Detection of Invalid Response Patterns on the California Psychological Inventory

Kevin Lanning
University of California, Berkeley, and Oregon State University

When faced with the task of responding to a personality questionnaire, an individual may respond with a number of strategies or test-taking attitudes. Among these, deceptive (fake) and disengaged (random) attitudes are of particular interest, for these can potentially mislead and misinform test users. A two-stage model was devised to detect deceptive and disengaged protocols on the California Psychological Inventory. Using parameters from signal detection theory, this model is found to be highly sensitive in detecting invalidity. *Index terms: California Psychological Inventory, expected utility, faking on personality inventories, personality assessment, random response patterns, signal detection theory.*

In personality testing, an implicit contract exists between respondent and administrator. In return for time and effort, the respondent may anticipate such rewards as self-knowledge in the form of an interpreted profile, vocational advice, diagnosis-specific psychotherapy, simple gratitude, a perceived benefit to science, or simply the completion of an experimental credit-hour requirement. To the extent that these and similar rewards are valued, respondents will become engaged in the task, and will respond in a subjectively honest fashion.

Unfortunately, careless or inappropriate test administration is likely to result in a breach of this contract. The perception of insufficient rewards will lead certain individuals to become disengaged from the task. Further, if individuals sense that they are being evaluated on the basis of their responses, many will become primarily or even exclusively concerned with making a specific type of impression. In the present paper, several potentially invalidating test-taking attitudes are explored on the revised California Psychological Inventory (CPI; Gough, 1987).

Test-Taking Attitudes and Validity

When the motive to present an unreal, positive façade exists, a *fake-good* strategy may be employed. A fake-good strategy reflects a bias toward an "ought self" rather than toward an "ideal self" (Higgins, Klein, & Strauman, 1985); it is manifested by the failure to acknowledge commonly held weaknesses (Hartshorne & May, 1928) and the report of a constellation of unusual virtues (Gough, 1952). Conceptually, faking good is a form of other-deception rather than self-deception (Sackeim & Gur, 1978); this distinction is imperfectly captured by the difference between the Lie (L) and Correction (K) scales of the Minnesota Multiphasic Personality Inventory (Paulhus, 1984).

A second test-taking strategy, *faking bad*, may also be considered. The fake-bad strategy is the product of the motive to appear poorly adjusted, overtly neurotic, or even psychotic. Although usually infrequent, faking bad may appeal to respon-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 1, March 1989, pp. 45–56
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/010045-12$1.85

dents in certain testing situations. For example, a prison inmate or accused felon may wish to appear incapable in order to attain the relative sanctuary of a psychiatric hospital, and a military inductee might wish to appear incompetent in order to avoid service. From the perspective of the respondent, faking bad may represent a cry for help, an attempt to extract aid or sympathy, or simply a manifestation of negativism toward the "system" as represented by the convenient target of the personality test.

Deceptive (fake-good and fake-bad) strategies may be contrasted with an apparently random (or stimulus-avoidant) response pattern, in which responses appear disengaged from the content of test items. This pattern may arise for various reasons, including scoring errors, problems in understanding the items or instructions, or a lack of willingness or ability to respond to item content.

Other test-taking attitudes. This brief list of potentially invalidating test-taking strategies or attitudes is not intended to be exhaustive. Because the self-concept is complex and multifaceted (Markus & Wurf, 1987), a wide range of strategies may come to bear on the response process. For example, Dunnette, McCartney, Carlson, and Kirchner (1962) investigated the ability to simulate the responses of an "ideal salesperson." Dicken (1959, 1960) investigated the ability to simulate high scores on traits such as dominance and flexibility. In contrast with these relatively narrow categories, the motives to appear either virtuous or incapable and the potential for disengaged responding are all relatively broad, and thus have potential consequence in a wide range of testing situations.

Invalidity and the CPI

Three scales have been developed on the CPI to assess fake-good, fake-bad, and random response patterns. The Good Impression scale (Gi) includes items that empirically distinguish responses made under instructions to present a highly favorable self-portrait from responses made under standard conditions (Gough, 1952). The Well Being scale (Wb) was designed to identify fake-bad protocols; this scale includes items that distinguish persons simulating neurosis or pathology from normal and genuinely pathological samples (Gough, 1954).

The third CPI scale to assess invalidity, Communality (Cm), is comprised of items with endorsement rates that deviate severely from .50. A person who responds without concern for item content can be expected to answer approximately half of the items in the scored direction, thereby attaining a score approximately five standard deviations (SDs) below the mean obtained by normative samples (Gough, 1957, 1987).

Recently, a substantial revision of the CPI was published (Gough, 1987). In addition to other changes, each of the three validity scales has been altered. On the Gi scale, 5 of the 40 items were replaced by other items. On the Wb scale, 6 (14%) of the 44 Wb items were eliminated. On the Cm scale, 3 of the 28 old items were eliminated, and 13 new items were added.

Although correlations between the new and old versions of the three scales are high, ranging from .88 to .98, the effectiveness of the modified scales in identifying faked or disengaged responding remains worthy of investigation. Further, these scales offer only a "first line of analysis" in the detection of invalidity (Gough, 1987, p. 36); it is likely that other CPI scales may contribute to the detection of invalid protocols.

