Multivariate Behavioral Research, *39* (2), 273-302

The Impact of Item Characteristics on Item and Scale Validity

John A. Johnson
Pennsylvania State University

This study describes the relation between personality items' *validities*, defined as the items' correlations with acquaintance ratings on the Big 5 personality factors, and other itemmetric properties including ambiguity, syntactic complexity, social desirability, content, and trait indicativity. Five external validity coefficients for each item on the California Psychological Inventory were correlated with a number of itemmetric variables often assumed to affect item validity. Item validity correlated positively with social desirability and trait indicativity and negatively with ambiguity across the five factors. Other characteristics had a more limited influence on item validity. Multiple regression analyses revealed trait indicativity — how obviously an item response indicates a trait — to be the most important determinant of item validity. Scales built from itemmetrically sound versus poor items showed differential validity in two additional samples. Implications for the psychological processes underlying responses to personality items are discussed.

Goldberg (1993) and Wiggins (1979) are probably best known for their influential work on the Big Five and circumplex taxonomies of personality traits. Unbeknownst to many, however, almost 40 years ago Wiggins and Goldberg (1965) were defining a new field called *itemmetrics* — the measurement of the properties of individual personality questionnaire items. Much of the early itemmetric research was limited to measuring and reporting

Much of this article was written while the author was a visiting research fellow at the University of Bielefeld, supported by a fellowship from the Alexander von Humboldt-Stiftung. I wish to express my gratitude to Alois Angleitner and Robert Wicklund for their help in arranging my research stay, and to the psychology department at Bielefeld and the von Humboldt-Stiftung for their support during that time.

I would also like to thank Michael Gerg and Kelly Horner for serving as judges on the item rating tasks in this study. Thanks go to Robert R. McCrae for providing the rational ratings of CPI items according to the five-factor model and to thanks go to the editor of the German CPI, Ansfried Weiner, for generously providing the authorized 462-item revised version of this inventory for the research project before the inventory actually went to press. Finally, I would like to thank Jack Block, Matthias Burisch, Lewis R. Goldberg, Harrison Gough, Robert Hogan, Ronald Holden, Eric Knowles, Del Paulhus, and Richard E. Zinbarg for their helpful suggestions on earlier version of the manuscript for this article.

Correspondence regarding this article can be sent to the author at Penn State DuBois, College Place, DuBois, PA 15801, or e-mailed to j5j@psu.edu.

single item properties such as ambiguity, social desirability, grammatical form, and retest consistency for all items on a given personality inventory (e.g., Angleitner, 1981; Goldberg & Rorer, 1964; Keren, 1979; Löhr, 1977). The early reports often read like accountants' ledger books, enumerating item properties without explaining the significance of these properties.

Other itemmetric research has examined the relations among test item parameters such as the relation between grammatical characteristics and item endorsement frequency (Wiggins & Goldberg, 1965). The "ultimate goal of itemmetric research," according to Goldberg (1968), "should be the discovery of the relationships between item properties and scale validity" (p. 273). It is curious, then, that one item characteristic that has been rarely studied is an item's *external validity* — its correlation with relevant nonself-report criteria such as acquaintances' personality ratings or work performance. The present article is a step toward Goldberg's ultimate goal by reporting the relations between external validity and an array of item characteristics.

The Speech Act Framework for Studying Item Properties

The present research was designed to accomplish more than simply catalog empirical relations between item properties and item/scale validity. This itemmetric research also used *speech act theory* (Austin, 1962; Searle, 1969) to make predictions about which item properties should be more relevant to validity than others. Jerry Wiggins (1974/1997) was the first personality psychologist to recognize the relevance of speech act theory for personality assessment, and his writings on this topic helped inspire the present research as well as my previous work on this topic (Johnson, 1997, 2002).

A central premise of speech act theory is that words can serve two fundamentally different purposes. One is propositional communication: to describe or depict the way things are. Austin (1962) coined the term *constative* to label purely propositional communications. Speech act theorists suggest that people rarely use constatives, and that purely propositional communication may exist only as an ideal in scientific discourse (van Oort, 1997). More often, people use language in a *performative* fashion. Performatives are acts designed to modify the present situation, including the behavior and state of mind of those to whom the words are directed (Allan, 1994).

Personality psychologists have debated whether endorsements of personality items such as "I read at least ten books a year" are constatives — mere description — or performatives — performances whose function is to create personality impressions on others. The constative and performative views of item responses make difference predictions about the item

properties most relevant to item validity. Empirical findings on the relations between item properties and validity will shed light on whether item responses are better viewed as constatives or performatives and should therefore help us to design better items.

The Constative View of Personality Item Responses

Most personality psychologists have treated personality item responses as if they are (or should be) constatives rather than performatives (Johnson, 1981, 1997). The constative view assumes that valid responses to personality items are obtained when the responses most accurately describe a person's actual thoughts, feelings, or behaviors.[1] For example, to produce a valid response to the item "I read at least ten books a year" the respondent must first scan his or her memory for the number of books he/she reads each year and then endorse the item if and only if the memory scan reveals that ten or more books are read each year.

