrections for social desirability on the basis of a meta-analytic investigation. In support of Ones et al.'s conclusion, Ellingson, Sackett, and Hough (1999) asked examinees to respond twice to a biodata measure of personality (Assessment of Background and Life Experiences; Hough et al., 1990), once honestly and once under instructions to fake. The standardized mean difference in scores across 11 scales ranged from 0.31 to 0.86 standard deviations. Correcting for measured social desirability erased most of these differences on average, but the corrections had no impact on who would have been selected if these tests had been used to identify the most competent people.

In reviewing the literature on the faking of biodata instruments that are the focus of this study, Lautenschlager (1994) concluded that biodata, like personality measures, were fakable. He also concluded that faking occurred even when the items were relatively verifiable and that attempts to control for faking had been unsuccessful. Kluger, Reilly, and Russell (1991) examined the susceptibility of items to faking under different methods of scoring biodata items. They found that when items were scored as Likert items, there was faking in a socially desirable direction, but that faking was not evident in items in which each option was scored separately. The difference between items in which each option was scored and Likert items was close to one standard deviation, but items reflecting attitudes or values cannot often be scored in this way. Finally, studies have shown that giving examinees a warning about the consequences of faking and the possibility of detection has been effective in reducing socially desirable responding (e.g., Dwight & Donovan, 1998).

Recently, Schmitt and Kunce (2002) have proposed another method to attempt to control for the inflation that may occur as a result of conscious self-deceptive distortions. Based on the finding that response distortion on application blanks is more likely on questions that are subjective and cannot be verified (Becker & Colquitt, 1992), their method requires examinees to elaborate on their answers to some of the biodata questions. The hope was that elaboration would create a respondent frame for less distortion on all items, including both those items requiring elaboration as well as those that did not require elaboration. Two examples of elaborated items are presented in the Appendix. Research in social cognition also indicates that people overstate their abilities when they believe their answers cannot be verified (Fiske & Taylor, 1991). Schmitt and Kunce found that performance on items for which some elaboration was required was 0.7 to 0.8 standard deviations lower than on items for which no elaboration was required. There was some carryover to nonelaborated items on the same biodata form, but the effect was only one third to one half as large as the effect for elaborated items.

Although the Schmitt and Kunce (2002) results are encouraging, there are several unanswered questions. First, the authors recognized that the attribution that lowered scores are a function of the removal of social desirability is untested. Scores on elaborated items may be lower for other reasons, such as unwillingness to elaborate answers, fatigue, or an inability to recall specific instances of a behavior. If requiring elaboration reduces the impact of social desirability on the answers to biodata or personality items, then the correlation of measures of social desirability with elaborated biodata items should be lower than the correlation of social desirability with nonelaborated versions of the same items as well as other nonelaborated items. We argue that this is true for trait measures of social desirability like those of Paulhus (1991) because respondents to elaborated items should recognize the potential of verification of their responses and control their natural inclination to inflate responses. Second, Schmitt and Kunce did not have performance outcome data in their study, so they could not test the impact that elaboration of items might have on the validity of those items. Third, there is some evidence that elaborations required in the context of completing accomplishment records (Hough, 1984) differentially discourage the response of members of various subgroups (Ryan, Ployhart, Greguras, & Schmit, 1998; Ryan, Sacco, McFarland, & Kriska, 2000). We are thus interested in evaluating the impact of elaboration on the responses of members of different racial/ethnic subgroups. It is our intent in this article to replicate the results of the Schmitt and Kunce study and to attempt to answer the three questions above. Specifically, we hypothesize the following:

> *Hypothesis 1:* Elaborated items will produce lower mean responses than nonelaborated versions of the same items.
>
> *Hypothesis 2:* Mean responses to nonelaborated items in the same form that includes elaborated items will be lower than responses to the same items in a form that does not include elaboration.
>
> *Hypothesis 3:* Nonelaborated items will correlate more highly with social desirability and impression management than will (a) elaborated versions of the same items or (b) other items that are not elaborated but that are on the same test as the elaborated items.
>
> *Hypothesis 4:* Elaborated items will correlate no differently with outcome measures than will nonelaborated versions of the same items.
>
> *Hypothesis 5:* There will be no significant subgroup mean differences in the impact of elaboration of items.

Tests of Hypotheses 1 and 2 represent a replication of the Schmitt and Kunce work with a different group of respondents; tests of Hypotheses 3–5 represent an extension of that work considering new questions about the meaning of differences produced by elaboration and the impact elaboration has on validity and subgroup differences.