Method

An approach to detecting three types of invalid protocols on the revised CPI is presented. First, regression equations are developed to detect invalid protocols. Second, cutting scores for these equations are derived in a framework of signal detection theory (SDT; Swets, Tanner, & Birdsall, 1961). Finally, these cutting scores are applied to several archival datasets. Changes in predictive validity as a consequence of excluding protocols labeled "invalid" are then investigated.

Examinees

Five of the samples used in the present study were datasets from the CPI archives (total $N =$ 9,644). This archival data included individual item responses, and thus permitted rescoring using the keys for the revised CPI (Gough, 1987). In addition to these actual protocols, an additional eight samples, each consisting of 100 cases, were computer-generated to simulate various forms of "random" responding.

Samples consisting of faked protocols. In 1956, the CPI was administered to several samples of university undergraduates. Following Ruch (1942), each of these samples took the CPI twice, first under standard instructions, and then with special instructions. The 100 university undergraduates (50 male and 50 female) in the Fake-Good condition received instructions that included the following: "Try to give just as favorable an impression of yourself as you would if you were actually applying for an important position, or were trying to create a very favorable impression."

In the Fake-Bad condition, an additional 100 undergraduates (50 male and 50 female) were instructed as follows: "In taking the CPI this time please try to simulate the responses of a person suffering from a neurotic disturbance. Try to imagine how this person would feel and think, and then complete the CPI as you think this person would complete it."

Control samples. Two heterogeneous samples of ostensibly valid protocols served as control groups to develop and validate the regression equations. The Control-1 sample consisted of 2,200 protocols, including married couples, college and professional school students, high school students, and prisoners. The Control-2 sample included the 1,850 (of 2,000) CPI normative cases reported in Gough (1987) that were not included in the Control-1 sample. The fifth (high school) sample included 5,394 students, and was used to examine empirical implications of invalid responding. In addition to the CPI, non-test information—teacher ratings of several personality characteristics, grade-point average (GPA), and subsequent college attendance—was gathered for many of these students (Gough, 1964, 1968).

Random datasets. Three datasets (Random-1, Random-2, and Random-3) each consisted of 100 protocols produced by a pseudo-random number generator; in these datasets, the item endorsement rate (probability of responding "true") was set to .50. In five additional datasets of 100 simulated protocols each, the item endorsement rate was manipulated to assess the ability of the CPI equations to detect various forms of random responding.

Development of Regression Equations

A series of discriminant analyses was undertaken to predict each of the three types of invalidity from 19 of the 20 profiled scales of the revised CPI. (The remaining scale, Femininity/Masculinity, was excluded from these analyses because of the distinctly differing implications of high and low scores for the two genders.) Pairwise regression analyses (e.g., "valid" vs. "fake-good" protocols) were employed rather than a single multiple discriminant function analysis because certain types of classification errors were more costly than others. For example, calling a Fake-Bad protocol valid is a more costly error than calling that same protocol random.

In these analyses, it was recognized that trivial increments in predictability might attain statistical significance simply because of the large sample sizes. Because of this, a conservative strategy was embraced in which the analyses were limited to a maximum of six steps. In several cases, prediction reached an asymptote prior to the inclusion of six CPI scales, and shorter equations were therefore used.

Regression analysis provides optimal discriminant functions for detecting each of the potentially invalidating test-taking attitudes on the CPI. However, the multiple correlation statistic R is not satisfactory as an index of the accuracy (or effectiveness) of these equations, because the magnitude of R is largely determined by the relative sample size of the valid and invalid groups. Following a discussion of the development of the equations, a measure of accuracy that is immune to this problem is introduced.

Results

Prediction of Faking Good. Scores on the Gi scale correlated .40 with the dichotomous criterion of membership in the Control-1 versus Fake-Good samples. With additional scales, discrimination was improved; a six-variable equation resulted in a multiple correlation of .46. Weights for this and all subsequent equations apply to raw, as opposed to standard, CPI scores. Adjusted to provide a mean of 50 (SD of 2.63), the equation is

$$\text{Fake Good} = 44.67 + .15\text{Do} + .18\text{Em} + .35\text{Gi} - .11\text{Wb} - .13\text{To} - .12\text{Fx} \quad . \quad (1)$$

Because no other Fake-Good sample was available to directly estimate the degree of shrinkage on cross-validation, several supplementary analyses examined the robustness of the equation. First, scores from the two equations developed within each gender group separately proved quite similar (r = .92); multiple correlations for these equations showed minimal shrinkage (.01 to .02) on application to the opposite gender. Second, the total-sample equation was used to distinguish between the Fake-Good group and a new group of controls (the Control-2 sample); this comparison produced a semi-cross-validated correlation that equaled the correlation obtained in the derivation sample (.46), and was substantially higher than the correlation that resulted from the Gi scale alone (.39). Given the high ratio of observations to predictors (1,950:19), the precise b weights in the equation can be expected to outperform simple unit weights on cross-validation (Dawes, 1979; Goldberg, 1972).

