Journal of Business and Psychology, Vol. 16, No. 4, Summer 2002 (©2002)

BIODATA VALIDITY DECAY AND SCORE INFLATION WITH FAKING: DO ITEM ATTRIBUTES EXPLAIN VARIANCE ACROSS ITEMS?

Kenneth E. Graham
Caliber Associates

Michael A. McDaniel
Virginia Commonwealth University

Elizabeth F. Douglas
Doylestown, OH

Andrea F. Snell
The University of Akron

ABSTRACT: Rating scales were developed to assess the biodata dimensions offered by Mael (1991). Biodata items assessing conscientiousness were administered under honest-responding and faking-good conditions. Item attributes were examined to determine their value in predicting item validity for honest respondents and item validity for faking respondents. Analyses were also conducted to determine whether the degree of item faking was related to item attributes. Item attributes associated with item validity for honest respondents are not the same as the item attributes indicative of item validity for the faking respondents. We suggest that this makes it very difficult to develop a biodata questionnaire which will be equally valid for both honest and faking respondents.

KEY WORDS: biodata; item attributes; faking; validity.

The topic of accuracy in self-report information has been an issue of substantial debate (Mabe & West, 1982). For biodata, the degree of

Address correspondence to Kenneth E. Graham, Caliber Associates, 231 Cherokee Street, PMB 165, Leavenworth, KS 66048-2818. e-mail: grahamk@calib.com.

0889-3268/02/0600-0573/0

prediction is likely enhanced by the accuracy of the self-report information. Although some authors (Hogan, 1991; Stokes, Hogan & Snell, 1994) have argued that some inaccuracy in responding, such as socially desirable responses, may be related to positive levels of performance criteria, it is reasonable to suggest that accurate presentations of past behavior will be more highly related to future behavior than inaccurate descriptions of past behavior.

There are several reasons why individuals may be motivated to provide inaccurate responses (Lautenschlager, 1994). An inaccurate response could be the result of conscious and intentional false reports. This false reporting may arise out of the respondents desire to gain social approval or to hide specific instances of undesirable behavior (e.g., fired from job). Another possible reason for response inaccuracy may be a problem with memory or lack of self knowledge on the part of the respondent. Inaccuracy may also occur from problems with the actual items (e.g., confusion concerning the meaning of the question). In addition, an accurate response may be judged inaccurate due to errors in verifying information using external sources (e.g., the response is accurate, but the verification criteria is in error).

Although the findings regarding the validity and generalizability of biodata are encouraging, relatively few studies have been directed at examining the potential problems of response inaccuracy in answers to biodata items. This literature we summarize below. Interested readers should also consult Lautenschlager (1994).

LITERATURE REVIEW OF RESPONSE INACCURACY

A study by Keating, Peterson and Stone (1950) was one of the first works regarding response accuracy on self-report work history information. These researchers reported correlations ranging from .90 to .98 between self reported values and information acquired from previous employers regarding wages and employment. A similar study by Mosel and Cozan (1952) reported results that were consistent with the Keating et al. (1950) study. Goldstein (1971) also examined the agreement between application blank information and information from previous employers. The results were in contrast to the Keating et al. (1950) study and the work of Mosel and Cozan (1952). A substantial number of discrepancies were found between the application blank information and information provided by previous employers.

A study by Doll (1971) also examined faking on biodata items. The items were categorized as continuous vs. noncontinuous, and as objective vs. subjective. A continuous categorization was based on whether the alternatives ranged across an obvious continuum. If an item was inde-

pendently verifiable through an objective means, it was classified as being objective. An item considered unverifiable was considered subjective. Participants were divided into one of three conditions: 1) fake good, but be prepared to defend any answers in an interview; 2) fake good, but be aware there is a lie scale to identify people who exaggerate or 3) fake to look as good as possible. The results showed a greater likelihood for subjective and continuous items to be faked. Overall, the most faking occurred in the fake to look as good as possible condition (condition three). Condition two (warning of a lie scale) showed the least amount of faking. Cascio (1975) also supplied evidence that verifiable items are less likely to be distorted than nonverifiable items.

Schrader and Osburn (1977) conducted a study in which some participants were warned of a lie scale and others were not. Results offered further support that the warning of a lie scale reduces response distortion. However, Dwight and Donovan (2001) note that warning does not directly tell us whether respondents respond more honestly when warned because warning may introduce its own systematic bias or even suggest faking as an option to test takers who might not otherwise have thought of it.

McManus and Masztal (1993) used Mael's (1991) taxonomy of biodata item attributes to examine the relationship between item validity and item attributes. The authors presented results suggesting that the historical, external, objective, and verifiable attributes are all strongly related to validity. They found that 45% of the variance in item-level validity is explained by these item attributes.