Method

Participants

Six hundred fifty-four first-year undergraduate students at a large Midwestern university volunteered for this study and received $40 each for their participation. Of these, 644 provided usable data after various screens for careless responses. Mean age was 18.45 years ($SD = 0.69$). Seventy-two percent were women. Seventy-nine percent were Caucasian; 9.5%, African American; 2.3%, Hispanic American; 5.3%, Asian American; and 3.9%, other. First-year students were recruited through their classes, their housing units, and the student newspaper. All measures were administered as part of a larger test battery in small group administrations ($M = 15.19$, $SD = 8.12$ participants in each group). Trained proctors adhered to a script and read test instructions verbatim, similar to standardized test procedures.

The motivation of participants to answer carefully and thoughtfully was important to the project, and several precautions were taken to maximize participant cooperation. First, participants were each paid $40 for less than 4 hr of their time. For undergraduates, this compensation was considered generous. Second, all participants were told of the purpose of the project:

> To test whether measures of judgment and background are related to your grades and other activities at MSU. . . . Because a major purpose of our study is to determine if your responses to the judgment and background measures are related to your performance as a student at MSU, we are asking your permission to allow the registrar to give us access to your grades . . . .

Obviously, we could not replicate the situation in which students were applying to college, but we encouraged participants to answer as seriously and honestly as would a college applicant. They were asked to consider their behavior as high school students in answering the questions. Third, all participants were given three breaks during the period to minimize the impact of fatigue. Fourth, the measures used in this study were part of a larger package of instruments to which examinees were asked to respond, but the biodata items were the first of this larger set of instruments to which the participants responded. Fifth, we used a carelessness scale to eliminate 10 of the 654 participants whose responses indicated a lack of attention to the questionnaire.

Participants were randomly assigned to groups that completed either the elaboration or nonelaboration forms of the biodata instruments that are explained below. The two biodata forms were assigned randomly at the group level rather than at the individual level, so that within each group no one would notice that some participants were doing substantially more or less writing than others.

Measures

The measures included two versions of a biodata instrument constructed to measure 12 dimensions of student performance. We developed these dimensions by examining the University Residence Life materials at the university attended by the participants and by interviewing a senior-level administrator in the residence halls. In addition, we gathered information on expectations, ideas, and definitions of college student performance taken from published educational literature, national educational reports, and college mission statements posted on the Internet. We independently sorted this information into rationally determined dimensions of college success and, after discussing the results of the sorting procedure, agreed that 12 dimensions represented the college student performance domain:

(1) knowledge, learning, and mastery of general principles;

(2) continuous learning, intellectual interest, and curiosity;

(3) artistic and cultural appreciation;

(4) multicultural tolerance and appreciation;

(5) leadership;

(6) interpersonal skills;

(7) social responsibility, citizenship, and involvement;

(8) physical and psychological health;

(9) career orientation;

(10) adaptability and life skills;

(11) perseverance; and

(12) ethics and integrity.

The development of these dimensions and the measures described below is detailed in two reports (Gillespie et al., 2002; Manheim et al., 2002). Although the dimensions related to college student performance, our questions and instructions requested that participants consider their previous experiences, including those in high school. Because participants were all first-year college students, some of the experiences on which they reported may have occurred in college.

We adapted biodata items reflecting these 12 dimensions from existing measures, and we wrote additional items to reflect these dimensions. For those existing item stems with response options that did not seem relevant to first-year college students, a sample of paid college-student participants were asked to write open-ended responses. We then piloted the complete set of items and student-generated responses, dropping those items that showed little variance. To obtain support for the dimensionality of the items, we resorted a randomized list of items back into the original dimensions. Items on which five of the six of us agreed with the original assignment of the item to a dimension were retained; those on which four of the six of us agreed were discussed and rewritten or discarded, and those with fewer than four of us in agreement were discarded. Items assigned to a different dimension by all six of us were reassigned to the new dimension. Using these criteria, we did not assign 11 items to any of our 12 dimensions. We retained them in the pool of nonelaborated items for exploratory purposes.

Two different forms of the same biodata items resulted from this process. The two forms consisted of 126 items and were identical, with the exception that the elaborated form contained some items that required respondents to provide written support for their multiple-choice responses (e.g., “If you answered b, c, d, or e, please list the student offices you held in high school”). Twenty-one items were selected for elaboration on the basis of their nature. Because items that referred to attitudes or opinions would require in-depth explanations, we requested elaboration of items that required a discrete answer and appeared to be verifiable. All biodata items were scored continuously on 4- or 5-point scales. For purposes of the analyses conducted below, we computed scores on the elaborated items and the nonelaborated items in both forms of the tests. Alpha coefficients for the elaborated and nonelaborated item sets for both groups of respondents are reported in Table 1.

