Journal of Applied Psychology
December 1991 Vol. 76, No. 6, 889-896

Faking Biodata Tests
Are Option-Keyed Instruments More Resistant?

Avraham N. Kluger
Department of Industrial Relations and Human Resources
Richard R. Reilly
Department of Management Stevens Institute of Technology
Craig J. Russell
Department of Industrial Relations and Human Resources

ABSTRACT

Response biases in biodata scores derived with option-keying and item-keying procedures were investigated. Results indicated that (a) when subjects simulated responding as job applicants they distorted their responses in a socially desirable direction; (b) item-keyed scores were susceptible to inflation due to socially desirable responding and specific job-title knowledge, but option-keyed scores were not; and (c) response biases were not reflected in response latencies. A supplementary analysis indicated that the two keying procedures may capture different aspects of criterion variance. Implications for reconciling conflicting reports about the susceptibility of biodata scores to response biases are discussed. Issues related to reliability and validity of the two keying procedures, and the generalizability of the results to personality tests, are also discussed.

This research was supported in part by a grant from the Gaudet Fund at Stevens Institute of Technology. Revisions were completed while Craig Russell held a position at Purdue University; he is now at the College of Business, Louisiana State University, Baton-Rouge, Louisiana 70803-63120.

We thank Mike McDaniel for his help in obtaining a biodata instrument, and we thank Philip Bobko, Adrienne Colella, Angleo DeNisi, and three anonymous reviewers for their very helpful comments. A copy of the software used for this study can be obtained from Avraham N. Kluger.

Correspondence may be addressed to Avraham N. Kluger, Institute of Management and Labor Relations, Rutgers University, Rockafeller Road, Kilmer Campus, New Brunswick, New Jersey, 08903-5062.

Received: April 23, 1990
Revised: June 13, 1991
Accepted: June 18, 1991

The use of biographical information, or biodata, has a long and successful history in personnel selection. Reviews by Reilly and Warech (1990) , Hunter and Hunter (1984) , Schmitt, Gooding, Noe, and Kirsch (1984) , and Reilly and Chao (1982) affirm the validity of biodata across a wide variety of jobs and criteria. As biodata instruments become more widely used, however, concerns have arisen regarding the accuracy of job applicants' self-reported data (see Fleishman, 1988).

Research on the susceptibility of both verifiable and nonverifiable biodata instruments to response biases has produced mixed results. Some researchers have found little response bias (Cascio, 1975 ; Colquitt & Becker, 1989 ; Keating, Patterson, & Stone, 1950 ; Mosel & Cozan, 1952), whereas others have reported evidence for substantial response bias (Goldstein, 1971 ; Hogan & Stokes, 1989 ; Weiss & Dawis, 1960).

Mumford and Owens (1987) speculated that differences in item-keying strategies may explain these inconsistent findings. Although it is often not clear which type of keying procedure was used (e.g., Klein & Owens, 1965 ; Schrader & Osburn, 1977), it seems that studies reporting problems with faking (e.g., Hogan & Stokes, 1989) used item-keying strategies (see Lecznar & Dailey, 1950), whereas studies reporting fewer problems with faking (e.g., Trent, Atwater, & Abrahams, 1986) used an option-keying strategy (see England, 1961).

The item-keying (IK) strategy assumes linear, monotonic relationships between item scores and the criterion, ignoring possible nonlinear relationships. For example, if a 5-point Likert-type biodata item is scored from 1 to 5 and that item has a positive correlation with the criterion, then a response of 1 will contribute one point toward the total biodata test score, a response of 2 will contribute two points, and so on (assuming unit item weights).

With an option-keying (OK) strategy, each item response *option* (alternative) is analyzed separately and contributes to the score only if it correlates significantly with the criterion. A common application of the OK strategy, called the contrasting groups method (England, 1961), involves comparing the frequency with which each option was chosen by high and low criterion groups. Options for which there are significant frequency differences are keyed with either 1 or – 1 (depending on whether high or low criterion groups choose it more frequently); other options are keyed with 0. An item with five Likert-scale points, for example, might be keyed so that a 2 contributes one negative point, 4 contributes one positive point, and 1, 3, and 5 are scored zero. Thus, option-keying methods offer the potential advantage of capturing both linear and nonlinear relationships between the item scale and the criterion.

The OK method obviously involves fitting a much larger number of parameters than the IK method. Therefore, when weights obtained from one sample are applied to a cross-validity sample, OK validities ought to shrink more than IK validities. Mumford and Owens (1987) noted that, despite differences in potential shrinkage, the two methods generally yield comparable cross-validities. Thus, on the basis of these psychometric criteria, there seems to be no reason to prefer one keying procedure over the other.

The two keying methods may also differ in susceptibility to inflation from "faking." Hogan and Stokes (1989) defined faking in terms of socially desirable responding. Indeed, Hogan and Stokes found that job applicants were more likely than incumbents to respond to biodata items in a socially desirable way. Alternatively, Trent et al. (1986) defined faking as responses made in a manner most likely to result in a job offer. Socially desirable responding is one strategy an applicant can use to maximize the chances of receiving a job offer. Regardless of bias source, the resultant inflation of the respondent's score may depend on the type of keying procedure used.