Terms such as social desirability, impression management, faking, intentional distortion, and self enhancement have been used throughout the non-cognitive literature to describe response distortion (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Lautenschlager, 1994; Viswesvaran & Ones, 1999; Ones, Viswesvaran, & Reiss, 1996). Non-cognitive measures such as personality measures and biodata inventories have been found to be susceptible to response bias due to faking (Hough et al., 1990; Kluger, Reilly, & Russell, 1991; Ones, Viswesvaran, & Korbin, 1995). For example, Kluger et al. (1991) found that when instructed to fake, subjects were able to improve their scores on a biodata measure by one standard deviation. Hough et al. (1990) found that respondents were able to distort successfully their self-descriptions when instructed to do so. Ones, Viswesvaran, and Korbin (1995) used meta-analysis to determine whether or not respondents are able to fake responses intentionally to personality inventories. Their results showed that, on average, respondents are able to improve their scores by nearly half a standard deviation through faking.

Much of the faking research has used scores on social-desirability scales as an operational definition of faking. Christiansen, Goffin, John-

ston, and Rothstein (1994) examined the effects of correcting for social desirability in the Sixteen Personality Factor Questionnaire. The results showed that correcting for social desirability is unlikely to affect the criterion-related validity of the measure. These results were supported by Ones, Viswesvaran, and Reiss (1996) who analyzed the cumulative literature on social desirability scales to assess the impact of such scale scores on the criterion-related validity of the Big Five personality factors. The results of their analyses indicated that correcting for social desirability scales does not improve the criterion-related validity of personality scales. In addition, they found that social desirability is related to stable personality constructs, specifically emotional stability and conscientiousness. They demonstrated that removing the effects of social desirability from personality scale scores would result in the loss of relevant variance from the measures. Based on this evidence, Ones, Viswesvaran, and Reiss (1996) concluded that socially-desirable responding is not a problem regarding the use of personality measures in personnel screening. However, the extent to which the social desirability studies have relevance for an understanding of the effects of faking on test validity is unclear. It is clear that social desirability scales do not capture all faking variance (Ellingson, Sackett, & Hough, 1999).

Faking can be situationally induced (Douglas, McDaniel, & Snell, 1996). For example, it is well documented that people can fake when instructed to do so by an experimenter (Hough et al., 1990; Kluger et al., 1991; Ones et al., 1995). Based on this evidence, it is reasonable to expect that people may also fake effectively when they are motivated to fake through self interest. For example, individuals may demonstrate faking behaviors, such as distorting resumes and answers in interviews and employment tests, when it is seen as necessary to get a valued outcome, such as employment. Stokes, Hogan, and Snell (1993) demonstrated that applicants were more likely to engage in impression management than incumbents due to motivation to increase their probabilities of obtaining employment. Staffing professionals who review applications and conduct interviews have provided anecdotal evidence supporting these results.

Effects of Response Inaccuracy

Douglas, McDaniel and Snell (1996) examined the effects of situationally-induced faking on the validity of non-cognitive measures. A personality instrument and a biodata instrument were used. Each instrument assessed conscientiousness and agreeableness. Participants completing the questionnaires were told to either fake good or to respond honestly. Douglas et al. found that under conditions of faking, construct and criterion-related validity decreased. They concluded that the con-

struct and criterion-related validity of non-cognitive tests is substantially harmed when respondents fake to look good.

The present study was designed to extend the findings of Douglas et al. (1996). Using biodata items designed to assess conscientiousness and agreeableness, we sought to explore the extent to which item attributes covary with the extent of faking on items and with their item validities in both honest and faking conditions. One applied goal of the research was to explore the possibility of identifying attributes of items that are resistant to faking and which have non-zero validities under conditions of faking.

METHOD

Measures

The focus of the present paper is on item-level analyses and results, however person-level measures were used to provide data at the item level. Person-level measures are those which assess characteristics of the respondent. Person-level measures included two biodata questionnaires, and job performance ratings.

Item-level measures are those that assess item attributes. These attributes are described in Table 1. These item attributes can be called construct-irrelevant item attributes because the item attributes are substantially independent of the constructs assessed by the items. Analyses associated with item-level variables concern the prediction of item validi-

Table 1
Item Attributes and Example Items (adapted from Mael, 1991)

Verifiable What was your High School grade point average?	*Nonverifiable* How many hours a day do you spend in physical activity?
Continuous Compared to others how organized are you? (Responses range from: a) not organized at all–b) extremely organized.)	*Noncontinuous* How would you describe yourself? (Responses may include: a) friendly; b) unorganized; c) intelligent; d) shy)
External Have you ever skipped a day of school?	*Internal* What is your attitude toward people who skip school?
Controllable How many attempts did it take you to pass your driver's test?	*Noncontrollable* How old are your parents?
Historical How many cars did you sell in the past month?	*Future or Hypothetical* For what reason would you leave your current job?
First Hand At what age did you receive your driver's license?	*Second Hand* My friends would most likely describe me as:
Objective How many times were you late for work last month?	*Subjective* Would you describe yourself as a hard worker?

ties by condition and the prediction of the extent to which item scores can be improved by faking. The person-level measures will be described first followed by item-level measures.

