Building a Better Mousetrap: Using Cart to Detect Response Distortion

Richard L. Frei Richard L. Griffith Michael A. McDaniel Andrea F. Snell
University of Akron

Laura S. Hamill Jill K. Wheeler
Bell Atlantic

Elizabeth F. Douglas
University of Akron

Paper presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology. St. Louis. April. 1997

Abstract

This paper demonstrates the use of classification and regression tress as a method of identifying items that differentiate between honest and faking respondents on non-cognitive measures. Such items may permit improved identification of applicants who misrepresent when responding to applicant screening tests.

Building a Better Mousetrap: Using Cart to Detect Response Distortion

Recent attention to personality constructs as predictors of job performance has brought the issue of response distortion to the forefront. It has been demonstrated that individuals can distort their responses to non-cognitive instruments in order to make themselves appear more attractive (Hough, Eaton, Dunnette, Kamp & McCloy, 1990; Ones, Viswesvaran, & Korbin, 1995; Schmit & Ryan, 1994). One method to detect faking is the use of lie scales. However, lie scales have been criticized because 1) the use of social desirability scores to correct personality scale scores removed significant variance from the measures and 2) they often use "unlikely virtues" items to identify socially desirable responding. These items are highly transparent and often utilize very extreme statements. The purpose of this paper is to examine the use of the Classification and Regression Trees statistical package (CART) as 1) a method of identifying fakers and 2) an alternative method for developing potentially more subtle or more effective response distortion scales.

Faking and Response Distortion

Response distortion is one of the most frequently cited criticisms of personality testing for personnel selection. Response distortion has been described in the literature with terms such as socially desirable responding, claiming unlikely virtues, impression management, faking, intentional distortion, and self enhancement (Hough et al., 1990; Hough & Paullin, 1994; Lautenschlager, 1994; Ones, Viswesvaran, & Reiss, 1996). Although researchers have differentiated between "faking good" and "faking bad", personnel psychologists are generally more concerned with applicants who attempt to present themselves in an overly positive light.

Research on faking has found that subjects can, in fact, distort their responses on non-cognitive measures when asked to do so (Hough et al., 1990, Lautenschlager, 1994). Ones, et al. (1995) found that, when instructed to fake good, respondents were able to change their responses by almost half a standard deviation for the Big Five factors. Both Hough et al. (1990) and Lautenschlager (1994) have identified several strategies that were developed to overcome the problems associated with intentional response distortion. For example, one method employed to reduce faking is the use of a forced-choice question format. Forced-choice questions are created

by pairing statements judged to be similar in social desirability. By making respondents choose between equally desirable (or equally undesirable) statements, the ability to self-distort is supposedly reduced. However, in an early review of forced-choice scales, Waters (1965) found that respondents could successfully distort their self-descriptions when instructed to do so.

A second strategy is the use of subtle items; that is, items for which the underlying construct is not apparent. Most personality inventories, especially those used in personnel selection, use questions that are fairly transparent with regards to the underlying constructs. For example, Cook (1993) has hypothesized that no one applying for a customer service job is likely to respond positively to "I don't like talking to strangers." It would seem obvious that subtle items are less susceptible to faking than transparent items. Alliger, Lilenfield, and Mitchell (1995) found that subtle questions were less easily faked than questions in which it was obvious what the measure was assessing. In a meta-analysis of fakability estimates, Ones et al. (1995) found that sets of obvious items were more susceptible to faking instructrions than subtle items.

In addition, researchers have used subtle items to actually detect faking. Dannenbaum and Lanyon (1993) found that subjects who faked psychopathology on the MMPI have lower scores on subtle subscales than non-faking subjects because the subtle subscales are keyed for psychopathology in a direction that is the opposite of their face validity. An example of a subtle item from the MMPI is "I enjoy detective or mystery stories." Thus, subjects who attempted to fake psychopathology did so on the basis of item content and achieved lower, rather than higher, scores. However, some researchers have found that respondents can, in fact, distort their answers on empirically-keyed measures (Haymaker & Erwin, 1980). Further, the use of subtle items with little face validity for selection purposes can potentially lead to negative applicant reactions. For example, Rynes and Connerley (1993) found that applicants prefer selection methods with high apparent content validity.