The following item, taken from a biodata inventory used to predict store manager job performance (Russell & Domm, 1990), [1] illustrates this point: "How often have you had problems getting a job done because you did not have the right kind of people?" The item's response alternatives are (1) *never,* (2) *seldom,* (3) *sometimes,* (4) *often,* and (5) *very often.* The example item had a negative correlation (– .26) with store manager job performance (in a sample of incumbents). Because it is probably more socially desirable to report a low frequency of this type of experience, respondents who do so receive higher scores on a key designed with the IK (correlational) method. However, the same item when option keyed produced a nonlinear relationship between the 5-point scale and the criterion. In the same sample of incumbents, the OK method produced weights of + 1 for Option 2, – 1 for Option 4, and 0 for Options 1, 3, and 5. For items keyed with the OK method, subjects who deflate their reported frequency toward the more desirable end do not necessarily receive higher scores and may even receive lower scores, depending on the empirical option weights. In fact, any change from an honest response to a more socially desirable response may decrease, increase, or leave unchanged the resultant OK score for a given item (e.g., a change from an honest response of 4 to a more socially desirable response of 5 when only option 3 is coded with 1 and all other options are coded with 0).

OK items present a more difficult task for a respondent who wishes to "fake good" because it is never completely clear which option will yield the maximal score for an item. A socially desirable responding (SDR) bias, for example, will inflate an OK score only if the most socially desired response is also the most highly weighted option. On the other hand, items keyed as a whole (IK method) will necessarily yield maximal scores when extreme responses are selected (assuming the sign is correct). A respondent, in fact, need only guess the direction of the item validity to effectively maximize the item score. Assuming that the sign of the item validity is the same as the sign of the correlation of the item with social desirability, SDR bias should result in score inflation for IK scores.

General social desirability is not the only bias that may influence applicant responses. Information about a particular job may result in a job-specific bias; that is, applicants will tend to present themselves as having characteristics deemed desirable for that specific job. Sometimes such a bias may be at odds with social desirability. For example, admission of past violent behavior, not usually socially desirable, may be deemed appropriate for the job of a bodyguard. In one of the few studies investigating this question, Schrader and Osburn (1977) did not find evidence of a job-specific bias. On the other hand, Longstaff and Jurgensen (1953) reported greater response bias when subjects knew the specific purpose of an inventory than when subjects were given only general instructions to fake good.

Unlike Schrader and Osburn (1977) , who used a quasi experiment, we investigated both SDR and job-specific biases in an experimental setting. We predicted that the job-specific bias, like the general social desirability bias, would be reflected in IK scores but not in OK scores.

Biodata keying strategies represent one approach to the study and possible control of score inflation due to response bias. An alternative approach was suggested by Mitchell (1987) , who surmised that applicants attempting to fake their answers may have longer response times than applicants who answer the items honestly. When answering honestly, one need only retrieve the information from memory to produce an answer. When one is motivated to distort the response, additional cognitive steps, involving the manipulation of information retrieved from memory prior to response, may increase response time. Two recent studies are of interest in the use of response latencies for the detection of faking. McManus (1990) failed to find significantly different latencies between subjects who were given instructions to fake and a control group, although she did find a significantly shorter mean latency for subjects given specific coaching on how to respond. McDaniel (1990) , however, reported significantly longer response times for subjects attempting to fake honesty-test items. Accordingly, we hypothesized that any motivated response bias should increase response time because of increased cognitive demand.

In the present study, we sought to determine whether IK scores, OK scores, and response latencies would differ when subjects were given instructions to simulate applying for a nonspecified job, to simulate applying for a specific job, or to answer honestly. We adapted two existing biodata instruments for a computerized laboratory experiment. The first was a Likert-type instrument developed and validated as a predictor of retail store manager performance (Russell & Domm, 1990). Both OK scores and IK scores were derived from the same Likert-type items. The second biodata instrument used a true/false response format and was designed to predict clerical job performance (

McDaniel, 1988). This instrument served two control purposes. First, scores obtained with a true/false format should be as susceptible to social desirability responding bias as IK scores, because a true/false format makes it easier for respondents to choose the more socially desirable alternative. Second, we instructed some of the subjects to imagine that they were applying for the job of store manager. We expected such instruction to affect only the instrument designed to select store managers but not the instrument designed to select clerks.

To further investigate the role of social desirability on score inflation, we asked independent judges to respond to both instruments by indicating the most socially desirable response for each item. This procedure allowed us to calculate a separate SDR score for all applicants. This SDR score allowed us to further differentiate between socially desirable responding per se and its resultant effects on empirically derived biodata scores.