Biodata Questionnaires. Two biodata inventories contributed data to this study. The one described by Graham (1996) was a 106-item measure. The other, drawn from Douglas et al. (1996), contained 56 items.

The items from Graham (1996) were designed to assess the respondents' level of conscientiousness. Conscientiousness has been found to be a consistent predictor of job performance across a wide range of occupations (Barrick & Mount, 1991; Hurtz & Donovan, 2000). A taxonomy of conscientiousness content areas based on Hogan and Hogan (1992), is shown in Table 2. Item development was guided by this taxonomy of conscientiousness. Typically, biodata items are construct-heterogeneous. Graham did not claim that any one of the items was tapping a specific conscientiousness taxonomy category and was irrelevant to all others. Rather, Graham used the taxonomy as a general guide in developing items to permit the pool of items to cover a broad range of conscientiousness.

The second biodata instrument was drawn from Douglas et al. (1996) and included 56 items rationally keyed with the goal of measuring conscientiousness and agreeableness. When combining the Douglas et al. measure with the 106-item instrument from Graham (1996), there was a total of 162 items. The analyses are based on these 162 items.

Performance Ratings. With permission from the participants, supervisors were sent a three-page performance appraisal form which contained nine items representing quantity, quality, accuracy, job knowledge, efficiency, dependability, loyalty, motivation, and overall performance. Responses were provided using a five-point scale. For example, on the quality dimension, supervisors were asked, "How good is the quality of the work?" and responses ranged from "Performance is inferior and almost never meets minimum quality standards" to "Performance is almost al-

Table 2
Hogan and Hogan (1992) Taxonomy of Conscientiousness

Conscientiousness Subscales	
Moralistic	Adhering strictly to conventional values
Mastery	Being hard working
Virtuous	Being perfectionistic
Not Autonomous	Concern about others' opinions of oneself
Not Spontaneous	Preference for predictability
Impulse Control	Lack of impulsivity
Avoids Trouble	Professed probity

ways of the highest quality." In addition, supervisors were asked to indicate how often they observed the employee working and how long they had worked with the employee. This performance appraisal form was an adaptation of the performance appraisal used successfully in the U.S. Department of Labor's validation effort for the General Aptitude Test Battery. An overall performance scale was calculated by summing all nine items.

Supervisors were sent a letter explaining that the ratings were being collected for research on an experimental personnel screening measure and that their responses were to be kept confidential. A job performance appraisal form and a copy of the participants signed inform consent sheet were also included. The item-level variables are described below.

Construct-Irrelevant Item Attributes. Each of the 162 items were rated by four independent raters using the dimensions outlined in Table 1. A review of the rating scales (see Appendix) indicates that we needed to use nominal scales to assess adequately the rated attributes. This was due to some dimensions being more relevant to some biodata items than others. Consider the external/internal dimension. External items concern expressed actions (e.g., "Have you ever skipped a day of school?") whereas internal items refer to attitudes, opinions, or emotional reactions (e.g., "What is your attitude toward people who skip school?). Where in this dimension does one place an item concerning an individual's behavior in a future situation? For example, consider the item: "If you had the opportunity to attend school next year, what would you do?" Such items reduced the reliability of our initial rating scales. We resolved the problem by creating a nominal-level rating category for the external/internal dimension which covered purported behavior in a future hypothetical situation. By creating nominal rating categories for the rating dimensions defined by Mael (1991), we obtained satisfactory reliability. However, the rating categories within a rating dimension seldom can be ordered on a continuum. Thus, using external/internal as an example, one cannot speak of an external/internal biodata dimension, but rather one can speak of a set of nominal categories organized around the concept of external/internal.

The raters consisted of three graduate students and one undergraduate student. All were familiar with biodata research. The four raters were given a rating scale for each of the seven attributes. Each rater was asked to classify each item into the most appropriate category for each dimension. The scales consisted of three example items for each rating category and a written description of the category.