A third strategy for reducing response distortion is to warn respondents about the consequences of distorting their responses. This is done by telling respondents that methods exist to detect faking and that test scores will be verified using outside sources. Wheeler, Hamill, and Tippins (1996) administered a personality battery to subjects over two time periods. Preceding the first administration, subjects were given a verbal warning not to fake. No such warning was given preceding the second administration. Scores for the second administration were substantially

higher, indicating that the warning given at Time 1 was effective in reducing faking. Other research investigating the effectiveness of such warnings (Haymaker & Erwin, 1980; Lautenschlager & Atwater, 1986) have found that they do reduce that amount of intentional distortion in self-report instruments.

A fourth strategy is the use of subscales to identify response distortion. Many common personality tests use response validity scales (or social desirability scores) to either identify suspicious self-descriptions or correct for socially-desirable responding. The Adjective Checklist, the CPI, the Hogan Personality Inventory, and the MMPI all contain response validity scales. Many of these lie scales use an "unlikely virtues" approach to identify socially desirable responding. Faking is assessed according to how frequently respondents endorse items that are high in virtue (i.e., "I have never stolen anything, even as a child"; "I'm always happy and pleasant."). However, correcting for response distortion has had negligible effects on predictive validity. For example, Christiansen, Goffin, Johnston, and Rothstein (1994) looked at the effects of correcting the 16PF for faking by administering the test to 500 assessment center candidates. They found no substantial differences in the multiple correlations associated with uncorrected and corrected scores, suggesting that correcting the 16PF scores for faking was unlikely to affect criterion-related validity. Results of a meta-analysis conducted by Ones et al., (1996) indicated that social desirability is substantially correlated with the personality factors of emotional stability and conscientiousness. Further, these authors found that the use of social desirability scores to correct personality scale scores removed significant variance from the measures.

Hogan (1991) stated that the ability to enhance scores on a personality inventory is, itself, an important personality variable. Both Hogan (1991) and Ones et al. (1996) concluded that response distortion is not a serious problem in personality testing, but rather an important individual difference variable. However, in a recent study on the effects of faking on validity, Douglas, McDaniel, and Snell (1996) found that faking decreased both the criterion and construct validity of these measures. Frei, Griffith, McDaniel, Snell, and Douglas (1997) conducted a test of factor invariance using LISREL and found major structural differences between faking and honest respondents. Similar results were found by Griffith, Frei, Snell, Hamill, and Wheeler (1997). Thus, these latter three studies have identified serious problems with the construct validity of personality tests when respondents intentionally distort their responses.

One potential problem with the use of lie scales is that unlikely virtues items are very transparent (and, thus, easier to fake). In addition, these items usually involve very extreme statements (i.e., "I have never gotten mad at my parents, even once"). It is also unclear how well social desirability scales can distinguish faking applicants from honest applicants. It might be helpful for the personnel practitioner to be able to identify potential response distortion using subtler items than those normally included in a lie scale. One technique that may aid in the development of subtler lie scales is classification and regression trees.

Classification and Regression Trees

CART (Classification and regression trees) is a statistical procedure that can be used to differentiate groups on item level data (Breiman, Friedman, Olshen, & Stone,1984). CART is a method of binary recursive partitioning and can be used to predict a dependent variable as a function of a number of independent variables. CART has been successfully used in a variety of research settings, including credit risk assessment, cancer diagnosis, infant mortality, and classification of radar images for the military. CART examines all of the independent variables and makes a series of binary splits based on the value of one independent variable, such that within groups variance on the dependent variable is minimized (Barnes, Welte, & Dintcheff, 1989). In this fashion, CART selects a subset of variables that optimally differentiate cases on the dependent variable.