The constative view predicts that item properties that interfere with objectively identifying and reporting one's actual thoughts, feelings, and behaviors will detract from the validity of the item. Many personality psychologists regard an item's *social desirability* as one such threat. Their concern has been that test-takers will tend to endorse items of high social

[1] The constative assumption that item responses must correspond to actual thoughts, feelings, or behavior to be valid is separate from the issue of interpreting the meaning of a response. If a person endorses the item "I enjoy hot, spicy foods," (and this accurately describes the person's food preferences), what does this imply about personality? Wiggins (1973) describes two approaches to item response interpretation, both of which assume a constative view. A *pretheoretical* approach (Wiggins, 1973, p. 401, Footnote 2) employs an intuitive, common-sense to item response interpretation. A pretheoretical interpretation of "I enjoy hot, spicy foods," relies on the folk wisdom of such preferences (perhaps that the person is bold). In contrast, a construct approach (Loevinger, 1957; Wiggins, 1973, p. 401) uses an explicit, formal psychological theory to interpret item responses. For example, a theory of extraversion that includes postulates about how the brains of extraverts and introverts leads them to different food preferences might interpret a preference for hot, spicy foods as a sign of extraversion. Nonetheless, in either case (pretheoretical or construct), the test interpreter needs the item response to correspond to the person's actual feelings about spicy foods in order to interpret that response.

The constative view also maintains that it is up to the test scorer and interpreter to make appropriate inferences about personality from the (hopefully) veridical reports, just as a medical doctor diagnoses diseases from patients' simple self-reports of symptoms (cf. Loevinger's 1957, p. 644) discussion of item responses as *signs* and *samples* of behavior). Constatively oriented psychologists would no more expect a person to answer items in order to create a personality impression than a doctor expects a patient to self-diagnose his or her medical condition. Psychologists operating under the constative view have assumed that responding in order to create a certain impression (i.e., performative responding) constitutes a "response style" that would interfere with valid assessment.

desirability and not endorse items of low social desirability *regardless of whether the item describes their actual behavior* (Edwards, 1990).

Another item property that has concerned constatively oriented psychologists is *item ambiguity*. Item ambiguity is uncertainty about the locution (literal significance) of the item. The constative view suggests that ambiguous items are problematic because they provide vague targets for comparing one's actual thoughts, feelings, or behaviors. One source of item ambiguity is the presence of indeterminate words such as *many* and *sometimes* (Simpson, 1944). Therefore, items such as "I read many books" would be predicted to be less valid than "I read at least ten books a year."

In addition to the presence of indeterminate words, other specific linguistic properties of items have been associated with item ambiguity. An item may also be ambiguous because it contains one or more of the following characteristics: (a) it possesses more than one deep structure ("Flattering people can be dangerous"), (b) it contains words with multiple meanings (*passionate* can refer to devotion, emotional volatility, or romantic inclinations), (c) it expresses multiple ideas through conjunctions and disjunctions, or (d) it contains complex constructions such as double negatives (Angleitner, John, & Löhr, 1986; Löhr, 1977; Löhr & Schütte, 1980).

Finally, constative theorists have suggested that item responses are more likely to correspond to someone's actual personality if the item refers to objective, verifiable facts. In the industrial-organizational literature (e.g., Becker & Colquitt, 1992), biographical items (e.g., "My high school grade point average was higher than 3.00") have been highly touted because such items are more verifiable than items that require subjective interpretation (e.g., "I did well in school"). Goldberg (1963) reported that people respond more consistently to items that reflect observable behavior than items reflecting attitudes, values, or other internal states of mind. Funder and Colvin's (1997) literature review indicates that highly visible (and therefore verifiable) traits show greater self-acquaintance and agreement and agreement among acquaintances. The constative view predicts that items that refer to verifiable behaviors will be more valid than items referring to subjective states of mind. Among the types of item content examined in this study, items referring to specific behaviors, behavioral traits, and biographical facts are more verifiable than items mentioning subjective characteristics (psychological reactions or traits, physiological reactions, and attitudes/opinions).

The Performative View of Personality Item Responses

Some dissenters to the constative view of personality item response have been those who emphasize the empirical correlates of responses over the

correspondence between item content and actual behavior (Buchwald, 1961; Meehl, 1945), as well as a few researchers who assume that personality item responses are in fact performatives (Hogan, 1991, 1996; Johnson, 1981, 1997). Although empirically oriented test authors (e.g., Gough, 1991, 1996) consider item properties when they *write* potential items, a purely empirical view is agnostic on the constative-performative issue and makes no predictions on item properties related to validity. The empirical view therefore will not be discussed further here.