On the basis of the foregoing discussion, we hypothesized the following:

- Subjects simulating job applicants will receive higher SDR scores than subjects responding honestly.
- SDR bias will inflate IK scores (and true/false scores) but not OK scores.
- Subjects who are given information about a specific target job will bias their responses. This bias will inflate IK scores of a biodata test designed for that job, but not OK scores derived from the same test, nor scores of a test designed for a job other than the target job.
- Subjects simulating job applicants will produce longer response latencies than subjects responding honestly.

Method

Subjects

Eighty-five graduate students enrolled in professional degree programs at two universities in the eastern United States participated in the study. These volunteer subjects were recruited from a human resources management class and from an applied psychology class. Instructors provided class time for participation. Students were taken to a computer lab, where they worked individually on a completely computerized task. One of us subsequently discussed the results with the students.

Dependent Measures OK and IK scores derived from a Likert-type scale.

To investigate the role of job specificity on response bias, we selected 25 valid Likert-type items that were not obviously related to the store manager job from a 67-item instrument (Russell & Domm, 1990). The item, "How often have you felt that you would do whatever it takes to get a job done?," falls into this category.

On the other hand, the question, "How often have you spoken at a store meeting?," is clearly job specific. The exclusion of job-specific items was necessary to avoid priming subjects for a particular job through clearly job-related item content.

For purposes of this study, we used the OK weights (for the 25 selected items) from Russell and Domm's original study, which were developed by keying options against store manager performance appraisal ratings. Russell and Domm added a constant of a 100 to the OK scores to avoid negative values; we followed their procedure. We also developed and cross-validated a new key for the same items using the IK method. For the IK score, each item received a positive or negative sign based on the item's Pearson correlation with the criterion; items with negative signs were reversed, and a unit-weight sum of the alternatives was calculated.

The unshrunken correlation between the 25-item OK scale (using Russell and Domm's, 1990 , original weights) and the criterion was .74 in an analysis sample (*N* = 608). In the same sample, we obtained an unshrunken correlation between the 25-item IK scale and the criterion of .72. A cross-validity of .32 for the OK method was obtained on a hold-out sample (*N* = 140), which was comparable to the cross-validity of .28 obtained for the IK scale in the same sample (*t* < 1 for differences between correlated correlations).

True/false scale.

A true/false scale was derived from an instrument based on a literature review of preexisting biodata items (McDaniel, 1988) and was designed to predict generic job performance. An example item was, "I have received a cash bonus due to my excellent job performance." All items were non-job specific and required a true/false response. We conducted a pilot test with an independent sample of graduate students (*N* = 37), drawn from the same student population, who simulated applying for a job. For some of the true/false items the same alternative was endorsed by all or almost all of the subjects. We therefore selected 25 of the 50 items from McDaniel's instrument that had endorsement rates of *true* approaching 50%. The removal of items with low or no variance reduced the chance of failing to support our hypotheses because of ceiling or floor effects. Keying was based on information provided by McDaniel; true and false alternatives were assigned weights of 0 and 1, respectively. A score for the true/false scale was created with a unit-weight sum of the item weights.

Social desirability scales.

Three advanced graduate students in the social sciences served as independent judges. Graduate students were considered as reasonable judges for two reasons: First, graduate students are familiar with the concept of social desirability, and second, graduate students' judgments are unlikely to be different from other populations. Wiggins (1973) reported that social desirability ratings obtained from graduate students are very similar to ratings obtained from raters from diverse cultural and subcultural backgrounds. Judges were first asked to answer both the true/false and Likert-type items in the most socially desirable way. Next, the judges were asked to indicate the *direction* of the most socially desirable response to the Likert-type items by using the extreme alternatives (1 and 5) only. All three judges agreed on 22 most desirable responses for 25 of the true/false alternatives in the true/false scale. In contrast, the judges agreed on only 4 most desirable alternatives for the 25 Likert-type items. However, the three judges agreed on the most desirable direction for 21 of these 25 Likert-type items.

On the basis of the judges' indication of the socially desirable responses for the true/false items and the socially desirable direction for the Likert-type items, two SDR keys were developed. For the few items on which the judges disagreed, the alternative chosen by two judges was keyed. The SDR key for the true/false items was very similar to the key developed by McDaniel (1988) ; that is, the socially desired alternative (true or false) was the same as the correct response for 22 out of the 25 items. For 21 items, all judges agreed on the correct response, for one item they all agreed on the incorrect response, and for one more item the majority chose the correct response, culminating in 22 keys identical to McDaniel's keys. The SDR direction key for the Likert-type items agreed with the IK weights (the sign of the item-criterion correlations) for 20 out of the 25 items. Although a highly transparent biodata instrument may seem to be invalid, Hogan and Stokes (1989) and Trent (1987) have reported biodata items for which social desirability was correlated with item validity. Furthermore the agreement between judges' ratings of social desirability and the IK keys is consistent with Hogan and Stokes' finding of a .66 correlation between item social desirability and empirically derived IK weight (predicting turnover).