This process results in the creation of a decision tree. Once the decision tree is completed CART tests the tree using cross validation. The sample is divided into 10 equal subsamples containing similar distributions of the dependent variable. The first 9 subsamples are used to construct a decision tree using all of the independent variables. This is referred to as the learning sample because CART utilizes this sample to create the decision rules used to classify cases in the tenth subsample, similar to a discriminant analysis. This tenth subsample is used to obtain initial misclassification rates. These error rates are then used to "prune" the tree of independent variables that are redundant or do not differentiate cases on the dependent variable. CART then creates a different set of subsamples and the process is repeated nine times. The result is an optimal tree that best differentiates cases using a minimal number of independent variables. Once a decision tree has been created, this tree can be used to classify new data.

In addition, CART provides *surrogate variables* which closely mimic the classification ability of each independent variable selected for the decision tree. To identify and rank potential surrogates, CART computes the *association*, a measure of how well the alternative split can predict the primary split. The highest possible value of the association is 1.0, which occurs when the surrogate produces exactly the same split, case by case, as the primary split. If no other variable can mimic the primary splitting criterion, no surrogates are provided.

CART has several advantages when compared to traditional statistical tools that may be used for similar types of analysis. For example, CART can use any combination of categorical and continuous variables rendering recoding of the variables unnecessary. CART is also extremely robust to the effects of outliers, and because CART develops alternative splits (surrogates) it can still classify a case when there are missing values for the primary splitting value (Steinberg & Colla, 1995). When CART analyses are compared to stepwise logistic regressions or discriminant analyses it typically performs about 10% to 15% better in classification accuracy.

Current Study

The following studies show how CART may be used to identify potential fakers and how CART can be used to develop lie scales with less transparent items. In Study 1, we tested whether CART could accurately identify subjects who were instructed to fake. In Study 2, we used a lie scale to classify applicants as fakers/non-fakers. We then conducted CART analyses to identify additional items that differentiated between the two groups beyond those transparent items traditionally included in lie scales.

STUDY 1

Method

Sample

The sample for this study consisted of 600 introduction to psychology students attending a large Midwest university. Subjects received extra credit for participating in the study. Subjects were randomly assigned to one of two conditions: 293 were placed into the honest condition and 307 were placed in the faking condition.

Measures

Independent Variables. A 64 item temperament scale measuring conscientiousness and agreeableness was administered. The scale consists of single statement items requiring a true/false response. In addition, a 60 item multiple response biodata questionnaire was administered which also tapped the constructs of conscientiousness and agreeableness. Responses to both instruments were treated as categorical in the CART analysis.

Dependent Variables. The dependent variable used in the CART analysis was the dichotomous assignment to honest and faking conditions used by Douglas, McDaniel, and Snell (1996). The dependent variable was entered in the CART analysis as a categorical variable with 2 levels.

Procedure

Subjects were randomly assigned to one of two conditions. In the first condition subjects were instructed to complete the temperament and biodata questionnaires as honestly as possible (honest condition). In the second condition subjects completed the same measures, but were instructed to answer the questions in such a way as to make as favorable an impression as possible (faking condition).

The CART analysis selected 3 items with 4 terminal nodes that optimally distinguished between the honest and faking groups. The first split occurred on the item "*When I have started projects, I have finished them...*". Alternatives ranged from *no matter what* (1) to *never* (4) and were treated as categorical variables. CART split the cases at 1.5, assigning subjects to the honest group if they scored higher than 1.5 on this item. Based on this split, 310 cases were classified as honest and assigned to a terminal node. The remaining 290 cases were assigned to a nonterminal node for further analysis.

CART then reclassified those two hundred and ninety cases, splitting them on the item "*The longest I have held a grudge is...*". Responses ranged from *since childhood* (1) to *a day, if that* (4) and again were treated as categorical variables. CART split the cases at 3.5, assigning subjects to the faking group if they scored higher than 3.5 on this item. Based on this split, 228 cases were classified as faking and assigned to a terminal node.The remaining 62 cases were assigned to a nonterminal node for further analysis.