Social desirability was measured by using a modified version of the Paulhus (1991) measure. This measure produces scores for social desirability, viewed as self-deception by Paulhus and impression management, which represents a deliberate attempt to present oneself favorably. Because of concerns about the intrusive nature of one item in each scale (i.e., “I have sometimes doubted my ability as a lover” and “I never read sexy books or magazines”), 2 items were not used. Each scale still consisted of 19 items and displayed alpha coefficients of .62 for social desirability and .80 for impression management.

We also obtained the grade point averages (GPAs) of our participants for their first academic year from the registrar’s office, and we asked participants to self-report the number of classes missed during the previous term by using a 5-point scale ranging from *less than five times* to *more than 30 times.* Although actual class attendance was impossible to verify, we believe that students honestly reported their attendance, because of the relationship between self-reported and actual GPA. We asked the same students to report their GPAs on a relatively coarse 10-option scale, and the correlation between these self-reported GPAs and the actual GPAs was .91.

The final set of student outcomes was their self-appraisals of performance on 12 behaviorally anchored rating scales developed to reflect the same 12 performance dimensions mentioned above. Exploratory factor analyses of these ratings indicated that a general factor accounted for a large portion of the variance (24%) in these ratings and that multifactor solutions were uninterpretable. The alpha coefficient for the composite of the 12 ratings was .80, so a summed composite of the 12 ratings was used in the analyses described below.

Table 1
Means, Standard Deviations, and Alpha Coefficients for Elaborated and Nonelaborated Items

Form	No. of items	*M*	*SD*	α	95% CI
Elaborated form–elaborated items	21	2.46	0.61	.80	2.39–2.52
Elaborated form–nonelaborated items	105	3.19	0.33	.91	3.16–3.23
Nonelaborated form–elaborated items[a]	21	2.92	0.56	.83	2.86–2.98
Nonelaborated form–nonelaborated items	105	3.23	0.31	.91	3.20–3.27

Note. Means are item means. CI = confidence interval.
[a] Elaboration was not required on these items, but they were the same items for which elaboration was required on the other form.

Data Analyses

Hypotheses 1 and 2 were evaluated by using a 2 (one group of participants from whom elaboration was requested and a second group of participants who were not asked to elaborate items) × 2 (elaborated items vs. nonelaborated items) analysis of variance (ANOVA) with repeated measures on the second factor. Two hundred eighty-seven participants were required to produce elaborated responses to 21 items, whereas the second group of 335 was not required to elaborate any responses. This group variable was the first factor in the ANOVA. Hypotheses 3 and 4 were evaluated by using a test for the significance of the difference between correlations. Hypothesis 5 was evaluated by using a 2 (one group of participants from whom elaboration was requested and a second group of participants who were not asked to elaborate items) × 3 (racial/ethnic group) × 2 (elaborated items vs. nonelaborated items) ANOVA with repeated measures on the last factor.

Results

Mean Differences in Responses to Elaborated and Nonelaborated Items

The 2 (groups of participants who were and were not requested to elaborate items) × 2 (item type: elaborated vs. nonelaborated) ANOVA with repeated measures on the second factor described above revealed a significant group effect, $F(1, 637) = 31622.78$, $p < .01$; a significant item-type effect, $F(1, 637) = 978.51$, $p < .01$; and a significant interaction, $F(1, 637) = 153.86$, $p < .01$, of Group × Item type. The nature of the interaction effect is obvious in Table 1, in which we present the means, standard deviations, and alpha coefficients of the elaborated and nonelaborated items for participants who responded to the biodata forms with and without elaboration. As can be seen, requiring participants to elaborate on their answers to items produced very different means across the two groups of respondents. This difference is equal to .80 standard deviations and was statistically significant ($p < .05$). Comparison of participants' responses to nonelaborated items across groups produced a small mean difference of .04, corresponding to a trivial difference of .13 standard deviation difference in means, which was statistically nonsignificant ($p > .05$). These data support the conclusion that elaboration of responses tends to produce lower scores on the elaborated items (Hypothesis 1), but they do not support the conclusion that responses to nonelaborated items within the same form are affected in the same way (Hypothesis 2).

Correlations of Elaborated and Nonelaborated Items With Social Desirability

To evaluate the extent to which social desirability and impression management were correlated with responses to elaborated and nonelaborated items and to evaluate Hypothesis 3, we correlated scores on these two measures with both sets of biodata items in both groups of participants. Relevant correlations are contained in the last two rows and columns of Tables 2 and 3. The expectation was that, for elaboration items, correlations with social desirability measures would be higher when participants were not required to elaborate than when they were required to elaborate on the same items. Correlations for the self-deception measure are in the predicted direction but are not significantly ($p > .05$) different from each other ($r = .23$ vs. .13). The other pair of correlations pertaining to impression management is virtually the same ($r = .15$ vs. .16). It is interesting to note that the correlations of the two social desirability indices with the nonelaborated items are substantially and significantly ($p < .05$) higher whether items were contained in a form that involved the elaboration of items ($r = .40$ and .41) or not ($r = .46$ and .39). This suggests that the items for which we required an elaborated response were, by their very nature, less susceptible to the response biases measured by the two Paulhus (1991) indices.