The performative view considers the act of endorsing a personality item to be formally identical to the act of uttering that statement in the presence of an audience (Hogan, Carpenter, Briggs, & Hansson, 1985). Hogan (1991) argues that in both cases these speech acts are performative. Item responses "reflect automatic and often nonconscious efforts on the part of test-takers to negotiate an identity with an anonymous interviewer (the test author)" (Hogan, 1991, p. 902). While not completely discounting the possibility of conscious impression management, Hogan's view stresses the unwitting nature of performative communication. Just as speakers are normally unaware of the grammar that guides the production of sentences, actors are unaware of the *constitutive rules* (Searle, 1969) through which words produce personality impressions. The largely automatic and unconscious nature of rule-guided behavior accounts for the consistency of expression that we call personality.

In his classic article on personality traits, Wiggins (1974/1997) noted that the production of personality impressions from constitutive rules applies to acts generally, not just speech acts. Constitutive rules take the form "*X* counts as *Y* in context *C*." Wiggins (1974/1997, p. 101) gave the following example of a constitutive rule for a non-speech act: "An action (pushing) that is likely to harm or injure another (*X*) counts as aggressive (*Y*) in the context of rules for classifying the consequences of social actions (*C*)." People, in turn, are described as "aggressive" if, under certain circumstances, they have behaved in a manner likely to harm or injure another.

The constitutive rules for deriving personality impressions from personality item responses are embodied in the key for the personality inventory. To produce a validly interpreted item response, the respondent must respond in accordance with the constitutive rules — regardless of the literal applicability of the item to his or her actual thoughts, feelings, or behavior. For example, for scores on an Intellect scale to correlate with knowledgeable acquaintances' impressions of intellectuality, persons regarded as intellectual must endorse many items such as "I read at least ten books a year," even if they read only two or three books per year.

Because, according to the performative view of item responses, "the literal truth or falsity of responses is irrelevant" (Hogan et al., 1985, p. 29), item

characteristics presumed by the constative view to enhance or detract from the correspondence between item responses and actual behavior (social desirability, ambiguity, verifiability) may not have much impact on item validity.

In the case of social desirability, Emler (1999) noted that individuals who consistently perform socially undesirable acts in everyday life also consistently describe themselves in socially undesirable ways on personality questionnaires. Thus, the social desirability/undesirability of personality items, rather than detracting from validity, allows individuals to make the same positive or negative impressions on personality tests as they do in everyday life.

From a performative perspective, ambiguous items may sometimes be more valid than unambiguous items, because the manner in which a person interprets an ambiguous item is itself a valid indicator of personality (Elias, 1951). For example, neurotic and nonneurotic people who experience exactly the same number of headaches will interpret and endorse the ambiguous item "I often get headaches" differently. Neurotics' greater tendency to endorse such an item, rather than literally reflecting a higher frequency of headaches, validly reveals their tendency toward neurotic complaint (see Costa & McCrae, 1987; Schroeder & Costa, 1984).

Finally, objectively verifiable items are not necessarily better vehicles for creating a personality impression on an audience. In fact, Funder (2001) points out that the most objective, verifiable facts about a person (what Cattell called *L*- or life data) are the most uninformative about personality because these data are influenced by too many factors unrelated to one's personality. For example, "I have never been fired from my job" may indicate something about an individual's work ethic or reliability, but it may say as much or more about the person's employers or the local economy.

According to the performative view, the item characteristic that should most impact item validity is its ability to generate a consistent personality impression across most members of an audience. An item that indicates the same trait to most people might be said to have strong "*trait-indicativity*" (McCrae, Costa, & Piedmont, 1993). If saying "I read at least ten books a year" makes most people tend to see the speaker as *intellectual*, then the personality item "I read at least ten books a year" would have strong indicativity for the trait, *intellectual.*

Summary of Predictions

Social Desirability and Validity. It is a well-established fact that the proportion of persons in a sample who endorse a personality item correlates strongly (.8-.9) with the item's mean rated social desirability (Edwards, 1957). Constative theorists suspect that this endorsement pattern represents

either self-deception about, or deliberate misrepresentation of, one's actual thoughts, feelings, and behaviors (Borkenau & Ostendorf, 1989). The constative view therefore predicts the lowest validity for highly undesirable or desirable items and highest validity for items with relatively neutral desirability. In contrast, the performative view holds that the desirability of one's item responses simply mirrors the desirability of one's behavior in everyday life (Hogan & Hogan, 1992; Nicholson & Hogan, 1990). Higher endorsement rates for desirable items merely reflects the higher frequency of prosocial behaviors over antisocial behaviors in our society. The performative view sees no reason why items of neutral desirability would be more valid than items of low or high desirability.

Ambiguity and Linguistic Characteristics. The constative view suggests that ambiguous items make poor targets for comparing one's actual thoughts, feelings, and behaviors. Therefore, perceived ambiguity should be inversely related to item validity, as should linguistic characteristics likely to be associated with ambiguity (item length, negations, and indeterminate frequency qualifiers). Unlike the constative view, the performative view does not see accurate comparison between item content and actual behavior as crucial. The performative view regards ambiguity as a problem only to the extent that it interferes with trait indicativity. In fact, this perspective suggests that in some cases the manner in which individuals interpret ambiguous items can communicate valid information about themselves.