These attributes are based on Mael's taxonomy of biodata items (Mael, 1991). Mael's taxonomy was not developed based on the items' ability to predict validity. Rather, it is a comprehensive taxonomy that provides a method for examining item attribute correlates of item valid-

ity and the degree to which item scores can be improved through faking. This classification scheme was used as a means to classify our 162 items according to multiple item attributes.

Faking Effect Size. The faking effect size was defined as the standardized mean difference *(d)* between the faking respondents and the honest respondents. A positive *d* indicates that the faking respondents reported more of the construct assessed than the honest respondents. The *d* statistic has a mean of zero and a standard deviation of one. Thus, if an item has a *d* score of one, the mean of the faking respondents is one standard deviation above the mean of the honest respondents.

Validity Coefficients. Two additional item-level variables concerned the criterion-related validity of the items. One variable consisted of the item validity calculated from those individuals assigned to the honest responding condition. The second variable consisted of the item validity calculated from the data provided by the faking respondents.

Summary of Item-Level Information. To summarize, the item-level variables consisted of ratings on each of the nominal categories for the seven item attributes, a faking effect size, and two validity coefficients. The seven item attributes, described in Table 1, assess attributes which are conceptually independent of the constructs measured by the item. The faking effect size is a standardized mean difference between the honest and faking groups. An item with a large positive faking effect size would be an item in which the faking group scored substantially higher on the item than the honest group. One validity coefficient was the criterion-related validity coefficient for the item for the honest respondents. The second validity coefficient was the criterion-related validity coefficient for the item for the faking respondents. Once again, the unit of analysis was the item; the variables analyzed were item-level attributes.

Participants

The unit of analysis for this study is the item. There were 162 items. However, item-level data originates within person-level data. For example, one item-level variable is the validity of the item for individuals instructed to respond honestly. To obtain this item-level validity, multiple individuals must respond to the item and have job performance criterion data available. To minimize the sampling error variance in the item-level variables, we used data from all available individuals. This resulted in some item-level variables being based on more person-level data than other item-level variables. To further complicate matters, the data draws on items from two separate biodata inventories (Douglas et al., 1996; Graham, 1996) which had different data collection histories. Table 3

Table 3
Number of People (sample size) Contributing Data to Various Item-Level Variables

Item Attribute	Item Source	Sample Size
Validity for honest respondents	Douglas et al. (1996)	97
Validity for honest respondents	Graham (1996)	55
Validity for faking respondents	Douglas et al. (1996)	111
Validity for faking respondents	Graham (1996)	47
Faking effect size	Douglas et al. (1996)	600
Faking effect size	Graham (1996)	273

Note: Not all sample members had criterion data. Thus the sample size for the validity statistics is smaller than the sample size for the faking effect sizes.

presents the sample sizes which contributed to each of the item-level variables except the construct-irrelevant item attributes. The construct-irrelevant item attributes were based on ratings provided by the four raters.

Procedure. Participants signed an informed consent form and were read a description of the study. Participants were administered the biodata inventory and were instructed to fake good or to answer honestly when answering the biodata questionnaire. If employed, participants were given the opportunity to authorize permission for the authors to send the participants' supervisors the performance appraisal form. Participants were assigned extra credit at the end of the session.

RESULTS

Reliability of Attribute Ratings

Table 4 presents the first set of results for the item-level analyses. The table shows the reliability of the attribute ratings. The reliability was calculated using a Kappa statistic (Fliess, 1971). Reliability for the verifiable, continuous, controllable, first-hand and objective scales were calculated using all four of the raters. The reliability of the external and historical scales were calculated after dropping two of the four raters. This adjustment was performed to increase reliability of the ratings to an acceptable level.

Predicting Item Validity and Item Faking

Table 5 shows for each item attribute category, the number of items in each category, the mean validity for the honest and faking respon-

Table 4
Reliability of the Construct-Irrelevant Attribute Ratings

Scale	Reliability
1. Continuous/Noncontinuous	.62
2. Verifiable/Nonverifiable	.61
3. External/Internal	.61
4. Controllable/Noncontrollable	.68
5. Historical/Future-Hypothetical	.71
6. First-hand/Second-hand	.87
7. Objective/Subjective	.68

dents and the mean faking effect size. The results in Table 5 address three questions:

(1) When respondents answer honestly, what is the mean item validity by item attribute categories?
(2) When respondents fake to look good, what is the mean item validity by item attribute categories?
(3) How does the extent to which the item scores can be improved through faking (i.e., the faking effect size) vary by item attribute categories?

We defined the faking effect size as the standardized mean difference between the honest and faking groups. What stands out in Table 5 is that the item characteristics associated with item validity for honest respondents are almost always different from the item characteristics that are associated with item validity for the faking respondents. The last column of Table 5 addresses what types of items are most likely to be faked by showing the mean faking effect sizes for each item attribute.