This unexpected result and the comments of two anonymous reviewers stimulated several additional analyses. One possibility was that participants were not motivated to write the short descriptions required for the elaborated items. If this is the case, then there should have been more zero responses to items that required elaboration than there were to the same items when no elaboration was required. This was the case. On average, only 16% of the participants indicated zeros for their answers to the 21 elaborated items in their nonelaborated form. Thirty-five percent indicated zero when an elaborated answer to these items was requested. However, this result is also consistent with the notion that an elaboration request diminishes socially desirable responding. It is also true that an average of 65% of the participants did provide elaborated responses. In addition, an average of 2% more participants provided answers of "4 or more" to the 21 items in the elaborated condition than to those in the nonelaborated conditions.

We also examined the degree to which the objectivity or verifiability (Mael, 1991) of the items might be an alternate explanation of the differences observed for elaborated and nonelaborated items. Five of the authors and 4 research assistants independently

Table 2
Means and Standard Deviations of Grade Point Average (GPA), Absenteeism, Self-Rating, Self-Deception, Impression Management, and the Elaborated Biodata Items

	No elaboration group		Elaboration group							
Variable	*M*	*SD*	*M*	*SD*	1	2	3	4	5	6
1. Biodata	2.92	0.56	2.46	0.61	—	.14	−.08	.46	.13	.14
2. GPA	3.03	0.67	3.00	0.71	.16	—	−.53	.15	−.09	.22
3. Absenteeism	2.03	1.08	1.92	1.08	−.01	−.33	—	−.26	−.12	−.32
4. Self-rating	58.82	8.52	58.24	8.51	.42	.09	−.14	—	.15	.34
5. Self-deception	3.10	0.36	3.11	0.33	.23	.00	−.09	.26	—	.29
6. Impression Management	3.19	0.48	3.18	0.44	.16	.18	−.29	.29	.41	—

Note. Correlations below the diagonal are those for the group ($N = 350$) that was not requested to elaborate items. Correlations above the diagonal are those for the group ($N = 290$) that was asked to elaborate the same 21 items. Correlations greater than .12 above the diagonal are statistically significant ($p < .05$). Below the diagonal, correlations greater than .11 are statistically significant ($p < .05$). The biodata means are item means.

provided ratings of the objectivity and verifiability of each of the 126 biodata items on 5-point Likert-type scales. A comparison of the means of elaborated and nonelaborated items on these two indices indicated that the elaborated items were judged to be significantly ($p < .05$) more objective ($d = .93$) and verifiable ($d = .77$) than were the nonelaborated items. Composite verifiability and objectivity judgments were correlated .77. To further explore the degree to which correlations with self-deception and impression management might be a function of objectivity and verifiability, we computed the correlation between these two composites for each item and the correlations of each item response with self-deception and impression management. The correlations of objectivity judgments and item correlations with self-deception and impression management were −.37 and −.12, respectively. Similar correlations involving the verifiability judgments were −.38 and −.29. With the exception of the −.12 correlation, all four correlations were statistically significant ($p < .05$), though not great in magnitude. All four correlations indicate that the more objective and verifiable the item, the lower the correlation with the participants' self-deception and impression management scores.

Finally, to assess whether the responses of our participants to the self-deception and impression management indices were similar to other groups' responses to these measures, we examined data reported in Paulhus (1991). For self-deception, Paulhus quoted a study by Quinn (1989), who reported means of 7.6 and 7.3 for 884 male and female "religious adults." Paulhus (1988) reported male and female means on the same measure of 7.6 and 6.8 for a group of 443 college students. After we rescored them to make them consistent with Paulhus' scoring of the scales and adjusted for the elimination of one item from this scale, the average of our participants' self-deception responses was 7.5. On the impression management index, Quinn reported means of 7.3 and 8.9 for men and women, whereas Paulhus reported 4.3 and 4.9. The average of our respondents' impression management responses when we made the same adjustments for differences in scoring was 6.0. There were nonsignificant ($p > .05$) male–female differences equal to .17 and .26 in standard deviation units for self-deception and impression management, respectively. Responses to these measures among our respondents seem very similar to those of these earlier groups.