Content Type. Items have been categorized according to whether they are general or specific and whether they refer to emotional reactions, behaviors, physiological reactions, thought patterns, or attitudes (Angleitner, et al., 1986; Blaney, 1991; Werner & Pervin, 1986). The constative view suggests that items with specific, publicly observable referents should be more valid than items that refer to unobservable mental states and traits. The performative view, in contrast, predicts that item validity is much more a function of trait-indicativity than of any particular item content.

Trait Indicativity. From the constative viewpoint, trait indicativity is not a necessary characteristic for items to show validity. In fact, some constatively oriented psychologists have worried that if respondents become aware of what is being measured, they might respond in ways that they wish to appear rather than reporting their actual behaviors. In contrast, the performative position assumes that the ways in which people unconsciously strive to appear on personality inventories mirrors the way they strive to appear in everyday life. The performative position therefore holds that

strong trait indicativity helps respondents to create the same personality impressions on personality inventories that they create in everyday life. The performative view therefore predicts that trait indicativity will be the most important determinant of item and scale validity.

Serial Position. Knowles' (1988) research has shown that experience with an inventory improves the respondent's understanding of the personality construct that is being measured by the items, leading to higher item-total consistency for items toward the end of an inventory. From the performative perspective, a better grasp of the relation between the measured construct and items appearing later within an inventory implies a positive correlation between serial position and item validity. The constative view predicts no relation between serial position and validity.

Method

Participants

Three participant samples were used. Sample 1 consisted of 79 (31 male, 48 female) American students from an introductory psychology course. Sample 1 was used to derive estimates of the effects of itemmetric characteristics on item validity. Sample 2 consisted of 75 (30 male, 45 female) American introductory psychology students, and Sample 3 consisted of 96 (31 male, 58 female, 7 sex not indicated) psychology students recruited by advertisement at the University of Bielefeld in Germany. Samples 2 and 3 were used to determine whether itemmetric characteristics related to validity at the item level in Sample 1 similarly affected validity at the scale level. All subjects received course credit for participating in the study.

Personality Measures

California Psychological Inventory (*CPI*; Gough, 1975, 1987). This True-False inventory was chosen because its items are more itemmetrically heterogeneous than the items of other omnibus inventories. Many of its items were chosen empirically, allowing for a greater degree of subtlety than what is found in rationally constructed inventories. The length and complexity of its items vary greatly, from items such as "I like poetry" to "The person who provides temptation by leaving valuable property unprotected is about as much to blame for its theft as the one who steals it." Such itemmetric variation is necessary to examine covariation between item properties and item validity.

Bipolar Adjective Rating Scales. All samples were rated by acquaintances who knew them well with the Bipolar Adjective Rating Scales (BARS; Johnson, 1991, 1997; Johnson & Ostendorf, 1993), designed explicitly to gather acquaintance ratings on the seven broad dimensions of personality assessed in the original Hogan Personality Inventory (Hogan & Johnson, 1981). Five of these scales correspond directly to what is today called the Five Factor Model (FFM; Costa & McCrae, 1992). German raters used the German translation of the BARS (Johnson, 1991). The BARS contains 49 7-step Likert items anchored by two trait terms. Item ratings can be summed as follows to yield five scores corresponding to the FFM: Factor I, *Sociality*; Factor II, *Likeableness*; Factor III, *Conventionality*; Factor IV, *Poise*; and Factor V, *Mentality*. The composition of the five scales and their reliability estimates in the present study are presented in Table 1. Further details on the construction, reliability, and validity of the BARS can be found in Johnson (1991, 1997).

Assessment Procedures

The American participants completed the original 480 item version of the CPI (Gough, 1975), but only responses to the 462 items retained in the revised version (Gough, 1987) were analyzed in the present study. The German participants completed the authorized German translation of the 462-item revised CPI. Subjects chose acquaintances to complete the BARS, and the acquaintances returned their ratings in sealed envelopes to the investigator. Subjects in sample 1 were rated by three acquaintances; subjects in samples 2 and 3 were rated by two acquaintances. Ratings were averaged within the samples. Table 1 presents alpha reliability estimates based on both rater intraclass correlations and inter-item correlations. All raters indicated they knew the ratees at least "somewhat" (3 on a 1-5 scale); most indicated they knew the ratee "fairly well" (4) or "very well" (5).

Itemmetric Characteristics Used as Predictor Variables

Social Desirability. *Social desirability* ratings for the CPI items were taken from Appendix D of Gough's (1987) manual. *Desirability extremity* (Holden & Fekken, 1990) was calculated by taking the absolute value of an item's desirability minus the mean (4.9, 4.6, 5.0, 4.4, 5.1 for Factors 1-5) desirability rating for the factor.