Although the SDR keys were very similar to the IK and true/false keys, they were not identical. Therefore, we developed separate SDR scores for both the Likert and the true/false items. A sign (positive or negative) was assigned to each item on the basis of the judges' ratings, negative items were reversed, and a unit-weight sum was used as an index of SDR. The calculation of SDR scores was similar to the calculation of IK and true/false scores except that the keys were derived from SDR judgments and *not* from the original weights of the items.

Response latency measures.

Two additional dependent variables were created from subjects' item response latencies. The response latencies for all Likert-type and true/false items were obtained by measuring the time between item presentation and the keying of a response (registered to the .01 s). Total time spent per questionnaire was computed by summing response latencies for the 25 items within each biodata test.

In summary, we created five biodata scores: (a) a biodata score based on the OK weights developed by Russell and Domm (1990) for the Likert-type items; (b) a biodata score computed with the IK method for the Likert-type items; (c) a biodata score for the true/false items using McDaniel's (1988) key; (d) a Likert SDR score, and (e) a true/false SDR score. In addition, we created (f) a response latency score for the Likert-type items and (g) for the true/false items.

Procedure

The experiment was conducted in three different computer laboratories. Between 18 and 32 subjects were tested per session. Instructions, the experimental task, measures, and manipulations were all administered on personal computers.

Subjects arrived at the facility and were randomly assigned to an experimental condition. The sample size within experimental cells was not equal (see Table 2) because each computer independently and randomly assigned subjects to the experimental cells. A (2) × 2 × 2 two-period crossover design was employed, with one within-subjects factor and two between-subjects factors. This design offers the advantage of an increase in statistical power under most circumstances. The within-subjects factor was manipulated by asking that subjects respond once honestly and for research purposes only (honest condition) and once as they would as job applicants (simulated applicant [SA] condition). Therefore, all subjects completed both 25-item biodata instruments adapted for this experiment twice. Subjects were randomly assigned the order of presentation of the honest and simulated applicant instructions. The order of this manipulation was one of the between-subjects factors. The second between-subjects factor, job specificity, was manipulated as follows: Under the general condition, subjects were asked to imagine that they were "applying for a job"; under the specific condition, subjects were asked to imagine that they were "applying for the job of retail store manager."

The honest and SA instructions paralleled procedures described by Trent (1987) . Specifically, in the honest condition, subjects were given the following message: "Now, please answer the following questionnaire keeping in mind that it is used only for research purposes. We are interested in the most HONEST response you can give." This message was augmented by a reminder: "REMEMBER, try to answer as honestly as you can." In the SA condition the following message was delivered: "Now, please answer the following questionnaire imagining that your responses will determine whether you will get a job or not. Assume that you really want the job." This message was augmented by the following reminder: "REMEMBER, assume that you are actually applying for a job." Subjects were then debriefed and dismissed.

Although a within-subjects crossover design is more complex than a between-subjects design, it offers the advantage of increased statistical power. When the correlation between a measure in Period 1 with the same measure in Period 2 is zero, the statistical power of the within-subjects crossover design is twice as large as the power of a between-subjects design (Cotton, 1989). When this correlation is positive, power further increases as a function of the correlation size. However, if the main effect of the variable of interest–in this case the effect of honest versus SA instructions–in the first period differs from the main effect in the second period, an asymmetric transfer problem arises. A significant Treatment × Period interaction, that is, asymmetric transfer (not to be confused with a Treatment × Order interaction), has at least two alternative causal interpretations (see Cotton). If a significant asymmetric transfer is found, Cotton recommended that only data from the first period be analyzed, as in a between-subjects design.

Results

Table 1 contains descriptive statistics and intercorrelations among the various scales (collapsed across periods of administration). Table 2 shows the means for all three experimental conditions.

Before testing for experimental effects, it was necessary to check for asymmetric transfer, that is, whether the size of the main effect of the honest versus SA manipulation differed as a function of period of presentation. Using the *F* tests adapted from Cotton (1989) , we found an asymmetric transfer effect for the OK measure (see Table 3 , column 1). This asymmetric transfer can be seen in the patterns of the OK means (Table 2). Under the general condition, for example, the first-period SA mean taken from the SA/honest column (99.85) was compared with the first-period honest mean in the honest/SA column (100.80). In a similar fashion, the means for the second period were compared. The means in both the general and specific manipulations suggest that, in the first period, the honest mean was higher than the SA mean but that in the second period the pattern was reversed. Consequently, we interpreted all tests of effects on the OK score for first-period data only. Thus, for the OK score, the test of the honest versus SA effect was between-subjects. For all other measures we used complete data.

Hypotheses 1 and 2: Simulated Applicant Effect on SDR and Biodata Scores

For the SDR scores, the IK score, and the true/false score, the SA condition yielded significantly higher scores than the honest condition (column 2 in Table 3). These findings support our first two hypotheses: Subjects under SA conditions responded in a more socially desirable way than did subjects responding honestly, and the higher SDR was paired with inflated IK and true/false scores. We conducted a separate analysis of data from the first period for the biodata items keyed with the OK method (due to the asymmetric transfer effect). Results indicate that the SA condition yielded a lower OK score than the honest condition, $F(1, 81) = 4.54$, $p < .05$. This finding is in agreement with the second hypothesis, which also suggested that OK biodata scores are not susceptible to inflation due to response bias.