Table 3
Means and Standard Deviations of Grade Point Average (GPA), Absenteeism, Self-Rating, Self-Deception, Impression Management, and the Nonelaborated Biodata Items

	No elaboration group		Elaboration group							
Variable	*M*	*SD*	*M*	*SD*	1	2	3	4	5	6
1. Biodata	3.23	0.31	3.19	0.33	—	.12	−.21	.55	.40	.41
2. GPA	3.03	0.67	3.00	0.71	.26	—	−.53	.15	−.09	.22
3. Absenteeism	2.03	1.08	1.92	1.08	−.18	−.33	—	−.26	−.12	−.32
4. Self-rating	58.82	8.52	58.24	8.51	.61	.09	−.14	—	.15	.34
5. Self-deception	3.10	0.36	3.11	0.33	.46	.00	−.09	.26	—	.29
6. Impression management	3.19	0.48	3.18	0.44	.39	.18	−.29	.29	.41	—

Note. Correlations below the diagonal are those for the group ($N = 350$) that was not requested to elaborate items. Correlations above the diagonal are those for the group ($N = 290$) that was asked to elaborate the same 21 items. Correlations greater than .12 above the diagonal are statistically significant ($p < .05$). Below the diagonal, correlations greater than .11 are statistically significant ($p < .05$). The biodata means are item means.

Taken as a whole, these post hoc examinations of elaborated and nonelaborated items and the social desirability measures indicates that (a) the elaborated items were more objective and verifiable; (b) item objectivity and verifiability were moderately related ($< .40$) to self-deception and impression management correlations with item responses; (c) respondents were more likely to indicate zero when requested to elaborate, but the majority of the respondents still provided an elaborated response and many provided multiple elaborated responses; and (d) our participants' motivation to respond in a socially desirable manner as indexed by the Paulhus (1991) measures was similar to those of participants who were the subjects of analyses contained in earlier studies.

Validity of Elaborated and Nonelaborated Items

We also tested the degree to which elaboration affected the criterion-related validity of the biodata items. The first two columns and rows of Table 2 contain the correlations of the elaborated items under conditions in which elaboration was and was not required with GPA, class attendance, and the self-reported performance composite. None of the differences in validity coefficients between the two groups of respondents were statistically significant ($p > .05$). The correlation with GPA was actually larger when elaboration was required than it was when respondents gave answers to the same items without elaboration. The reverse was true for the self-rating composite, but not significantly so.

In Table 3, we present the same correlations for the biodata items for which no elaboration was required of either group of respondents. Any differences in validity coefficients across groups in this table would presumably be a function of some carryover effect, because no elaboration of these items was required. The correlation of biodata with GPA was somewhat higher (.26 vs. .12) for the group in which elaboration was required, but this difference was not significantly different ($p > .05$). Validity coefficients in this table are higher than those reported in Table 2, probably partly because the biodata composite in the case of the data reported in Table 3 consists of more items and is more reliable (see Table 1).

Subgroup Differences in Response to Elaborated and Nonelaborated Items

Hypothesis 5 was tested by using a 2 (one group of participants from whom elaboration was requested and a second group of participants who were not asked to elaborate items) × 3 (ethnic group status) × 2 (elaborated items vs. nonelaborated items) ANOVA with repeated measures on the last factor. Only Caucasians ($n = 505$), African Americans ($n = 59$), and Asian Americans ($n = 33$) were used in this analysis, as numbers of other participant groups were quite small ($n < 13$). The means and standard deviations for the three groups are presented in Table 4. As in the ANOVA presented above, the Group and Group × Item Type interaction was statistically significant ($p < .01$), but neither the Ethnic group factor nor any of the interactions with ethnic group were statistically significant ($p > .05$).

For both elaborated and nonelaborated cases, Table 4 shows that mean scores of Asian Americans were the lowest and that mean scores of Caucasians were the highest. Table 5 shows the magnitude of the mean differences for both sets of biodata items. The largest subgroup mean differences were equal to approximately one third of a standard deviation, with Asian Americans tending to score lower than Caucasians on both elaborated and nonelaborated items and African Americans tending to score higher than Asian Americans on nonelaborated items. The mean difference between African Americans and Caucasians on both elaborated and nonelaborated items was slight to nonexistent. Although racial/ethnic status was not significant in these analyses, it should be pointed out that sample sizes for all but the Caucasian group were quite small.

Although not of central concern in this study, the means and standard deviations for all variables in the study are presented in Table 4 as well. Tests of ethnic group differences on these variables revealed statistically significant ($p < .05$) differences on all but the impression management variable. African Americans and Asian Americans did worse on GPA than did Caucasians, but on the self-rating index, African Americans and Caucasians perceived themselves to be doing equally well relative to each other. Asian Americans rated their performance lower than did the other two groups. Asian Americans also indicated that they missed more classes than did the other two groups. On the self-deception index, African Americans appeared to be responding in a more socially desirable manner than did the other two groups. The African American and Asian American groups are too small to justify differential prediction analyses, but if these mean differences are representative of these subgroups, African American and Asian American GPAs would be overpredicted because the biodata mean differences are substantially smaller than GPA mean differences.