Item Ambiguity. Proposed indices of item ambiguity have included direct Likert ratings (Johnson, 1986a), response inconsistency (Benton, 1935), balanced endorsement frequency (Fricke, 1957; Hanley, 1962), an "ambiguity

Table 1
Composition and Reliabilities of Bipolar Adjective Rating Scales

	Sample 1		Sample 2		Sample 3	
	IR[a]	II[b]	IR[a]	II[b]	IR[a]	II[b]
Sociality	.59	.88	.55	.84	.53	.88
talkative-quiet						
sociable-solitary						
outgoing-reserved						
extraverted-introverted						
Likeableness	.59	.85	.53	.85	.52	.78
good natured-irritable						
tactful-blunt						
warm-cold						
diplomatic-outspoken						
cooperative-stubborn						
tolerant-impatient						
agreeable-critical						
empathic-self-centered						
Conventionality	.50	.88	.62	.86	.54	.82
responsible-undependable						
trustworthy-unreliable						
conscientious-negligent						
careful-careless						
persevering-quitting						
rule-abiding-rule-avoiding						
Poise	.67	.84	.41	.83	.59	.78
poised-nervous						
confident-worried						
cheerful-depressed						
self-assured-shy						
healthy-frail						
composed-moody						
Mentality	.60	.67	.50	.75	.61	.76
imaginative-down-to-earth						
aesthetic-inartistic						
creative-ordinary						
intellectual-unreflective						
complex-simple						
clever-naïve						
well-read-unlettered						
learned-unlearned						

[a] IR = Coefficient alpha based on average member intraclass correlations across raters.
[b] II = Coefficient alpha based on inter-item correlations

index" (AMBDEX; Goldberg, 1963) combining inconsistency and endorsement frequency, and the number of responses left blank (Goldberg, 1968). The present study employed all of these indices except blank responses, which showed an extremely skewed distribution and zero reliability in these samples.

Ambiguity ratings were collected from the participants in Sample 1, who rated each CPI item's ambiguity on a 1-5 Likert scale. In a review of ambiguity indices, Johnson (1986a) found rated ambiguity to be more reliable than the other, more indirect, indices of ambiguity.

Response inconsistency (changing one's answer from one administration to the next) was offered by Benton (1935) as an indirect indicator of ambiguity. Benton suggested that persons may respond in one direction to one perceived meaning of an ambiguous item the first time and in the opposite direction to another perceived meaning the second time. The present study used values for the percentage of persons failing to answer CPI items in identical fashion on two occasions in Appendix D of Gough's (1987) manual.

Balanced endorsement frequency is the degree to which endorsement frequency approaches 50 per cent "True" and 50 per cent "False" for items in a dichotomous scoring format such as the CPI's. Fricke (1957) and Hanley (1962) argued that the tendency toward balanced endorsement frequency represents random responding to ambiguous items. The tendency toward balanced endorsement frequency correlates highly with response inconsistency (Goldberg, 1963). Endorsement frequencies for CPI items are available in Gough's (1987) manual. The tendency toward balanced endorsement frequency was computed as .50 – |(endorsement frequency – .50)|. Gough's appendix contains endorsement frequencies for males and females separately, and these separate values were used in some initial analyses. However, results were highly similar for male and female values, so male and female values for these indices were averaged.

AMBDEX (short for *ambiguity index*; Goldberg, 1963) is the percentage individuals changing their response from one administration to the next, divided by the value of the ordinate of the unit normal curve at the point of the item's average endorsement frequency. Goldberg did not believe balanced endorsement frequency was a valid indicator of ambiguity and saw it as a contaminant whose relation to response inconsistency should be statistically removed (see Goldberg, 1963, for a detailed explanation). CPI AMBDEX values for males and females tabled in Goldberg and Rorer's (1964) monograph were used in the current study. Results based on separate male and female AMBDEX values were highly similar so the two values were averaged into an overall AMBDEX index. AMBDEX was found to correlate moderately (about $r = .30$) with ambiguity ratings in a previous study (Johnson, 1988).

Linguistic Properties of Items. Several linguistic itemmetric variables previously studied by Holden, Fekken, and Jackson (1985) and Wiggins and Goldberg (1965) were judged by the author and an advanced psychology student. Interjudge agreement was extremely high (all interjudge correlations were above .9). *Item length* was defined by the number of morphemes. Judges also counted the following types of negations: (a) *direct negatives*, which are instances of *not*, *n't*, and *never*; (b) *implicit negatives*, which include words with the prefixes *in-*, *im-*, and *un-*; and (c) the number of *negative qualifiers* such as seldom, rarely, etc. The sum of all three types became the variable *total negations*. Judges also counted the total number of *frequency qualifiers* (words such as *frequently*, *occasionally*, *often*, and *seldom*; Simpson, 1944).

Item Content Type. Following the methodology of Angleitner et al. (1986) and Werner and Pervin (1986), the content type of CPI items was rated by three judges (the author and two advanced psychology students). Judges assigned each CPI item to one of the following categories: (a) *specific, observable behaviors*; (b) *broad, observable behavioral traits*; (c) *specific, nonobservable psychological reactions* to particular situations, either real or imagined; (d) *broad, nonobservable psychological traits* (prevailing moods, typical thought patterns, abilities) that endure over significant periods of time; (e) *physiological reactions*; (f) *biographical facts* about the past; and (g) *attitude*/belief/opinion statements where the self is not the focus of reference.