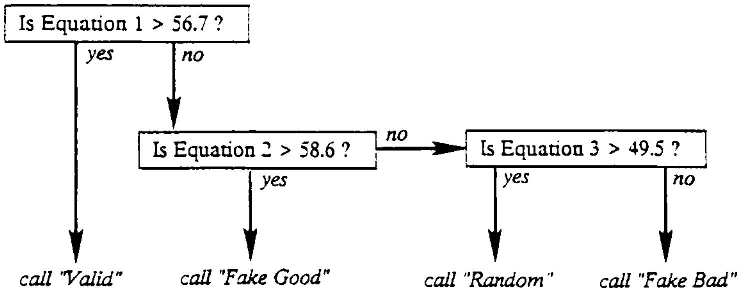
Prediction of Fake-Bad and Random response patterns. Initially, two separate equations were developed to predict Fake-Bad and Random response patterns. A five-variable equation to predict membership in the Fake-Bad sample obtained a multiple correlation of .74. A five-variable equation to predict membership in the Random-1 sample was also quite successful, resulting in a multiple correlation of .72. In both cases, the equations appeared to be robust: Scores on the within-gender equations correlated at least .97, and showed little shrinkage (no more than .02) on application to the other gender group. In short, both Fake-Bad and Random protocols could be reliably distinguished from samples of valid protocols by using these two equations.

However, a problem appeared in that the Fake-Bad and Random equations were very similar to each other: In the Control-1 sample, scores on these two equations correlated .93. Further, an attempt to apply reasonable cutting scores led to the less-than-ideal result that every case identified as "fake-bad" was also called "random." In other words, *when contrasted with valid protocols,* Fake-Bad and Random protocols appeared highly similar.

It remains possible, however, that Fake-Bad and Random protocols may be distinguished from each other in the absence of valid protocols. For this reason, a sequential or decision-tree strategy was investigated.

Untangling Random and Fake-Bad protocols. Figure 1 illustrates a sequential strategy for detecting faked and random protocols. First, protocols are evaluated with respect to the Fake-Good equation (Equation 1). Protocols not identified as "fake-good" are then evaluated with respect to a second equation (other-invalid, Equation 2) that classifies protocols as either "valid" or belonging to a combined "random and fake-bad" category. For protocols with sufficiently high scores on the

Figure 1
Decision Tree for Evaluating CPI Profile Validity



second equation, Equation 3 (random vs. fake-bad) is then applied. The development of the cutting scores shown in Figure 1 is discussed below.

The equation that determines whether a profile should be considered as belonging in a combined "random and fake-bad" classification was developed by contrasting the combined Random-1 and Fake-Bad samples with the Control-1 sample. The raw-score form of this three-scale equation (X = 50, SD = 2.51) is

$$\text{Other Invalid} = 75.77 - .68\text{Cm} - .18\text{Wb} + .12\text{Ac} \quad . \tag{2}$$

The multiple correlation for this three-scale equation (.80) was only slightly larger than the Pearson correlation obtained by the Cm scale alone (.79). The three-step solution was preferred because it surpassed the single Cm scale in distinguishing a validational sample of controls (Control-2) from a combined original Fake-Bad and additional Random group (Fake-Bad + Random-2), which had correlations of .83 and .82, respectively.

The third equation, which distinguishes between Random and Fake-Bad protocols, was developed by contrasting the combined Random-1 and Random-2 samples with the Fake-Bad sample. This five-scale equation (X = 50, SD = 2.2) also attained a multiple correlation of .80:

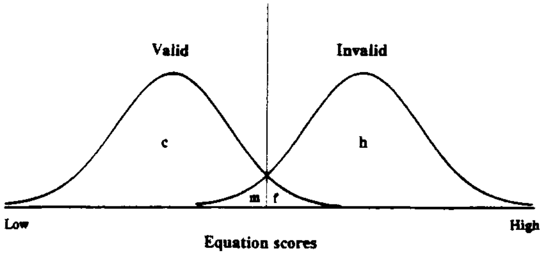
$$\text{Random vs. Fake Bad} = 41.95 + .13\text{In} + .22\text{Gi} - .06\text{Cm} + .14\text{Py} + .13\text{Fx} \quad . \tag{3}$$

Determination of Cutting Scores

As is true of a number of problems in personality measurement (Stenson, Kleinmuntz, & Scott, 1975; Szucko & Kleinmuntz, 1981), assessing protocol validity may be considered as a problem in accuracy analysis (Swets & Pickett, 1982) or as a diagnostic problem in signal detection (Swets et al., 1961).

Scores on Equations 1 and 2 may be obtained by either invalid (signal + noise) or valid (noise) protocols (see Figure 2). Because both kinds of protocols may be correctly or incorrectly identified, four outcomes are possible: An invalid protocol classified as invalid is a hit (h); a valid protocol classified as invalid is a false positive (f); an invalid protocol classified as valid is a miss (m); and a valid protocol classified as valid is a correct rejection (c). For Equation 3, the problem is fully analogous, differing only in that the problem is to distinguish between random and fake-bad protocols, rather than between invalid and valid protocols.