Lastly, CART reclassified the remaining sixty-two cases , splitting them on a true/false item *"I tend to hold grudges"*. The item was entered as a categorical variable with a value of 1 assigned to true responses and 2 assigned to false responses. CART split the item at 1.5, assigning those who answered true to this item into the honest group and those who answered false to this item in the faking group. Based on this split, 23 cases were classified as honest and the remaining 39 cases were classified as faking.

The items were able to classify honest respondents correctly at a rate of 83% and correctly classify subjects in the faking condition at a rate of 71%. In this data set, the surrogates CART identified had low to moderate association values (.165 to .486) for each splitting variable, indicating that the three items identified by CART best differentiated between groups.

STUDY 2

Method

Sample

Subjects were 4500 applicants for non-managerial positions in a large East coast communications firm.

Measures

Independent Variables. The independent variables were collected from a 116 item biodata scale designed to measure persuasiveness, ambition, energy, reliability, people orientation, and social adjustment. Eight transparent unlikely virtues items were used in calculating the dependent variable and, thus, were not included in the CART analysis. Six non-random response items were also deleted from the scale. The remaining 102 items were included in the analysis.

Dependent Variable. Subjects were classified as honest or faking based on their responses to the unlikely virtues scale. The eight unlikely virtues items were combined into a scale score, ranging from 8 to 24. Respondents scoring high on the unlikely virtues scale (21 and above) were used as the faking group, while subjects scoring low (11 and below) served as the non-faking group. A total of 353 applicants were included in the CART analysis, 161 in the faking group and 192 in the non-faking group and the dependent variable was entered as a categorical variable with 2 levels.

Results

The CART analysis selected 3 items with 4 terminal nodes that optimally distinguished between the honest and faking groups. Since this scale is currently in use as a selection tool, exact item content cannot be revealed in order to maintain test integrity. The first split occurred on the item measuring the social adjustment construct. Alternatives ranged from *strongly agree* (1) to *disagree* (3) and were treated as categorical variables. CART split the cases at 2.5, assigning subjects to the honest group if they scored higher than 2.5 on this item. Based on this split, 127 cases were classified as honest and assigned to a terminal node. The remaining 226 cases were assigned to a nonterminal node for further analysis.

CART then reclassified those two hundred and twenty six cases, splitting them on the item measuring the construct of ambition. Responses ranged from *strongly agree* (1) to *disagree* (3) and were treated as categorical variables. CART split the cases at 1.5, assigning subjects to the honest group if they scored lower than 1.5 on this item. Based on this split, 50 cases were classified as honest and assigned to a terminal node. The remaining 176 cases were assigned to a nonterminal node for further analysis.

Lastly, CART reclassified the remaining 176, splitting them on another item measuring the social adjustment construct. Responses ranged from *strongly agree* (1) to *disagree* (3) and were treated as categorical variables. CART split the item at 2.5, assigning those who scored less than 2.5 into the honest group and those who scored higher to the faking group. Based on this split, 10 cases were classified as honest, while the remaining 166 cases were classified as faking.

Based on the responses to these three items, honest respondents were correctly classified at a rate of 76% and subjects in the faking condition were correctly classified at a rate of 75%. In this data set, the surrogate items identified by CART had low association values (.05 to .37) for each splitting variable, indicating that the three items identified by CART best differentiated between groups.

Discussion

CART has been widely used in natural and social science research settings, but is relatively new to the field of industrial/organizational psychology. It gives both the practitioner and the

researcher a powerful statistical tool with which to differentiate between groups using a minimum number of variables.

CART also provides surrogate variables that mimic the classification ability of the primary splitting variables. The surrogate variables provide practitioners with a method of choosing alternative indicators that may be cheaper, shorter, less intrusive, or easier to collect. For example, a personnel psychologist could use CART to determine whether an inexpensive questionnaire could classify successful/unsuccessful applicants with reasonably the same accuracy as a more expensive interview.