The kappa coefficient of agreement between judges was .58, a level highly comparable to values found in similar studies (Angleitner & Riemann, 1991; Löhr, 1977; Werner & Pervin, 1986). Agreement was perfect for 54% of the items and two of three judges agreed on 38% of the items. Items with no agreement were discussed until a decision on the most appropriate category was reached. If a majority of judges assigned the item to a content type, a dummy variable for that content type was given a value of 1; otherwise the value for the dummy variable was 0.

Item-Trait-Indicativity. Item indicativity is a specific case of act prototypicality[2] (Buss & Craik, 1983; Borkenau, 1990). To measure act prototypicality, Borkenau (1990) told research participants that a fictitious

[2] Trait indicativity is similar to, but subtly different from, the property of *item prototypicality* discussed by Broughton (1984). Broughton's items were single trait adjectives from an adjective checklist. He considered these items to represent trait descriptions that belonged to a more abstract, superordinate, dispositional category. Item prototypicality was conceptualized as the degree to which an item was a good exemplar or core member of the superordinate

person *A* had performed act *X* and then asked the participants to what extent that Person *A* had Trait *Y*. Acts perceived by most individuals as strongly indicating the trait were considered prototypical for that trait. In the present study, the acts being judged are speech acts, namely, the endorsement of personality items.

Item indicativity is also similar to the concept of item subtlety (Holden, Fekken, & Jackson, 1985; Holden & Jackson, 1979). Subtlety is often measured by the percentage of individuals who can not identify the standard keying of an item or by ratings of lack of indicativity for the trait the scale is purported to measure. These typical ways of assessing subtlety, however, measure indicativity only for the single trait corresponding to the item's standard keying. The present research, rather than assuming that items imply only one trait, measures each item's indicativity across all five domains of the FFM.

The present study used two measures of trait indicativity. The first was McCrae, Costa, and Piedmont's (1993) indicativity ratings for the CPI. McCrae et al. had two professional psychologists and two psychology students rate independently the degree to which an endorsement of each item from the CPI indicated the positive or negative pole of each dimension of the FFM. Ratings were made on a five-point scale from –2 (strongly indicative of the negative pole) to 0 (irrelevant to the dimension) to +2 (strongly indicative of the positive pole). Correlations between judges' ratings ranged from .23 to .74 (median = .56), all significant at the .001 level. The ratings were then averaged and rescaled such that a –1 meant that all four judges thought than an endorsement of the item was strongly indicative of the negative pole and +1 meant that all four judges felt the item was strongly indicative of the positive pole of the dimension being considered.

The second measure followed Duff's (1965) procedure. Participants in sample one — after completing the CPI in normal fashion — were instructed to judge which dimension from the five-factor model each CPI item indicated the most. The percentage of subjects choosing a dimension indicates how strongly the item indicates a single FFM dimension. To compare the two measures of indicativity, these five percentages were correlated with the absolute value of McCrae et al.'s FFM ratings from four judges. (Absolute values were used because the present study's subjects indicated only the dimension and not the direction of scoring for each items.) Despite the

---

category or a weak exemplar or peripheral member of the category. Although Broughton referred to personality items as "*indicants* of dispositional constructs" (p. 1336, my emphasis), Broughton's study relied on semantic judgments about the relation between item and dispositional category rather than personality judgments of a hypothetical person who endorsed a personality item in sentence format.

slightly different descriptions of the five factors used by the psychologists and the student judges and the ipsatized scoring of the student judgments, the following correlations were observed: Factor I, (r =.80, $p < .01$); Factor II, $r = .28$, $p < .01$); Factor III, ($r = .49$, $p < .01$); Factor IV, ($r = .59$, $p < .01$); and Factor V, $r = .55$, $p < .01$).

Validity Coefficients Used as Criterion Variables

Validity is a multidimensional concept (Johnson, 1981); one might say that an item has as many "validities" as there are types of validity and criteria for establishing validity. The validities of CPI items for the present study were computed according to one of the major purposes of the CPI as described by its author, Harrison Gough. According to Gough (1987), the CPI was designed: "… to identify individuals who will be evaluated and described in particular and interpersonally significant ways" (p. 4). Within this stated purpose, personality ratings generated by knowledgeable acquaintances would be an appropriate criterion for computing external validity coefficients. Other writers (Funder, 1991; Hofstee, 1994) have underscored the fundamental importance of acquaintance ratings for establishing validity.

Although the FFM has its critics (Block, 1995, 2001; Briggs, 1989), factor analyses of personality ratings (Goldberg, 1993) and self-report questionnaire items (Ostendorf & Angleitner, 1992) indicate sufficient redundancy within the realm of personality description such that this realm can be well represented at a broad level by five general domains. Because the robustness and comprehensiveness of the five factors are widely accepted today (Digman, 1996), the FFM-based BARS was chosen as the instrument for collecting acquaintance ratings.