Figure 2
Assessment of Protocol Invalidity as a Signal Detection Task
(h = Proportion of Invalid Protocols Correctly Classified, c = Proportion of Valid Protocols Correctly Classified, $m = 1 - h, f = 1 - c$)



Working within this signal-detection framework, the determination of optimal cutting scores on each equation is a function of three parameters: (1) the accuracy of the assessment device, (2) the perceived proportion of persons belonging to the two groups (i.e., the prior probability that a given person will respond in an invalid fashion), and (3) the relative utility assigned to each of the set of possible outcomes, $h, f, m,$ and c. Of these parameters, only accuracy can be assessed empirically; estimates of likelihood and cost must of necessity be made on an a priori basis.

Accuracy of the selection device. Numerous indices that gauge the discriminability, effectiveness, or accuracy of selection devices are available (Swets, 1986a). Many of these, including simple and multiple point-biserial correlations, vary as a function of the relative size of the two distributions being evaluated (e.g., fake and valid protocols). Other indices, such as d', assume that the SDs of the two distributions being compared are equal. Of the measures that do not make these assumptions, A_z is in many respects the most attractive (Swets, 1986a; Swets & Pickett, 1982).

A_z may be defined as the proportion of the area of the graph that lies beneath and to the right of the parabolic form of the receiver operating characteristic (ROC), when h and f are each plotted on linear probability scales. Conceptually, A_z assesses the probability of a correct answer in a two-alternative forced-choice trial. If a person guesses (without information) in such a trial, then h and f are equally likely, the ROC curve is a straight line that bisects the graph along the negative diagonal, and $A_z = .5$. As discrimination approaches perfection, the probability of h relative to that of f increases, a greater proportion of the graph lies below the chance diagonal, and A_z approaches its limit of 1.

The computation of A_z is facilitated by plotting the estimates of h and f on binormal graphs, on which both axes are scaled so that corresponding normal deviates (upper and right-hand axes) are equally spaced (see Figure 3). ROCs plotted on such graphs are linear or nearly linear in form (Swets, 1986b; Swets & Pickett, 1982). A_z is a function of the slope of the linear ROC (s) and the distance between the means of the ''signal + noise'' and ''noise'' distributions (Δm), with the latter parameter expressed in SD units of the noise distribution.

In Figure 3, Δm is estimated by the absolute value of $z(F)$, as shown on the upper scale, at the intersection of the ROC curve with the horizontal $z(H)$ of 0. A_z is then taken as the proportion of the area under the normal curve corresponding to the value $[s(\Delta m)/(1 + s^2)^{1/2}]$ (Swets & Pickett, 1982).

For each of the three equations, two independent judges each plotted linear ROC curves from five pairs of values of h and f. Estimates of A_z computed on these curves differed by no more than .005; values of A_z were .91, .99, and .95 for Equations 1 through 3, respectively.

Comparison of these values with those found in other areas of investigation can be useful. In a recent review, Swets (1986b) described typical values of A_z found in problems of medical imaging (A_z between .85 and .95), information retrieval (.85 to .95), weather forecasting (.70 to .90), aptitude testing (.66 to .72), and polygraph performance (.80 to .95). Clearly, the present values of A_z (.91 to .99) are high: Valid and invalid protocols appear to be readily distinguishable on the CPI.

Figure 3 also provides values for the second ROC parameter, the slope (s), which describes the relative variation in the two samples being evaluated. For the first two equations, these values are less than 1, indicating a greater SD for scores among invalid cases than for protocols from the control group. The third equation shows a steep slope, indicating a substantially greater SD for equation scores in the Fake-Bad sample (SD = 1.92) than in the cross-validational Random-3 sample (SD = .89).