DISCUSSION

Implications for Assigning Items to the Mael (1991) Taxonomy

Anecdotal evidence from other biodata researchers suggests that developing reliable attribute ratings of biodata items is typically quite difficult. The biodata item pool used in this study was fairly diverse with respect to item format as we pursued an ambitious goal of evaluating the most critical attributes of Mael's (1991) taxonomy. A set of rating scales were developed which can be used to categorize biodata items in a reliable manner. Thus, these rating scales will be useful to future researchers. However, in the course of developing these scales, it was found that biodata items can be written in an amazing variety of ways.

Table 5
The Number of Items Represented in Each Item Attribute Category, Mean Validity for the Honest and Faking Respondents, and the Mean Faking Effect Size

Construct-Irrelevant Variables	# of Items in Category	*Mean Item Validities*		Mean Faking Effect Size
		Honest	Faking	
Verifiable-A (verifiable through hard records)	.19	.16	.02	.37
Verifiable-B (non-verifiable items)	105	.08	.04	.36
Verifiable-C (verifiable through supervisors or coworkers)	30	.12	.03	.28
Verifiable-D (verifiable through friends)	8	.08	.11	.52
Continuous-A (alternatives represent a clear continuum of some construct; considered continuous)	67	.11	.05	.42
Continuous-B (alternatives do not represent a continuum of some construct; considered noncontinuous)	37	.11	.01	.21
Continuous-C (it is possible to detect an underlying continuum but the alternatives do not represent a clear continuum; considered grey continuous)	58	.07	.05	.38
External-A (items asking for expressed actions to events; considered external)	41	.06	.06	.42
External-B (items asking for attitudes and opinions; considered internal)	115	.11	.04	.33
External-C (hypothetical assessment of situations; considered internal)	6	.03	–.06	.41
Controllable-A (items concerned with actions that the individual chooses to perform; considered controllable)	43	.06	.07	.45
Controllable-B (items concerned with actions over which the individual has no control; considered noncontrollable)	2	.22	–.02	.13
Controllable-C (items concerned with the individual's feelings or attitudes)	64	.09	.03	.30
Controllable-D (items concerned with situations where control is shared)	53	.13	.04	.35
Historical-A (items asking about past behavior; considered historical)	76	.09	.06	.38
Historical-B (items asking about situations that may occur in the future; considered future or hypothetical)	12	.06	–.06	.26
Historical-C (items asking for present feelings and personal knowledge; considered present)	74	.10	.04	.35
First-Hand (items which require personal knowledge of the individual)	114	.08	.05	.36
Second-Hand (items which require the respondent to consider other's evaluations or opinions of the respondent's performance or attributes)	48	.13	.03	.34

Table 5 (*Continued*)

Construct-Irrelevant Variables	# of Items in Category	*Mean Item Validities*		Mean Faking Effect Size
		Honest	Faking	
Objective-A (items asking for the assessment of actual behavior; considered objective)	26	.08	.07	.44
Objective-B (asking for judgments that may be biased by personal feelings or interpretations; considered subjective)	86	.08	.04	.34
Objective-C (asking the respondent to assess other people's feelings and opinions; considered subjective)	50	.13	.03	.35
All 162 items	162	.09	.04	.35

Thus, when these rating scales are applied to new item sets, further refining of the scales may be required.

What Item Attributes Are Associated with High Validity for Honest Respondents?

Table 5 shows the mean item validity by item attribute category when individuals respond honestly. The mean validity for all 162 items for honest respondents was .09. The pattern of mean item validities by attribute category form a discernable pattern for the honest respondents. Items asking for information derived from sources outside oneself tend to be more valid than items focusing on the respondents own perceptions of themselves. Consider the verifiable categories. Both verifiable-A (verifiable through hard records) and verifiable-C (verifiable through supervisors and coworkers) contain items with above average validities (.16 and .12), whereas verifiable-B (nonverifiable items concerning the respondent's thoughts) have below average validity (.08). Next consider external category-B, which includes items concerning information in personnel records and the assessment of the applicant by his current supervisor. These items have above average item validity (.11). In contrast, external categories A and C do not primarily assess information from sources external to the applicant. Items in these categories have below average item validity (.06 and .03). A similar pattern is found for the controllable item attribute. Items in category controllable-A are primarily based on the individual's own perceptions of themselves and past events. These items have lower than average item validity (.06). In contrast, items categorized as controllable-D, which includes items concerning others actions with respect to the respondent as well as information in personnel records, have above average item validity (.13). When one

compares first-hand items, which require information based on the respondent as the source, with second-hand items which ask the respondent to provide information from other sources, such as a supervisor or coworker's opinion, the first-hand items have below average validity (.08) whereas the second-hand items have above average validity (.13). Finally, the same pattern is seen with the objective items. Objective-B items ask for the respondents opinions about themselves. These items are associated with low validity (.08). In contrast, items in category objective-C require the respondent to report others opinions of the respondent and these items have above average validity (.13). Collectively, for honest respondents, this pattern of mean validities provides evidence that self-reports of the respondents own feelings, opinions, and behaviors are less valid then the respondents self-reports about how others view the respondent.