We performed an additional analysis on the Likert-type items to facilitate interpretation of the effects of the SA condition on biodata scores. A count of extreme scores was computed for each subject (i.e., the frequency of choosing 1 or 5 on the 25-item instrument). Subjects in the SA condition gave significantly more extreme responses (*M* = 4.41 items; *SD* = 2.42) than did subjects in the honest condition (*M* = 2.21 items; *SD* = 1.59), *F* (1, 81) = 60.45, *p* < .001 (repeated measure). This pattern partially explains the difference between OK and IK scoring methods in susceptibility to biases; that is, the increase in extreme scores entails an increase in IK scores, but it may increase, decrease, or leave unchanged OK scores.

Hypothesis 3: Specific Job Bias

As can be seen in Table 2 , within each of the other manipulations, all the respective Likert IK and Likert SDR means of subjects who were given the specific target job instruction were higher than the means of subjects who received the general instruction. This effect of job specificity on Likert IK scores, *F* (1, 83) = 1.94, *p* < .16, and Likert SDR scores, *F* (1, 83) = 2.84, *p* < .10, was in the predicted direction but only marginally statistically significant. However, in Period 1, these effects were statistically significant, *F* (1, 81) = 5.11, *p* < .05, and *F* (1, 81) = 5.43, *p* < .05, respectively. (The IK means in Period 1 were 80.5 [*SD* = 6.7] for the job-specific instructions versus 77.6 [*SD* = 5.7] for the general instructions. The respective SDR means were 83.9 [*SD* = 8.9] and 80.3 [*SD* = 6.6]). It well may be that when the second-period manipulation was introduced–reversal of SA and honest instructions–subjects' attention was focused on the novel instruction and not on the job-specificity instruction already delivered in the first period. Instruction specificity had no additional main effects or interaction effects on any other dependent variable. As expected, the true/false score designed to predict clerical job performance was not affected by the job-specificity (store manager) manipulation, nor was the OK score. This pattern of findings lends some support to Hypothesis 3.

Hypothesis 4: Response Latencies

We found no significant differences in response latencies for either honest/SA or job-specificity manipulations and thus could not support Hypothesis 4. The response-latency measures were only slightly positively skewed (skewedness in all cases < 1.3) and were sensitive enough to detect period effects (see next paragraph and Table 3). Removal of a few outliers and a separate square root transformation reduced skewedness but did not change the results.

Finally, the period effects (column 3 in Table 3) suggest that IK, true/false and both SDR scores were higher in the second administration, regardless of other conditions. The period effects also indicate that subjects responded more quickly to items during the second administration.

Discussion

The results supported our first three hypotheses: Strong effects, approaching one standard deviation, of an instruction to simulate applying for a job were observed for the SDR, IK and true/false scores. SDR bias occurred simultaneously with inflated IK scores but not with OK scores. OK scores did not rise and actually decreased under the SA condition, demonstrating that keying strategy has considerable implications for inflationary response bias. Our results support Mumford and Owens's (1987) speculation that differences in biodata instruments' susceptibility to faking may be related to the type of keying procedure used.

This interpretation is supported by our findings that, under the SA condition, subjects chose more extreme alternatives. With the present empirical OK weights, only 6 out of 25 extreme responses led to maximal item scores. Second, because OK weights were not highly correlated with SDR (see Table 1), the keys were not clearly transparent to simulated applicants. Lack of transparency was evidenced by the inability of our judges to agree on the exact SDR categories. This was in sharp contrast to the judges' high rate of agreement on the extreme SDR direction, and the high rate of overlap between the judges' SDR direction and the signs of the IK weights.

A recent paper by Crosby (1990) supports these points by demonstrating that the tendency to respond in a socially desirable way is uncorrelated with OK biodata scores. Crosby found a nonsignificant correlation between scores on the Marlowe-Crown Social Desirability Scale and a biodata instrument developed with the OK method. Our results suggest that IK weights are transparent (and the IK method susceptible to inflationary response bias), but not OK weights. Our results are consistent with and extend Crosby's finding, indicating that the same Likert-type items are differentially susceptible to inflationary biases depending on the scoring method used.