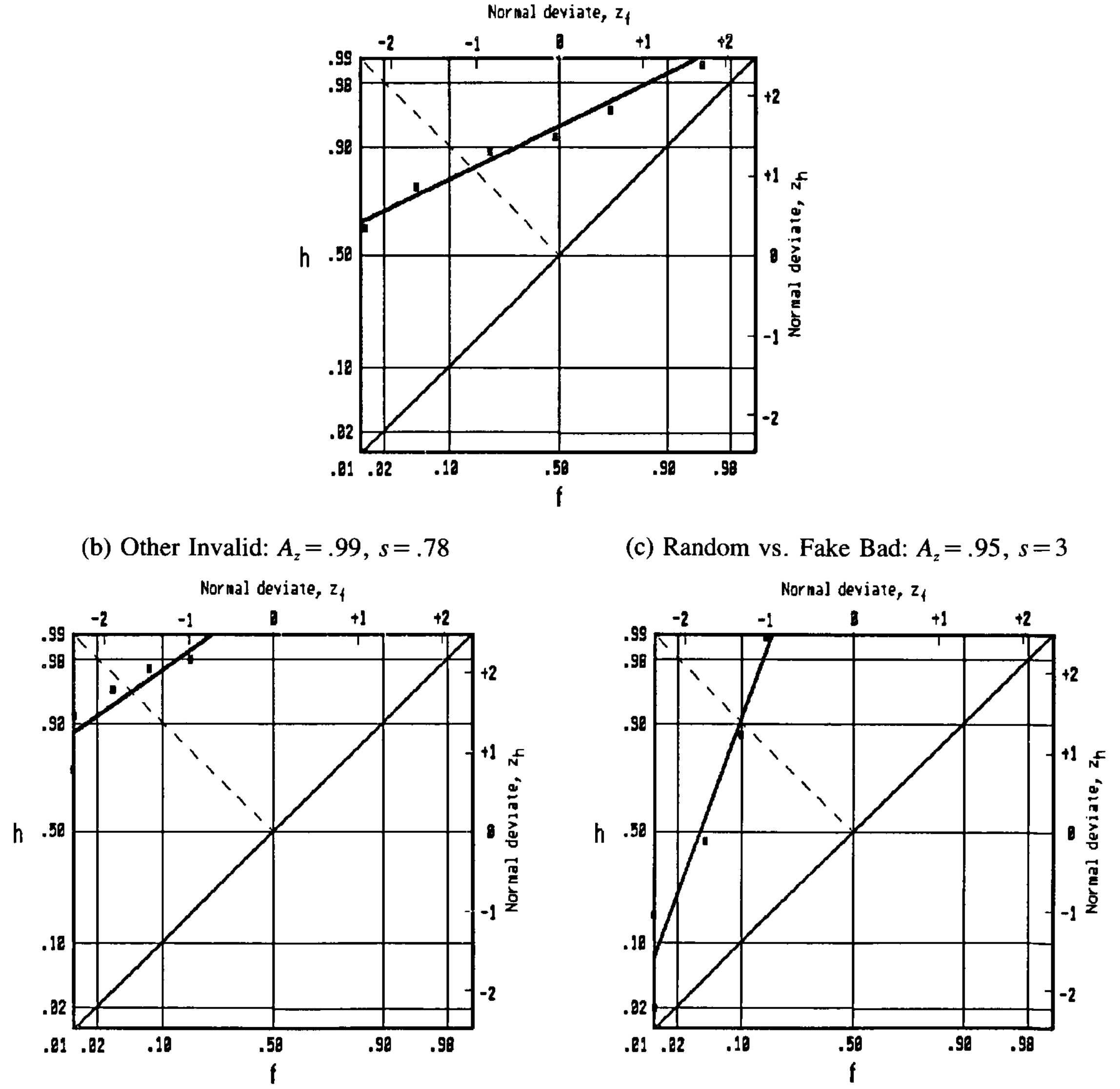
Incidence of invalidity. For any particular application, the determination of optimal cutting scores is not simply a function of the accuracy of the selection device, but also depends on the proportion of persons belonging to the two groups (Meehl & Rosen, 1954). That is, the overall frequency of errors depends on the relative incidence of valid and invalid protocols.

Unfortunately, the expected incidence of invalidity is difficult to estimate. Because faking good, faking bad, and disengaged responding are largely situation-dependent, no single estimate of prior probability will be fully satisfactory. Dunnette et al. (1962) estimated that in typical selection situations, no more than one in seven individuals fake good. For most purposes, this estimate is probably high. Particularly in research settings, the percentage of examinees whose primary motivation is to deceive will likely be less than this. Schwab (1971) has argued that the magnitude of faking has been overemphasized. In the great majority of testing situations, the incidence of faking may be presumed to be no greater than 20% and is much more likely to be in the vicinity of 1%.

In order to arrive at relatively robust cutting scores for each of the proposed equations, the total percent incorrectly classified may be ascertained for various prior estimates of invalidity, and over a range of possible cutting scores. For Equation 1, cutting scores in the range of 56 to 57 minimize error for estimates of invalidity between .01 and .20, leading to overall misclassification rates between 0% and 7% (Table 1). As a single cutting score, the value of 56.7 is advocated; this score is biased

Figure 3
Receiver Operating Characteristics (ROCs)
for the CPI Invalidity Equations

(a) Fake Good: $A_z = .91$, $s = .51$



(b) Other Invalid: $A_z = .99$, $s = .78$

(c) Random vs. Fake Bad: $A_z = .95$, $s = 3$

slightly above the midpoint of the 56 to 57 range, reflecting the fact that situations in which faking is infrequent are presumed to be more numerous than those in which faking is frequent.

For Equation 2, applying a wide range of potential cutting scores (56 to 61) results in error rates no greater than 5%; here, the cutting score of 58.6 is proposed. For both of these equations, where it

Table 1
Misclassification Rates for the Fake-Good Equation (Equation 1) and Other-Invalid Equation (Equation 2)

Cutting	Error Type		Invalidity Incidence			
Score	*f*	*m*	.20	.10	.05	.01
Fake-Good						
57	.00	.42	8%	4%	2%	0%
56	.01	.33	7%	4%	3%	1%
55	.04	.26	8%	6%	5%	4%
54	.07	.23	10%	9%	8%	7%
Other-Invalid						
60	.01	.21	5%	3%	2%	1%
59	.01	.14	4%	2%	2%	1%
58	.01	.09	3%	2%	1%	1%
57	.02	.06	3%	2%	2%	2%
56	.03	.05	3%	3%	3%	3%

Note. Percentages refer to overall misclassification rates, assuming that values equal to or greater than cutting scores are identified as "invalid."

is anticipated that invalidity rates will be outside of the 1% to 20% range, or where a precise estimate of invalidity exists, suitably adjusted cutting scores should be employed to minimize misclassifications.