In this paper, we used CART to identify students who were instructed to fake and to pinpoint items that could still differentiate between groups of faking and non-faking applicants beyond those transparent items included in the lie scale. Although CART classified our fakers at a rate of between 75%-80%, it still made classification errors. Practitioners might use this technique to flag applicants who have a propensity to inflate their qualifications and place additional emphasis on verifying their responses on other selection tools, such as resumes and application blanks. Additional research needs to be conducted to determine the relationship between CART-identified items and other response distortion scales.

References

Alliger, G.M., Lilenfield, S.O., & Mitchell, K.E. (1995). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7,* 32-39.

Barnes, G. M., Welte, J.W., & Dintcheff, B. (1991). Drinking among subgroups in the adult population of New York State: A classification analysis using CART. *Journal of Studies on Alcohol.* 52, 338-344.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and regression trees.* Pacific Grove. Wadsworth.

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.

Cook, M. (1993). *Personnel Selection and Productivity (rev. ed.).* New York: John Wiley and Sons Ltd.

Dannenbaum, S. E. & Lanyon, R. I. (1993). The use of subtle items in detecting deception. *Journal of Personality Assessment, 61,* 501-510.

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

Frei, R.L., Griffith, R.L.,McDaniel, M.A., Snell, A., & Douglas, E. F. (1997). *Faking of non-cognitive measures: factor invariance using multiple groups LISREL.* Paper presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology. April. St. Louis.

Griffith, R.L., Frei, R.L., Snell, A.F., Hamilll, L.S., & Wheeler, J.K. (1997). *Warning versus no warnings: Differential effect of method bias.* Paper presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology. April. St. Louis.

Haymaker, J. C. & Erwin, F W. (1980). Investigation of applicant responses and falsification detection procedures for the Military Applicant Profile (Final project report, work unit no. DA644520). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hogan, R. (1991). Personality and personality measurement. In M.D. Dunnette & L.M. Hough (Eds.) *Handbook of industrial and organizational psychology* (Vol. 2, pp. 327-396). Palo Alto, CA: Consulting Psychologists Press.

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

Hough, L.M. & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G.S. Stokes, M.D. Mumford, & W.A. Owens (eds.) *Biodata Handbook: Theory, Research, and Use of Biographical Information in Selection and Performance Prediction.* Palo Alto: Consulting Psychologist Press.

Lautenschlager, G.J. (1994). Accuracy and faking of background data. In G.S. Stokes, M.D. Mumford, & W.A. Owens (eds.) *Biodata Handbook: Theory, Research, and Use of Biographical Information in Selection and Performance Prediction.* Palo Alto: Consulting Psychologist Press.

Lautenschlager, G. J. & Atwater, D. C. (1986). Controlling response distortion on an empirically keyed biodata questionnaire. Unpublished manuscript.

Ones, D.S., Viswesvaran, C. & Korbin, W.P. (1995). Meta-analysis of fakability estimates: Between-subjects vs. within-subjects design. In F.L. Schmidt (Chair) *Response Distortion and Social Desirability in Personality Testing and Personnel Selection.* Symposium conducted at the 10th annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

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.

Rynes, S. L. (1993). Who's selecting whom? Effects of selection practices on applicant attitudes and behavior. In N. Schmitt & W. C. Borman (Eds.), *Personnel Selection in Organizations.* San Francisco: Jossey Bass Publishers.

Rynes, S. L., & Connerley, M. L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology, 7,* 261-277.

Schmit, M.J. & Ryan, A.M. (1993). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*, 966-974.

Steinberg, D. & Colla, P. (1995). *CART: Tree-structured non-parametric data analysis.* San Diego, CA. Salford Systems.

Waters, L. K. (1965). A note on the "fakability" of forced-choice scales. *Personnel Psychology, 18*, 187-191.

Wheeler, J. K., Hamill, L. S., & Tippins, N. T. (1996). *Warnings against candidate misrepresentations: Do they work?* Paper presented at the Eleventh Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.