Pearson correlation coefficients were computed within Sample 1 between the five BARS ratings and responses to each CPI item (where a "true" response was coded as 2, a "false" response as 1, and missing responses — constituting .57% of the responses — were coded as 1.5). These validity coefficients were converted to z-scores by Fisher's r-to-z transformation and became the dependent variables to be predicted by other properties of the CPI items.

Analyses

Because the direction of scoring for the dummy-coded, true-false CPI items is arbitrary, a decision had to be reached on how to handle the sign of the item validity coefficients. For example, CPI item 124, "I am likely not to speak to people until they speak to me" correlated –.31 with the BARS Sociality

rating. The only itemmetric variable that predicts the sign of the item validity coefficient is trait indicativity. In this case, the –.75 indicativity judgment for Sociality correctly predicts the sign of the correlation between Sociality ratings and responses to this item. But predictions for itemmetric variables such as item ambiguity and length do not specify whether the expected correlation between acquaintance ratings and item responses is positive or negative. Depending on how one considers a validity coefficient's sign, CPI item 124 could be regarded as a valid item for predicting Sociality if scored in reverse or an exceptionally bad item showing "inverse validity" for that criterion.

Several methods for dealing with the sign of validity coefficients were explored, including using the absolute values of validity coefficients, applying the sign from the indicativity ratings to the other itemmetric predictors, reversing the sign on non-indicativity predictors for negatively-signed validity coefficients, and analyzing positively and negatively signed validity coefficients separately. The most interpretable and consistent results were found by applying the sign from the indicativity ratings to the other itemmetric predictors. For example, because the Sociality indicativity judgment for CPI item 124 was negative, negative signs were applied to the other itemmetric variables.

Pearson correlation coefficients were then computed among all itemmetric variables in Sample 1, including the validity coefficients, across the 462 CPI items. The initial correlational analyses were followed by forward multiple regression analyses in Sample 1 for each factor dimension. The validity coefficients were regressed on the itemmetric properties to determine the amount of unique variance in validity explained by each itemmetric variable. Criteria used were $p = .05$ to enter, $p = .10$ to remove, and tolerance = .0001.

Next, scoring keys were constructed by identifying items predicted to have high or low validity from the optimal regression questions. For each factor, the 40 items with the largest predicted validity coefficients (regardless of sign) and the 40 items with predicted validity coefficients closest to zero were identified. The former items were labeled "itemmetrically favorable," and the latter, "itemmetrically unfavorable." Items with negatively signed validities were marked as reverse scored in the scoring key. This procedure yielded ten scoring keys for scoring CPI protocols in Samples 2 and 3. Participants from Samples 2 and 3 thus received 10 scores: five CPI scores (one for each factor) that were predicted to correlate significantly with informant ratings for each respective factor and five CPI scores that were predicted to correlate zero with informant ratings for their respective factors. McNemar's (1969, p. 158) *t*-test for correlated correlation coefficients was used to assess whether the scores based on itemmetrically favorable items actually predicted informant ratings better than scores based on itemmetrically unfavorable items.

Results

Simple Correlates of Item Validity

Correlations between the itemmetric indices and the validity coefficients are presented in Table 2. Social desirability, ambiguity, item length, and trait indicativity all correlated positively with item validity across the five factors. Social desirability extremity correlated with validity across all factors except Conventionality, replicating findings that Holden and Fekken (1990) obtained for an inventory of psychopathology. Item content referring to specific behaviors, behavioral traits, psychological reactions, and biographical facts was associated with higher validity for the Sociality factor. Content referring to psychological reactions was also associated with higher validity for the Mentality factor. Items mentioning psychological traits showed higher validity for the Likeableness and Poise factors. Physiological item content correlated positively with validity only for the Poise factor. Attitude/opinion content was inversely related to validity for the Sociality factor and positively related to validity for the Likeableness, Poise, and Mentality factors.

The significant positive correlations between item validity and social desirability, desirability extremity, ambiguity, and item length are more consistent with the predictions of the performative view than the constative view of item responses. The variable considered to be most important from the performative view, trait indicativity, correlated with validity as predicted. The results for content type depended on the factor dimension under consideration. For Factor I, Sociality, some results seemed to favor the constative view. Three forms of content referring to overt behavior correlated positively with validity and one covert content type correlated negatively. However, the highest correlation for sociality was with psychological reaction content type, which is contrary to the predictions of the constative view. The remaining significant correlations for the other four factor dimensions were with covert content type, which is again contrary to the constative view.

The magnitudes of most of the simple correlations are small. The next questions concern (a) how much better can item validity be predicted by a combination of itemmetric variables in multiple regression and (b) which itemmetric variables contribute most to the prediction of validity.