For Equation 3, the relative prevalence of these two types of invalid protocols should be estimated. If these differ by a ratio of no more than 3:1 (or 1:3), cutting scores in the vicinity of 49 to 50 appear optimal. Here, the single value of 49.5 is proposed. Although low in an absolute sense, the minimum overall error rates in the present case (4% to 10%) are slightly higher than those found for prior equations (see Table 2). These error rates are higher, in part, because the moderate prior probabilities (.25 to .75) used in distinguishing random from fake-bad protocols are less informative than the relatively extreme values (.01 to .20) used for Equations 1 and 2.

Utilities. The expected utility of a given cutting score is the sum, over the four possible outcomes *h, f, m,* and *c*, of the likelihood of the outcome multiplied by its value or cost. Because estimates of utility are subjective, it is appropriate to examine the robustness of particular cutting scores for several sets of perceived values and costs.

Two major settings in which the CPI is applied are counseling and selection. The counseling setting may be characterized as one in which each *f* is twice as costly as each *m*. By labeling a valid protocol as "invalid," the counselor not only loses the possibility of a substantive interpretation, but may also disappoint or alienate the client. Conversely, the selection setting, from the standpoint of the administrator, may be characterized as having each *m* twice as costly as each *f*.

For the present purposes, the values ascribed to correct decisions (*h* and *c*) may be taken as null for both counseling and selection situations. The expected utility of a particular cutting score is then the product of the frequency of the two types of error multiplied by their respective costs. These relative costs (2:1, 1:2) may be combined with the prior probabilities of Tables 1 and 2 to assess the expected utility of each of the proposed cutting scores (56.7 for Equation 1, 58.6 for Equation 2, and 49.5 for Equation 3).

When these utilities are applied to the rates of *f* and *m* in Tables 1 and 2, the cutting scores proposed for Equations 1 and 2 appear quite robust. That is, for both equations the cutting scores appear optimal or nearly optimal, regardless of the relative cost of *f* and *m*, in all situations where faking is

Table 2
Misclassification Rates for the Random Versus Fake-Bad Equation (Equation 3), Assuming Only Random and Fake-Bad Protocols

Cutting	Error Type		Random Incidence		
Score	*f*	*m*	.75	.50	.25
52	.01	.84	63%	43%	22%
51	.04	.52	40%	28%	16%
50	.09	.11	11%	10%	10%
49	.14	.01	4%	8%	11%

Note. Percentages refer to total percent incorrectly classified, assuming that values equal to or greater than cutting scores are identified as "random."

perceived as unlikely (prior probability of 1%). Only in the situation where faking is perceived as likely (prior of 20%) and each *m* is perceived as costly (selection setting) are these cutting scores suboptimal. In these circumstances, the optimal cutting scores for each equation drop by 2 points, leading more protocols to be identified as "invalid." This results in a reduction of expected cost-weighted error rates of approximately 2%.

For Equation 3, weighted error rates may also be examined, for the case where each *m* is seen as twice as costly as each *f* and for its inverse, where each *f* is seen as twice as costly as each *m*. In both cases, over the range of prior probabilities specified in Table 2, optimal cutting scores remain between 49 and 50. As with the first two equations, the cutting score proposed above appears satisfactory.

The proposed cutting scores are not intended to be universal. For some users, the perception of relative cost will be outside of the range discussed here. In addition, some users may prefer a decision rule that conflicts with maximizing expected utility (see, e.g., Lopes, 1981). In these circumstances, appropriate cutting scores will differ from the values proposed here.

Application of Cutting Scores to Additional Random Datasets

The equations to assess random responding were developed and validated on samples in which the probability of item endorsement was set to .50. (That is, the expected number of "true" responses to the 462 CPI items was 231.) In this analysis, five additional sets of 100 protocols were generated in which item endorsement rates were set at .01, .20, .50, .80, and .99. These additional datasets permitted an assessment of the effectiveness of the proposed algorithms in detecting various forms of random responding.

In the Random $P = .01$ sample, all 100 cases were correctly identified as random. In the Random $P = .20$ sample, 82 cases were correctly identified as random, with the remainder classified as valid. In the Random $P = .50$ sample, 87 cases were identified as random, 12 as valid, and 1 as fake-bad. Thus, for low to moderate endorsement rates, the proposed equations and cutting scores worked well, correctly identifying between 82 and 100 percent of the random cases.

For high endorsement rates, however, errors were frequently made, in which protocols were identified as fake-bad rather than as random. In the Random $P = .80$ sample, 92 cases were identified as fake-bad, with the remainder correctly identified as random. In the Random $P = .99$ sample, all 100 cases were identified as fake-bad. No case in either of these samples was identified as valid. This suggests that a random CPI protocol in which the "true" response is very frequent is difficult to distinguish from a fake-bad protocol. However, the unusual profile that results from this strategy will not be classified as valid.

Empirical Implications of Invalidity

Do individuals who respond in an invalid fashion differ from those who respond "honestly" on a personality questionnaire? To the extent that these individuals differ, the protocols that emerge from faked and disengaged strategies are not altogether invalid, for the protocols carry potentially diagnostic information.