Crosby's (1990) results and the relatively low correlations between the OK and IK scores for our experimental data (see Table 1) led us to speculate that the two keying procedures may be capturing different aspects of criterion variance. We tested this speculation by assessing the incremental validity of IK scores after the OK scores had been entered in a hierarchical regression analysis. The 25 experimental items taken from Russell and Domm's (1990) original cross-validation sample were used for this analysis. The criterion for this analysis was the overall performance rating provided by the store managers' supervisors. The relatively low correlation between the IK and OK scores ($r = .32$) and the significant increment (.07) in the squared multiple correlation ($R^2 = .39$), F (1, 137) = 5.58: $p < .05$ suggest an intriguing notion; combining two different types of keys derived from the same items may increase validity. (The validities of the IK and OK scores were .28 and .32, respectively.) In this analysis both the OK and IK regression coefficients were positive. Yet, Crosby (1990) found that the Marlowe—Crown Social Desirability Scale had a negative weight. This difference in the direction of the contribution of IK or SDR scores above and beyond the OK scores may be job specific.

Although the present research emphasized the susceptibility of biodata scores to inflation, our results suggest that SDR may lower the validity of biodata instruments regardless of the keying method used. Inflation, observed for the IK and true/false keys, is essentially a systematic bias that should lower validity by reducing valid predictor variance. Increased SDR did not systematically inflate OK scores, but did produce an effect resembling the introduction of random error. For all scores, correlations between the same respondent's scores under SA and honest conditions were not as high as might be expected for a test—retest reliability. The .39 correlation between the OK scores obtained under honest and SA conditions (see the diagonal of the square matrix in Table 1) indicates that OK scores were affected, even if not inflated. Therefore, keys developed under research conditions (e.g., with incumbents) seem to be susceptible to lowered validity from SDR even though OK scores are not, on average, inflated. Consistent with this interpretation, Hogan and Stokes (1989) found a correlation of .40 between concurrent and predictive item validities, and no overlap between the items entering the two different empirical keys.

The third hypothesis was supported by a significant effect of the job-knowledge manipulation on IK scores in Period 1. As expected, the scores on the true/false items that were designed for a different (clerical) job were not affected by the specific job-knowledge manipulation. Also, the specific job-knowledge manipulation did not interact with the SA manipulation. Paulhus (1984) distinguished two components of SDR: unconscious self-deception and impression management. The lack of interaction between the job-knowledge condition and the SA condition suggests that the job-knowledge effect was an unconscious bias; if conscious, it would have been found exclusively in the SA condition. However, Paulhus (1984) found that unconscious self-deception (versus impression management) was not affected by a laboratory manipulation, suggesting that self-deception may be resistant to laboratory manipulation. Although the nature of the specific job bias we observed is not clear, if this result is replicated it will suggest that SDR is not the only threat to the accuracy of biodata scores. Other biases, for example, job-specific bias, may also reduce the validity of these instruments.

Our job-knowledge manipulation was weak; only slight changes in the instructional sets were made to convey the job title of the target position. Researchers will need to examine the effects of alternative manipulations reflecting naturally occurring conditions. For example, different recruiting methods and information are often used in different labor markets (Wanous & Colella, 1989). The likelihood of both SDR and specific response biases occurring may vary depending on the type of recruiting and the job information given.

Although our results confirm prior findings regarding the presence of faking in biodata item responses, it would be useful to know how various concrete indicators of motivation may influence faking (e.g., number of months unemployed, size of status or pay increases relative to current position, etc.). In the field, motivation to fake may be a more important factor than our laboratory experiment indicates. On the other hand, a propensity to fake among actual applicants may be mitigated by the fear of being caught lying. The potential differences between laboratory and actual applicants call for a replication of our results in the field. For example, the effect of a warning not to fake on IK and OK scores could be studied in a field experiment, as could the validity of combining the alternative keying strategies.

We were unable to detect any effect of the experimental manipulation on response time, but it is interesting that significant changes in the subjects' behavior left no traces in the response latency measures. Indeed, a review of the research on nonverbal cues of deception also suggests that response latencies are not predictive of deceptive behavior (Zuckerman, DePaulo, & Rosenthal, 1986), although other dimensions of responses, such as number of words in the response, were found to be correlated with deception (Zuckerman et al., 1986). In contrast, McDaniel (1990) reported significant differences in response times between subjects who were faking integrity test items and subjects who were answering honestly. However, he indicated that the items he used were extremely emotionally charged. Mere processing time for faked and honest responses may not differ, but responses about behaviors at odds with basic moral imperatives may cause emotional conflict that slows down production time. McManus (1990) also investigated response latencies to biodata items. Her results were similar to ours when fake and honest groups were compared, that is, no significant difference between these conditions. However, McManus did find a significantly lower mean response latency for a group that was coached on how to answer the items. Apparently, briefing on how to respond allows subjects to know what to expect and results in fewer cognitive steps in making a response. McManus's results are consistent with our finding that mean response latencies for all subjects were significantly shorter during the second administration of the instrument. Familiarity with the items reduced the amount of cognitive processing involved. Our results, taken together with the results of McManus (1990) and McDaniel (1990) , suggest that using response time to detect faking may be difficult and complex. Response latencies may depend on a number of factors, including the item type, the amount of information given to the candidate about the instrument, and the candidate's previous experience with the instrument. These issues require further empirical investigation.