What Item Attributes Are Associated with High Validity for Faking Respondents?

Table 5 also shows the mean item validity by item attribute category for the faking respondents. The mean validity for all 162 items for faking respondents was .04. This was less than half the validity of the same items for the honest respondents (.09). For 20 of the 22 item attribute categories, the mean validity was always higher for the honest respondents than the faking respondents. For faking respondents, the validities for all item attribute categories were essentially zero. Therefore, unlike the results for the honest respondents, we could not identify any item attributes that were resistant to faking.

What Item Attributes Are Associated with the Degree to Which Items Can Be Faked?

We have operationalized the extent to which items can be faked as the standardized mean difference between groups instructed to fake and groups instructed to respond honestly. Across all 162 items, faking respondents could improve their item scores over that of the honest respondents by .35 standard deviations. One of the strongest indicators of the extent of faking is the degree to which the item responses are continuous. Continuous items offer response alternatives that clearly fall along a continuum (e.g., I do this behavior never, sometimes, often). This is reasonable because continuous items, more so than less continuous items, offer a clear path for faking. Specifically, one can fake a continuous item by indicating that one does the maximum amount of a positive behavior and the minimum amount of an undesirable behavior. Although there are only eight items in verifiable-D (verifiable through friends) and thus the results may be spurious, these items appear to be

easily faked. Other item categories with above average levels of fakability include external-A (items asking for expressed actions to events), external-C (items concerning hypothetical assessment of situations), controllable-A (items concerned with actions the individual chooses to perform), and objective-A (items asking for the assessment of actual behavior).

Item Attributes Associated with Item Validity for Honest and Faking Respondents

There was little overlap between item attributes associated with item validity for honest respondents and item attributes associated with item validities for faking respondents. We had hoped to identify item attributes that were strongly related to item validity for both honest and faking respondents. One could then build biodata tests that would be valid predictors for both faking and honest respondents. These analyses suggest that this task may be harder than initially believed.

This study has clearly shown that when individuals fake as much as they can, their responses to all types of biodata items (e.g., continuous, verifiable, objective, etc.) are not predictive of their job performance. Although this does not bode well for creating a nonfakable biodata instrument, it does not preclude the discovery of item attributes that result in a less fakable measure when individuals use less extreme faking strategies.

CONCLUSION

This paper has examined the relationships between item attributes and the validity of the items for honest and faking respondents as well as the extent to which the items can be faked. For an honest responding individual, one obtains higher validities when the respondent is asked to report how others view her or how her performance is revealed in records. One obtains lower validities when one asks the respondent to use herself as the source of information about her performance or capabilities. This finding is contrary to the prevalent measurement strategy of asking individuals to respond directly to questions about themselves.

For faking respondents, the average item validity is quite low (.04). No interpretable set of item attributes that resulted in acceptable validities when individuals fake their responses were identified. If the current study had identified a set of item attributes that are valid for both honest and faking respondents, the feasibility of building biodata items which have useful levels of validity for both honest and faking respondents would have been increased. However, given these data, we are less optimistic concerning the feasibility of building biodata items that are valid for both honest and faking respondents.

Appendix A
Rating Scales for Construct-Irrelevant Item Attributes

Rating Code	Description	Sample Item
Scale 1		
Verifiable vs. Nonverifiable		
A	Verifiable items are items that can be corroborated from an independent source. This corroboration can come from archival data, such as work records and supervisory ratings. The item must specify a concrete source and not rely on opinions of any kind.	According to your personnel records, how often are you late for work? Your personnel file would most likely show that you have: Supervisory ratings would most likely reveal that you:
B	Nonverifiable items are items about the individual's thoughts, feelings, and behaviors that are only accessible to the individual. The verification for these items can only come from the respondent. Hypothetical questions are also considered to be Nonverifiable.	According to you, your best quality as an employee is your: What is the one thing you hate most about your job? If you had to go out of town at the last minute and could not make it to work you would:
C	These items ask for a *supervisor's or a coworker's* perception of the individual's behavior. These items are considered verifiable but the verification is not as easy to obtain as hard records such as that described in category A.	My coworkers would say that I am: Coworkers would describe the quality of your work as: My supervisor would describe my work ethic as:
D	These items ask for a "friend's" or "other's" perception of the individual's behavior. These items are also considered to be verifiable but the verification is not as easy to obtain as hard records or coworker's and supervisor's assessments.	Others would describe me as: Your friends would describe you as a: Friends would most likely say that I:
Scale 2		
Continuous vs. Noncontinuous		
A	Continuous items are characterized by the items alternatives representing a clear continuum of some construct. The alternatives will represent the construct in varying degrees. It is common for the alternatives to range from "all the time" to "never."	How often do you volunteer to do community work? a. Always b. Often c. Sometimes d. Rarely e. Never When I start a project I finish it: a. all the time. b. often. c. sometimes. d. rarely. e. never.