A sample of 5,394 CPI protocols from high school students was available to investigate the non-test correlates of protocol validity. In this sample, the apparent incidence of both faking good (.2%, $n = 13$) and faking bad (.4%, $n = 22$) was slight. Cases identified as random were more frequent (2.2%, $n = 121$). The relatively high incidence of disengaged responding is to be expected, as the test was primarily administered in group settings in which there was relatively little incentive to students to respond frankly to item content.

Information in "random" protocols. Although the number of cases identified as fake-good and fake-bad was too small to permit reliable inferences, the number of cases identified as random was large enough to permit a comparison with the group of protocols that appear valid. Analyses of variance examined the relationships of profile status (valid vs. random) and gender with three non-test variables: (1) GPA ($n = 4,149$), (2) whether the student subsequently enrolled in college ($n =$

2,620), and (3) a composite teacher rating of socialization (best citizen rating minus disciplinary problem rating; $n = 4,073$).

For each of these analyses, both the main effect of profile status and the interaction of profile status with gender were statistically significant. A main effect of gender was present for the analyses of GPA and college attendance; no main effect of gender could emerge for the socialization rating because this variable was a composite of teacher nominations, each of which had been equally distributed in the two gender groups. (For GPA, F ratios for profile status, gender, and the profile status × gender interaction were 9.67, 120.15, and 22.52, respectively, each with $df = 1$, 4145. For college enrollment, these values were 6.91, 71.43, and 5.35, each with $df = 1$, 2616. For ratings of socialization, F ratios were 5.57, .09, and 5.83, each with $df = 1$, 4069.)

The cell means revealed no significant or substantial differences between female students with protocols identified as valid and female students with protocols identified as random. The means of the two groups were separated by only .01 SD for GPA (Ns of 2,175 and 79 for valid and random protocols, respectively), .17 SD for continuation to college (Ns of 1,354 and 75), and .11 SD on rated socialization (Ns of 2,136 and 80).

The protocols of male high school students were less likely to be identified as random than those of female high school students. Within the male sample, however, random CPI responding was accompanied by depressed scores on each of the dependent variables. The average GPA of 31 male students with random CPIs was 1.05 SDs below that of 1,864 students with valid CPIs. Only 2 of 17 males with CPIs identified as random went on to college (12%); this contrasts with the 50% continuation rate for 1,174 males with CPIs identified as valid. Similarly, the teacher ratings of socialization were .61 SDs higher for the valid males than for the random males. Correlations between the dichotomous criterion of profile validity and these dependent variables were misleadingly low (.13, .09, and .07, respectively), as these statistics were severely limited by differences in the shapes of the distributions being examined.

Invalidity and CPI research. What happens to the correlations between CPI scales and non-test measures when the cutting scores are first employed to screen out invalid protocols? Consistent with the prior analysis, correlations between CPI scales and GPA, college enrollment, and ratings of socialization marginally increased for females, and marginally decreased for males. For females, the average of the absolute values of 69 correlations between 23 CPI scales (the 20 profiled scales and the three structural scales) and target measures increased (by .009) when the invalid cases were dropped. For males, the average correlation decreased (by .005) when these cases were dropped.

Although the magnitudes of these changes in correlation are small, their direction is consistent. For females, 60 of the 64 correlations that changed showed an increase ($P < .001$, sign test); for males, 46 of the 66 correlations that changed showed a decrease ($P < .001$). For females, random protocols are truly invalid, and their removal results in greater predictability of target measures. For males, the decrease in correlations is, in part, attributable to the relation previously found between profile status and these target measures. Because of this relation, removing the invalid protocols resulted in a reduction in the variance of the target measures as well.

Discussion

The concern with protocol validity is not new (Gough, 1947), and the use of signal detection parameters is best understood as a simple formalization of techniques that have long been used by interpreters of diagnostic tests (Cureton, 1957; Gough, 1950). Recently, other techniques have emerged to detect invalid protocols. For example, Schmolck (1987) has argued that random (stimulus-avoidant) protocols may be identified by the absence of a relationship between the individual's array of responses and any substantive dimensions underlying the inventory. The availability of alternative methods notwithstanding, a substantial degree of empirical success has been obtained in the present study, with accuracy coefficients (A_z) between .91 and .99.

Across personality questionnaires, invalidity can be expected to occur more frequently when the set of items is jarring, offensive, embarrassing, or egodystonic than when items are inoffensive and egosyntonic (Gough, 1987; Schmolck, 1987). For a given personality questionnaire, the likelihood of an invalid pattern of responses is a function of the circumstances of test administration, the personality of the respondent, and the interaction of these variables.

For the CPI, the algorithms proposed in this paper have been used to estimate the prevalence of invalid responding in each of 57 samples (Gough, 1987). (It should be noted that using Equation 3, Gough reported a cutting score of 50, rather than the score of 49.5 used here.) Estimates of the incidence of faking good range from 0% (numerous samples, the majority of which took the test as part of in-depth assessments at the Institute of Personality Assessment and Research) to 10.7% (9 of 84 male police applicants). The protocols of psychiatric patients of both genders were most likely to be identified as fake-bad (4 of 34 females, or 11.8%) and random (3 of 41 males, or 7.3%). Among larger samples, male prison inmates were most frequently identified as fake-bad (4 of 196, or 2.0%), and the protocols of male military academy students were the most likely to appear random (92 of 1,414, or 6.5%).