From a different perspective, our findings may be generalized to other individual difference measures. For example, personality measures are known to correlate with measures of social desirability (e.g., Nicholson & Hogan, 1990 ; see also Edwards, 1990 , and Walsh, 1990). It is an empirical question whether the robustness of the OK method against SDR inflation found here will generalize to personality variables.

Conclusion

Our results allow some insight into the problem of distortion on empirically keyed biodata forms. Our data suggest that OK biodata scores are not susceptible to response biases, whereas IK biodata scores are. This finding explains some of the inconsistencies reported in the literature regarding how susceptible biodata scores are to faking. In addition, we provide some evidence for the existence of response biases other than SDR, such as job-specific bias. Also, the null finding of no response latencies as a function of type of instruction (honest vs. SA) may be useful. Combined with the results of other studies, our data also suggest caution in the use of response latencies as a method of detecting faking or cheating.

Finally, the differences between IK and OK investigated here have several important implications for the use of biodata in selection. First, our post hoc analysis suggests that OK and IK scores may be capturing different aspects of criterion variance. Key developers should routinely check whether a combination of both types of keys improves validity. Second, the response distortion of (simulated) job applicants seems to take the form of increased SDR, which should in turn lead to lowered validity, regardless of the type of keying used. The results suggest that keys should be developed on applicant samples in which the same motivational sets are operating. Third, one way to control invalid SDR responses is to warn subjects that attempting to fake will offer them no advantage. Trent et al. (1986) demonstrated that warnings can effectively mitigate the tendency to fake. Because OK scores are not inflated by response biases, it is morally defensible to warn applicants against faking that offers them no advantage. Finally, the effect of faking on validity may depend on the specific job performance being predicted. Jobs that require socially desired behavior, for example, working with other people, may be better predicted by IK scores, whereas jobs that do not require socially desired behavior may be better predicted by OK scores.

References

Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology, 60*, 576-580.

Colquitt, A. L. & Becker, T. E. (1989, April). *Faking of a biodata form in use: A field study.* (Paper presented at the Fourth Annual Meeting of the Society for Industrial/Organizational Psychology, Boston, MA)

Cotton, J. W. (1989). Interpreting data from two-period crossover design (also termed the replicated 2 × 2 Latin square design). *Psychological Bulletin, 106*, 503-515.

Crosby, M. M. (1990, April). *Social desirability and biodata: Predicting sales success.* (Paper presented at the Fifth Annual Meeting of the Society for Industrial/Organizational Psychology, Miami, FL)

Edwards, A. L. (1990). Construct validity and social desirability. *American Psychologist, 45*, 287-289.

England, G. W. (1961). *Development and use of weighted application blanks.* (Dubuque, IA: W. C. Brown)

Fleishman, E. A. (1988). Some new frontiers in personnel selection research. *Personnel Psychology, 41*, 679-701.

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

Hogan, J. B. & Stokes, G. S. (1989, April). *The influence of socially desirable responding on biographical data of applicant versus incumbent samples: Implications for predictive and concurrent research designs.* (Paper presented at the Fourth Annual Meeting of the Society for Industrial/Organizational Psychology, Boston, MA)

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Keating, H. G., Patterson, D. G. & Stone, C. H. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34*, 1-5.

Klein, S. P. & Owens, W. A. (1965). Faking of a scored life history blank as a function of criterion objectivity. *Journal of Applied Psychology, 49*, 452-454.

Lecznar, W. B. & Dailey, J. T. (1950). Keying biographical inventories in classification test batteries. *American Psychologist, 5*, -279.

Longstaff, H. P. & Jurgensen, C. E. (1953). Fakability of the Jurgensen classification inventory. *Journal of Applied Psychology, 37*, 86-89.

McDaniel, M. A. (1988). *Experimental personnel survey.* (Unpublished manuscript)

McDaniel, M. A. (1990). *Lying takes time: Predicting deception in biodata using response latencies.* (Paper presented at the 98th Annual Convention of the American Psychological Association, Boston, MA)

McManus, M. A. (1990). *Detection of faking on an empirically keyed biodata instrument.* (Paper presented at the Fifth Annual Conference of the Society for Industrial/Organizational Psychology, Miami, FL)

Mitchell, T. W. (1987). *Electronic mechanisms for controlling false biodata in computerized selection testing.* (Paper presented at the 95th Annual Convention of the American Psychological Association, New York)

Mosel, J. L. & Cozan, L. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 36*, 365-369.

Mumford, M. D. & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1-31.

Nicholson, R. A. & Hogan, R. (1990). The construct validity of social desirability. *American Psychologist, 45*, 290-291.

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

Reilly, R. R. & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-63.

Reilly, R. R. & Warech, M. W. (1990). *The validity and fairness of alternatives to cognitive test* [Report to the Commission on Testing and Public Policy, Berkeley, CA].(Unpublished manuscript)

Russell, C. J. & Domm, D. R. (1990, April). *On the construct validity of biographical information: Evaluation of a theory-based method of item generation.* (Paper presented at the Fifth Annual Meeting of the Society for Industrial/Organizational Psychology, Miami Beach, FL)

Schmitt, N., Gooding, R. Z., Noe, R. A. & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schrader, A. D. & Osburn, H. G. (1977). Biodata faking: Effects of induced subtlety and position specificity. *Personnel Psychology, 30*, 395-404.