Table 4
Means and Standard Deviations for Predictors, Social Desirability Measures, and Outcomes for African American, Asian American, and Caucasian Subgroups

	African Americans		Asian Americans		Caucasians	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Elaborated biodata	2.63	0.58	2.55	0.56	2.73	0.62
2. Nonelaborated biodata	3.20	0.31	3.10	0.41	3.20	0.31
3. GPA	2.46	0.73	2.50	0.93	3.14	0.61
4. Absenteeism	1.83	1.00	2.85	1.37	1.93	1.03
5. Self-rating	58.46	9.24	54.65	11.68	58.48	8.57
6. Self-deception	3.25	0.31	3.02	0.42	3.08	0.33
7. Impression management	3.19	0.53	3.36	0.57	3.17	0.44

Note. The means for the biodata measures are item means.

Table 5
Effect Sizes (d) Comparing African American, Asian American, and Caucasian Subgroups

Groups compared	Elaborated biodata	Nonelaborated biodata	GPA	Absenteeism	Self-rating	Self-deception	Impression management
African American–Caucasian	−.15	−.08	−1.00	−.10	−.04	.53	.04
Asian American–Caucasian	−.29	−.42	−.91	.84	−.48	−.18	.41
African American–Asian American	.14	.32	−.05	−.89	.38	.65	−.31

Note. A positive *d* value reflects a higher group mean for the first group in the pair listed; a negative *d* value reflects a higher group mean for the second group. *d* values are based on the pooled standard deviation of the groups.

A much different pattern of over- and underprediction would be observed for the absenteeism and self-rating outcomes, however, because the mean subgroup differences on the biodata measures and absenteeism and self-rated performance are more nearly comparable.

Discussion

The results of this article indicate a partial replication of Schmitt and Kunce's (2002) results in that generally, the mean scores on items for which elaboration was required were substantially lower than the scores on items for which elaboration was not required. However, in the biodata form with items for which elaboration was required, there was minimal carryover to the other items for which elaboration was not required. This may be due to several factors. It is possible that the proportion of items for which elaboration was required affected the size of the carryover effect. The proportion of items on which elaboration was required in this study was about one sixth, whereas in the Schmitt and Kunce study, elaboration was required of one fifth to two fifths, depending on the experimental condition.

We also have a different sample of participants, and the motivation to fake in this instance may have been less. Participants were paid to participate, and significant effort was made to motivate them in the written and oral instructions provided to them. In addition, participants whose responses indicated that they might have been careless were removed from the analyses described in this article. However, the participants' scores were obviously not being used to make admissions or selection decisions. The notion that our participants did make efforts to protect their self-images by making socially desirable responses is supported by the fact that their responses to the Paulhus (1991) measures of self-deception and impression management are similar to those in two large sample studies used in the original development of these measures.

One of our reviewers suggested an alternative explanation for our main findings that attributed depressed scores on elaborated items to participants' lack of effortful responding or laziness. If this is true, one would expect to see a large number of participants choosing the multiple-choice option corresponding to events that never occurred or that occurred once so that they would not be required to elaborate on their answers very frequently. Thus they could expend little or no effort in recalling situations and writing about them. We did indeed find a considerably larger percentage of zero answers across the 21 items in the elaborated condition (35%) as opposed to the nonelaborated condition (16%). However, as indicated above, a very large proportion of the participants provided answers that required elaboration, and many provided answers that required the largest number of elaborated incidents. It is also unlikely that respondents were fatigued and therefore elaborated fewer answers. The biodata questions were the first questions in the test battery to which responses were made, and most respondents answered these items easily in less than an hour.

Correlations of the social desirability and impression management scales with participant responses to the elaborated and nonelaborated items produced some unexpected results. First, the correlations of the social desirability measures were in the direction predicted (i.e., higher for items in their nonelaborated than in their elaborated form), but not significantly so. Perhaps more surprising was the fact that the correlations of these two measures with the nonelaborated items were much higher (.40 and .41 for self-deception and impression management, respectively, in the nonelaboration group, and .46 and .39 in the elaboration group). Contrary to the conclusion reached by Lautenschlager (1994), but consistent with the findings of Becker and Colquitt (1992), these results indicate that there may be differences in response bias depending on the verifiability or concreteness of the items. The results, though, do not provide strong support for the notion that social desirability is reduced by the requirement that respondents elaborate on their answers.