Across these samples, the prevalence of invalidity appears to reflect the circumstances of test administration (e.g., police applicants) as well as sample characteristics (prison inmates and psychiatric patients). Within each sample, however, people differ in the extent to which the demand characteristics of the test situation are salient, and differ in their tendency to embrace an invalid strategy. Because these differences are lawful, faked protocols may carry diagnostic significance, and substantive relationships may exist between test-taking attitudes and non-test measures (McCrae & Costa, 1983).

For these reasons, relations between protocol validity and several non-test measures were explored, in the present study, in a large sample of high school students. For females, random answering was found to be unrelated to the likelihood of college attendance, GPA, and socialization. For males, however, responding in a disengaged fashion indicated a lower probability of going on to college, a lower GPA, and a greater likelihood of being perceived as delinquent. Disengaged responding carries information about these respondents and is, in this important sense, not completely random. This may be taken as a demonstration that even the most extreme and non-disclosing of test-taking attitudes may carry empirical implications, and testifies to the soundness of an empirical approach to interpreting personality questionnaire data (Meehl, 1945).

It should be recognized that invalid responding is an infrequent event. Problems in detecting or diagnosing infrequent events are notoriously difficult (Meehl & Rosen, 1954). Despite this hurdle, the proposed algorithms appear useful. In most circumstances, these will outperform simple base rates as low as 1% when appropriate cutting scores are applied.

References

Cureton, E. E. (1957). Recipe for a cookbook. *Psychological Bulletin, 54,* 494–497.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dicken, C. F. (1959). Simulated patterns on the Edwards Personal Preference Schedule. *Journal of Applied Psychology, 43,* 372–378.

Dicken, C. F. (1960). Simulated patterns on the California Psychological Inventory. *Journal of Counseling Psychology, 7,* 24–31.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15,* 13–24.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs, 7*(2).

Gough, H. G. (1947). Simulated patterns on the Minnesota Multiphasic Personality Inventory. *Journal of Abnormal and Social Psychology, 42,* 215–225.

Gough, H. G. (1950). The *F* minus *K* Dissimulation Index for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology, 14,* 408–413.

Gough, H. G. (1952). On making a good impression. *Journal of Educational Research, 46,* 33–42.

Gough, H. G. (1954). Some common misconceptions about neuroticism. *Journal of Consulting Psychology, 18,* 287–292.

Gough, H. G. (1957). *Manual for the California Psychological Inventory.* Palo Alto CA: Consulting Psychologists Press.

Gough, H. G. (1964). Academic achievement in high school as predicted from the California Psychological Inventory. *Journal of Educational Psychology, 55,* 174–180.

Gough, H. G. (1968). College attendance among high-aptitude students as predicted from the California Psychological Inventory. *Journal of Counseling Psychology, 15,* 269–278.

Gough, H. G. (1987). *California Psychological Inventory administrator's guide.* Palo Alto CA: Consulting Psychologists Press.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character. Volume 1: Studies in deceit.* New York: Macmillan.

Higgins, E. T., Klein, R., & Strauman, T. (1985). Self-concept discrepancy theory: A psychological model for distinguishing among different aspects of depression and anxiety. *Social Cognition, 3,* 51–76.

Lopes, L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 377–385.

Markus, H., & Wurf, E. (1987). The dynamic self concept: A social psychological perspective. *Annual Review of Psychology, 38,* 299–338.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882–888.

Meehl, P. E. (1945). The dynamics of "structured" personality tests. *Journal of Clinical Psychology, 1,* 296–303.

Meehl, P. E., & Rosen, A. (1954). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

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

Ruch, F. L. (1942). A technique for detecting attempts to fake performance on the self-inventory type of personality test. In Q. McNemår & M. A. Merrill (Eds.), *Studies in personality* (pp. 229–234). New York: McGraw-Hill.

Sackeim, H. A., & Gur, R. C. (1978). Self-deception, self-confrontation, and self-consciousness. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory, Volume 2* (pp. 139–197). New York: Plenum.

Schmolck, P. (1987). *How to verify that an MMPI test protocol reflects stimulus-induced item responses instead of stimulus avoiding response sheet marking.* Unpublished manuscript, Department of Education, University of the Federal Armed Forces, Neubiberg, West Germany.

Schwab, D. P. (1971). Issues in response distortion studies of personality inventories: A critique and replicated study. *Personnel Psychology, 24,* 637–647.

Stenson, H., Kleinmuntz, B., & Scott, B. (1975). Personality assessment as a signal detection task. *Journal of Consulting and Clinical Psychology, 43,* 794–799.

Swets, J. A. (1986a). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Swets, J. A. (1986b). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340.

Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36,* 488–496.

Acknowledgments

This article has greatly benefitted from the comments of Lew Goldberg, Harrison Gough, Del Paulhus, and Peter Schmolck.

Author's Address

Send requests for reprints or further information to Kevin Lanning, Department of Psychology, Oregon State University, Corvallis OR 97331, U.S.A.