Trent, T. T. (1987, August). *Armed Forces adaptability screening: The problem of item response distortion.* (Paper presented at the 95th Annual Convention of the American Psychological Association, New York)

Trent, T. T., Atwater, D. C. & Abrahams, N. M. (1986, May). *Biographical screening of military applicants: Experimental assessment of item response distortion.* (Paper presented at the Tenth Psychology in the Department of Defense Symposium, Denver, CO)

Walsh, J. A. (1990). Comment on social desirability. *American Psychologist, 45*, 290-291.

Wanous, J. P. & Colella, A. (1989). Organizational entry research: Current status and future directions.(In K. Rowland & G. Ferris (Eds.), *Research in personnel and human resource management* (Vol. 7, pp. 59—120). Greenwich, CT: JAI Press.)

Weiss, D. J. & Dawis, R. V. (1960). An objective validation of factual interview data. *Journal of Applied Psychology, 44*, 381-385.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* (Reading, MA: Addison Wesley)

Zuckerman, M., DePaulo, B. & Rosenthal, R. (1986). Humans as deceivers and lie detectors.(In P. D. Blanck, R. Buck, & R. Rosenthal (Eds.), *Nonverbal communication in the clinical context* (pp. 1—35). University Park, PA: Pennsylvania State University Press.)

1

Russell and Domm (1990) made their data available to us, allowing us to perform several secondary analyses for the present paper. Russell and Domm's instrument can be obtained from Craig Russell.

Table 1.

[illegible]

Table 2.

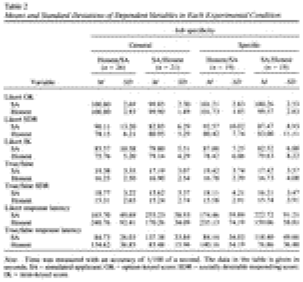

Table 2

Means and Standard Deviations of Dependent Variables in Each Experimental Condition

	Job specificity							
	General				Specific			
	Honest/SA (*n* = 26)		SA/Honest (*n* = 21)		Honest/SA (*n* = 19)		SA/Honest (*n* = 19)	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Likert OK								
SA	100.60	2.69	99.85	2.50	101.21	3.63	100.26	2.53
Honest	100.80	2.43	99.90	1.89	101.73	1.65	99.57	2.63
Likert SDR								
SA	90.11	13.20	82.85	6.29	92.57	10.02	87.47	8.93
Honest	78.15	6.21	80.95	5.29	80.42	7.76	83.00	11.11
Likert IK								
SA	83.57	10.58	78.80	5.51	87.00	7.25	82.52	4.88
Honest	73.76	5.30	75.14	4.29	78.42	6.00	75.83	8.33
True/false								
SA	19.38	3.51	17.19	3.07	19.42	3.74	17.42	3.57
Honest	16.23	2.50	16.90	2.54	16.78	3.59	16.73	4.00
True/false SDR								
SA	18.77	3.22	15.62	3.37	18.11	4.21	16.21	3.47
Honest	15.31	2.63	15.24	2.74	15.06	2.91	15.74	3.91
Likert response latency								
SA	165.70	40.69	255.23	38.53	174.46	39.89	222.72	91.21
Honest	240.76	92.41	170.26	34.09	235.13	74.19	159.06	58.01
True/false response latency								
SA	84.73	26.03	137.38	33.84	84.16	34.03	118.89	49.00
Honest	134.62	36.83	83.48	15.94	140.16	34.19	78.84	38.40

Note. Time was measured with an accuracy of 1/100 of a second. The data in the table is given in seconds. SA = simulated applicant; OK = option-keyed score; SDR = socially desirable responding score; IK = item-keyed score.

Table 3.

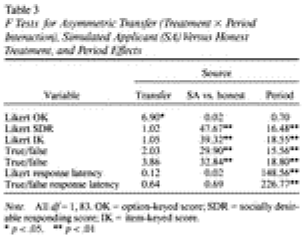

Table 3

F Tests for Asymmetric Transfer (Treatment × Period Interaction), Simulated Applicant (SA) Versus Honest Treatment, and Period Effects

	Source		
Variable	Transfer	SA vs. honest	Period
Likert OK	6.90*	0.02	0.70
Likert SDR	1.02	47.67**	16.48**
Likert IK	1.05	39.32**	18.55**
True/false	2.03	29.90**	15.56**
True/false	3.86	32.84**	18.80**
Likert response latency	0.12	0.02	148.56**
True/false response latency	0.64	0.69	226.77**

Note. All *df* = 1, 83. OK = option-keyed score; SDR = socially desirable responding score; IK = item-keyed score.

* $p < .05$. ** $p < .01$