These results, though, raise the possibility that the elaboration manipulation was confounded with the subset of items chosen for elaboration. The items that were chosen for elaboration were more objective and verifiable (Mael, 1991), as indicated by our judgments and those of a set of research assistants. Furthermore, greater item objectivity and verifiability were related to lower correlations with self-deception and impression management scores. Together, these findings suggest that the types of biodata items chosen for elaboration did confound our results but not to a degree that would explain mean differences on elaborated and nonelaborated versions of the same biodata items, which were approximately .80 standard deviations in magnitude.

Criterion-related validity findings were consistent with our hypotheses, in that validity appears to be unaffected by the fact that elaboration is required. Correlations of the elaborated items in both the elaboration and nonelaboration conditions were virtually identical across GPA, class attendance, and self-ratings on the performance composite. The validity of nonelaborated items was actually a little higher, though this may be partly a function of the higher reliability of the nonelaborated biodata items. Similarly, there appears to have been minimal impact on the validity of nonelaborated item responses when they are part of a form in which some items require elaboration. The finding that elaboration does not affect validity is consistent with the conclusion of Ones et al.'s (1996) meta-analysis that the impact of social desirability corrections is minimal. The corrections in this case, though, come in the form of a manipulation of the response required rather than statistical control using a social desirability measure, as was true in the studies examined by Ones et al.

The fact that elaboration reduces mean responses to biodata but seems to have no impact on the validity of biodata is similar to the findings of a meta-analysis by Ones et al. (1996) comparing the responses of incumbents and applicants to integrity tests. Elaboration also did not affect the correlation of biodata with self-deception or impression management. This provides evidence that elaboration does not affect the relative ordering of scores but only reduces the mean. In contrast, the verifiable or nonverifiable nature of the items appears to substantially impact the correlation between biodata scores and self-deception and impression management. One reviewer and the action editor of our paper suggested that this may be an indication that separate processes are involved in the effects of elaboration and item type. The elaboration effect (like the differences observed between applicants and incumbents) may be motivational in nature (individuals want a job or college admissions), whereas the effect of the verifiability of items may be primarily cognitive (perhaps a function of memory). These hypotheses about the different processes involved may serve as the basis of informative future research.

One reviewer also noted that the relationship among GPA, self-deception, and both elaborated and nonelaborated biodata measures represented a pattern of correlations that suggest that self-deception is acting as a suppressor in the prediction of GPA. The notion is that self-deception is not related to GPA, but that because of its relationship with biodata, it suppresses invalid variance in biodata that enhances the prediction of GPA. Although this was not central to the purpose of this article, we performed regressions for each group (elaborated and nonelaborated) of participants for both the elaborated and nonelaborated items. GPA was regressed on both biodata and self-deception. In all four regressions, the regression weight for self-deception was negative; this supports the suppression hypothesis. In three of these regressions, the suppressor effect was statistically significant ($p < .10$); the suppressor effect was not significant for the group that responded to the nonelaborated version of the 21 elaborated items. This pattern of relationships is consistent with some previous researchers' hypotheses (see Ones et al., 1996, p. 662), although Ones et al. (1996) did not find support for this hypothesis in their meta-analyses of integrity tests.

Our final set of analyses addressed the concern that there would be racial/ethnic mean differences in the impact of elaboration on scores on the biodata instrument. None of the mean differences for the three relatively large subgroups were significantly different. Asian Americans were the lowest scoring subgroup, and Caucasians were the highest. Furthermore, mean differences between Asian Americans and Caucasians were about one third of a standard deviation different on both the elaborated and nonelaborated items. This finding, along with an absence of a significant interaction between ethnic group status and test form, indicates that any observed differences between subgroups were not produced by elaboration. These results are not consistent with the findings of Ryan et al. (1998, 2000), but the extra work required of respondents to elaborate answers in our situation did not require as much effort or commitment, as was true in these two earlier studies in which greater effort required of participants resulted in greater ethnic group differences.

Conclusions and Implications

The results of this study indicate that elaborated responses do produce lower scores than do similar nonelaborated versions of the same items and that these differences are quite large. The results do not confirm the existence of a sizable carryover effect on nonelaborated items. Furthermore, it does not appear that elaboration is related to the size of the correlation between social desirability and responses to biodata items, though the transparency or verifiability of the items may be. Finally, validity and subgroup mean differences appear to be unaffected by the requirement that items be elaborated.