Appendix A (*Continued*)

Rating Code	Description	Sample Item
		Being on time for work is: a. very important to you. b. somewhat important. c. not very important. d. not at all important to you.
B	Noncontinuous items contain alternatives which do not represent a continuum of some construct. Here, the alternatives cannot be arranged along a continuum of varying degrees. Each alternative represents an entire construct in and of itself. *The respondent is not capable of detecting which alternative is best* because often times every alternative represents a possible correct answer. Many of the alternatives will seem equal in their desirability.	The best quality an employee can possess is: a. dependability. b. being hard working. c. responsibility. d. maturity. I am most often absent from work because of: a. illness. b. car trouble. c. vacation. d. family obligations. What do you enjoy doing most in your spare time? a. Exercising b. Reading c. Traveling d. Shopping
C	Grey continuous items are items whose alternatives vary in their desirability but do not represent a clear continuum. *It is possible to rank the alternatives from best to worst* even though they are not representative of the same construct. With these items it is possible to detect an underlying continuum in the alternatives but they do not all vary along a clear continuum. You must be able to choose the best alternative, the worst alternative, and rank the ones in between.	If you are late for an appointment you would: a. call and let the person know you are late. b. drive faster to get there. c. not worry about it. d. just decide not to go at all. If you witness a coworker drinking on the job you would: a. ignore the situation. b. confront the coworker. c. tell your buddies about it. d. notify your supervisor. When I make an error I: a. fix it immediately. b. leave it for someone else to worry about. c. blame it on someone else. d. inform my supervisor and ask what I should do.
Scale 3 External vs. Internal		
A	External items concern expressed actions. Some items concern actions involved in prior occurrences of real life events.	Of your past jobs, what is the most common way you ended your employment? When your employer tells you to do something you: In the past six months, how many times have you been absent from work?

Appendix A (*Continued*)

Rating Code	Description	Sample Item
B	Internal items refer to attitudes, opinions, and emotional reactions to events in general. Some items will require the individual's subjective assessment of other's attitudes or opinions. Other items will require self-judgments of skills, knowledge, or attitudes, possibly in comparison with others. This category also asks the respondent's opinion regarding information contained in personnel or supervisory records.	My coworkers would most likely describe me as: My best attribute as an employee is: My supervisor would say that my best asset is:
C	Other internal items are concerned with the hypothetical assessment of situations. The stem must be asking a question about the respondent's *future* action.	If faced with a dilemma, I would: If I could change one thing about myself it would be: If given the opportunity, I would:
Scale 4 Controllable vs. Noncontrollable		
A	Controllable items are concerned with *actions* that the individual chooses to perform or not to perform. These questions may be future oriented, historical, or hypothetical.	When you are reprimanded at work you most often: How many tries did it take you to pass your driver's test? If you witness a coworker stealing, you would:
B	Noncontrollable items are concerned with actions over which the individual has no control. These items can be either historical or future oriented. Hypothetical items *would not* be included in this category.	During your school years, how much emphasis did your parents place on academics? How many cousins do you have? When my parents are senior citizens they expect me to:
C	Some items are not concerned with the controllability of actions. These items will question the individual's *feelings or attitudes* about situations, articles, or persons other than themselves. They may also require the individual to self-report skills or personal knowledge.	What is the most important characteristic of a good employee? A sure sign of a job well-done is:
D	Some items will ask the individual to describe interactions with others over which control is shared. Often, the situation is one where the individual is approached by another and the only control they have is to ignore the person or refuse to interact with them. These items ask about another's actions where you may have had some influence on them, such as a supervisor's opinion.	How often do people tell you their problems? Personnel records would describe you as:

Appendix A (*Continued*)