Future research should be directed to a determination of the reason and implications for the relatively lower scores on elaborated biodata items. The confounding of the requirement that items be elaborated with their objectivity or verifiability must be further examined, as should the willingness of respondents to engage in the extra effort required to recall and elaborate their answers. Practically, the most important question that must be addressed is the impact of elaboration when respondents are applicants in an actual selection situation in which the stakes are higher than they were in the present study and in which selection test performance would tend to be more maximal than typical in nature. Another related question is whether and when any true carryover effects from elaboration to nonelaboration items exist. For both issues, it is conceivable that motivational differences would be present in an actual selection situation and may produce a very different pattern of results. If so, it would be appropriate to measure and build motivational and cognitive indices, along with elaboration effects, into a theoretical model of influences on responses to biodata measures and, perhaps, to other noncognitive measures.

Finally, it should be noted that the elevation of mean scores on biodata or other noncognitive measures can have serious consequences for their use in selection. Passing scores on a biodata test developed under a concurrent validation model with job incumbents or current college students, as in the study described in this article, may lead to serious underestimation of the scores of applicants. In these cases, organizations will have to renormalize the biodata measures or abandon their use. An alternative might be to use the elaboration approach described in this article.

References

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions in job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Becker, T. E., & Colquitt, A. L. (1992). Potential versus actual faking of a biodata form: An analysis along several dimensions of item type. *Personnel Psychology, 45,* 389–406.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561–590.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10,* 99–109.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual* *differences and behavior in organizations* (pp. 258–299). San Francisco: Jossey-Bass.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey-Bass.

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

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (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, S. A., & Donovan, J. J. (1998, April). *Warning: Proceed with caution when warning applicants not to dissimulate.* Paper presented at the 13th Annual Meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84,* 155–166.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition.* New York: McGraw-Hill.

Gillespie, M. A., Kim, B. H., Manheim, L. J., Yoo, T. Y., Oswald, F. L., & Schmitt, N. (2002, June). *The development and validation of biographical data and situational inventories in the prediction of college student success.* Paper presented at the 14th Annual Meeting of the American Psychological Society, New Orleans, LA.

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82,* 656–664.

Hough, L. M. (1984). Development and validation of the “accomplishment record” method of selecting and promoting professionals. *Journal of Applied Psychology, 69,* 135–146.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13,* 371–388.

Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76,* 889–896.

Lautenschlager, G. J. (1994). Accuracy and faking of background data. In G. A. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook* (pp. 391–419). Palo Alto, CA: Consulting Psychologists Press.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44,* 763–792.

Manheim, L. J., Oswald, F. J., Kim, B. H., Gillespie, M. A., Yoo, T. Y., & Schmitt, N. (2002, June). *Expanding the criterion space of college student success: Beyond GPA.* Paper presented at the 14th Annual Meeting of the American Psychological Society, New Orleans, LA.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual actions: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 61–138). Palo Alto, CA: Consulting Psychologists Press.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50,* 823–854.

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

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

Paulhus, D. L. (1988). *Assessing self deception and impression management in self reports: The balanced inventory of desirable responding.* Unpublished manual. University of British Columbia, Vancouver, Canada.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. L. Shaver, L. Wrightsman, & F. M. Andrews (Eds.), *Measures of personality and social-psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Quinn, B. A. (1989). *Religiousness and psychological well-being: An empirical investigation.* Unpublished doctoral dissertation. Wayne State University, Detroit, MI.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634–644.

Ryan, A. M., Ployhart, R. E., Greguras, G. J., & Schmit, M. J. (1998). Test preparation programs in selection contexts: Self-selection and program effectiveness. *Personnel Psychology, 51,* 599–622.

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal form a multiple hurdle process. *Journal of Applied Psychology, 85,* 163–179.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-related constructs. In C. R. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 11, pp. 115–140). New York: Wiley.

Schmitt, N., & Kunce, C. (2002). The effects of required elaboration of answers to biodata questions. *Personnel Psychology, 55,* 569–588.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency using various predictor combinations. *Journal of Applied Psychology, 82,* 719–730.

Taber, T. D., & Hackman, J. R. (1976). Dimensions of undergraduate college performance. *Journal of Applied Psychology, 61,* 546–558.

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

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability.* New York: College Entrance Examination Board.

(*Appendix follows*)

Appendix

Examples of Two Elaborated Biodata Questions

How many times did you lead class discussions during your senior year in high school?

a. Never
b. Once
c. Twice
d. Three or four times
e. Five or more times

If you answered b, c, d, or e, please list the classes and discussion topics you led. Do not list more than five.

In how many different languages besides English can you converse well enough to order a meal?

a. None
b. One
c. Two
d. Three
e. Four or more

If you answered b, c, d, or e, please list the languages. Do not list more than four.

Received July 25, 2002
Revision received February 27, 2003
Accepted March 6, 2003 ■