Rating Code	Description	Sample Item
Scale 5		
Historical vs. Present vs. Future/Hypothetical		
A	Historical items require that the individual consider their actions, reactions, and events that have occurred in their past. These items also include information that may be contained in personnel records regarding past performance or behaviors. To answer the item, the respondent *must think about their past behavior*.	The longest I have held a job is: How many dates did you go on last month? In high school, I was generally thought of as: *****This item is asking for a present opinion about past performance or behaviors. Items such as this need to rated as historical because the respondent must consider their behavior in the past.*
B	Future or hypothetical items are about *situations* that may occur in the future. They can cover reactions, attitudes, actions, and opinions. These items are often in the format of "If-Then" statements.	If a coworker is struggling I would: In order to advance in one's career, one should: *****This item should be rated as hypothetical because it is asking the respondent to consider the behavior of a hypothetical person (i.e. one).* If I witnessed a coworker drinking on the job I would:
C	Present items require the individual to consider their present feelings and personal knowledge as well as others feelings and opinions about the individual.	The one thing that irritates me most is: Coworkers would describe me as:
Scale 6		
First-Hand vs. Second-Hand		
A	First-hand items are items that require personal knowledge or judgment of the individual.	At what age did you receive your driver's license? How many days were you on vacation last year? Which of the following best describes you?
B	Second hand items require the individual to assess other's evaluations or opinions of the individual's performance or attributes.	Other people find me irritating: My supervisor is most likely to say that I am: My friends would most likely describe me as:
Scale 7		
Objective vs. Subjective		
A	Objective items are concerned with the individual's first hand knowledge of events. The items require only recall of facts without the influence of feelings, opinions or judgments. *These items ask for the actual behavior or incident that occurred*.	In the past if I have become angry at a friend, I have: How many times did you work on the weekend last month? How many times were you late for work last month?

Appendix A (*Continued*)

Rating Code	Description	Sample Item
B	Subjective items are concerned with *judgments that may be biased by personal feelings or interpretations*. Some items question the individual's feelings or judgements about situations, persons, or articles. These include *hypothetical* questions and questions which require self judgements.	The most important characteristic of a good worker is: Your best asset as an individual is: Would you describe yourself as a hard worker?
C	Other subjective items will ask the individual to assess other people's feelings and opinions. These may include the feelings or opinions of a supervisor or coworkers. These also may include opinions as to the information a supervisor recorded personnel files or work records.	Most people think that I am: My supervisor would most likely report that I am: My friends would probably describe me as:

REFERENCES

Barrick, M. R., & Mount, M. D. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

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

Christiansen, N. D., Goffin, R. D., Johnston, J. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47,* 847–860.

Doll, R. E. (1971). Item susceptibility to attempted faking as related to item characteristic and adopted fake set. *Journal of Psychology, 77*, 9–16.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). *The validity of non-cognitive measures decays when applicants fake*. Academy of Management Proceedings, Cincinnati, OH.

Dwight, S. A., & Donovan, J. J. (2002). Does warning not to fake actually reduce faking? *Human Performance*, Manuscript under review.

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.

Fliess, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

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

Graham, K. E. (1996). *Biodata validity decay and score inflation with faking: Do item attributes explain variance across items?* Unpublished Master's Thesis. The University of Akron, Akron, Ohio.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory Manual*. Tulsa, OK: Hogan Assessment Systems.

Hogan, R. T. (1991). Personality and personality assessment. In M. D. Dunnette and L. M. Hough (Eds.), *Handbook of industrial and organizational psychology. Second edition. Volume 2* (pp. 873–919). Palo Alto, CA: Consulting Psychologist Press.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D. & McCloy, R. A. (1990). Crite-

rion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581–595.
Hurtz, G. M. & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869–879.
Keating, E., Paterson, D. G. & Stone, C. H. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34*, 6–11.
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 Stokes, G. S., Mumford, M. D. *et al. The Biodata Handbook: Theory, Research and Applications* (pp. 391–419). Palo Alto, CA: CPP Books.
Mabe, P. A., III., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280–296.
Mael, F. A. (1991). A conceptual rational for the domain and attributes of biodata items. *Personnel Psychology, 44,* 763–792.
McManus, M. A. & Masztal, J. J. (1993). *Attributes of biodata: Relationships to validity and socially desirable responding*. Paper presented at the 8th Annual Conference for the Society of Industrial and Organizational Psychology, San Francisco.
Mosel, J. M. & Cozan, L. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 36,* 365–369.
Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The Red Herring. *Journal of Applied Psychology, 81*, 660–679.
Schrader, A. D. & Osburn, H. G. (1977). Biodata faking: Effects of induce subtlety and position specificity. *Personnel Psychology, 30,* 395–404.
Stokes, G. S., Hogan, J. B., & Snell, A. F. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: The influence of social desirability. *Personnel Psychology, 46,* 739–762.
Viswesvaran, C. & Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.