Multiple Regression and Cross-Validation at the Scale Level

The variables showing significant beta weights in the prediction of item validity are reported in Table 3. Trait indicativity entered the regression for all

Table 2
Pearson Correlations between Itemmetric Properties and Item Validities

Itemmetric Properties	Item Validity Coefficients				
	Sociality	Likeableness	Conventionality	Poise	Mentality
Social Desirability	.39***	.17***	.15**	.18***	.25***
Desirability Extremity	.30***	.16***	.07	.18***	.19***
Ambiguity Indices					
Likert Ratings	.23***	.13**	.09	.21***	.16**
Balanced Endorsement	.30***	.14**	.12*	.20***	.19***
AMBDEX	.22**	.14**	.12*	.21***	.17***
Linguistic Variables					
Item Length	.22***	.14**	.09*	.20***	.15**
Direct Negatives	.00	.01	.03	.05	.03
Implicit Negatives	.06	.06	.05	.08	.06
Negative Qualifiers	.04	.05	.06	.04	–.01
Total Negatives	–.02	.06	.07	.08	.05
Frequency Qualifiers	.09	.06	.07	.01	–.02
Content Type					
Specific Behavior	.25***	.08	.03	.05	.00
Behavioral Trait	.12**	.03	–.02	.08	.01
Psychological Reaction	.29***	.03	.09	.09	.07***
Psychological Trait	.04	.13**	.03	.10*	.03
Physiological Reaction	.00	–.06	.07	.17***	.03
Biographical Fact	.11*	.01	.09	–.03	.01
Attitude/Opinion	–.10*	.13**	.00	.17***	.12*

Table 2 (*cont'd*)

Itemmetric Properties	Item Validity Coefficients				
	Sociality	Likeableness	Conventionality	Poise	Mentality
Trait Indicativity – Professional Ratings					
Extraversion	.51***	–.08	–.25***	.16**	.09
Agreeableness	.13**	.18**	.09	.06	.14**
Conscientiousness	.16***	.04	.09*	.03	.19***
Neuroticism (reversed)	.25***	.28***	.04	.23***	.19***
Openness to Experience	.21***	.10	.03	.19***	.30***
Trait Indicativity – Lay Category Judgments					
Social Assertiveness	.47***	–.11	–.29***	.16**	.05
Empathy	.15**	.12**	.09	.04	.18***
Conventionality	.01	.10*	.16***	.02	.14**
Emotional Maturity	.16**	.27***	.05	.21***	.16**
Intelligence-Imagination	.15**	.03	.05	.16***	.27***
Serial Position	.18***	.11*	.06	.22***	.16***

* $p < .05$. ** $p < .01$. ***$p < .001$ (all two-tailed; difference in significance levels for apparently identical correlation coefficients due to rounding).

Table 3
Itemmetric Properties with Significant Beta Weights for Predicting Validity Coefficients

Itemmetric Properties	Beta Weights				
	Factor I	Factor II	Factor III	Factor IV	Factor V
Social Desirability	.49				
Desirability Extremity					
Ambiguity Indices					
Ambiguity Rating	–.35				
Linguistic Variables					
Frequency				–.10	
Total Negatives	–.13				
Content Type					
Attitude/Opinion				.14	
Physiological				.12	
Trait Indicativity					
Professional Ratings	.44	.18		.22	.30
Lay Judgments			.16		
Multiple Correlation	.58	.18	.16	.31	.30

five factor dimensions. For three factor dimensions (Likeableness, Conscientiousness, and Mentality), no itemmetric variable was able to contribute additional predictive value beyond trait indicativity. For the Sociality and Poise domain, three additional variables beyond trait indicativity contributed to the prediction of validity. Valid items for Sociality were identified with a positive weight for social desirability and negative weights for ambiguity and the number of linguistic negations. The regression equation for predicting item validity for Poise included a negative weight for frequency modifiers and positive weights for attitude/opinion content and physiological content.

Table 4 compares the validities of scales built from items predicted to be valid versus scales built from items predicted to have zero validity on the basis of their itemmetric properties. When all of the best itemmetric predictors were used in regression equations to predict item validity, the scales predicted to be valid showed statistically significant correlations with acquaintance ratings for all five factor dimensions in the American sample and for all factor dimensions except Mentality in the German sample. None of the scales built from items

Table 4
Acquaintance Rating Correlates of Itemmetrically Favorable and Unfavorable Scales

	Scales Predicted to be Valid	Scales Predicted to Invalid
American Sample, $N = 75$		
Factor I	.44***[b](.46***[c])	.11(–.07)
Factor II	.24*	.06
Factor III	.39***[b]	.02
Factor IV	.24*(.33**[c])	.16(–.16)
Factor V	.27*[b]	–.05
German Sample, $N = 96$		
Factor I	.35***[c](.38***[b])	–.37***(.05)
Factor II	.38***[c]	.07
Factor III	.37***[c]	.19
Factor IV	.50***[d](.56***[d])	.03(–.08)
Factor V	.18	.07

Note. Correlations in parentheses are for scales whose items were selected by trait indicativity alone, rather than by regression with all significant itemmetric predictors.
[a]predicted high r greater than predicted low r, two-tailed t, $p < .10$. [b]predicted high r greater than predicted low r, two-tailed t, $p < .05$. [c]predicted high r greater than predicted low r, two-tailed t, $p < .01$. [d]predicted high r greater than predicted low r, two-tailed t, $p < .001$.
*$p < .05$. **$p < .01$. ***$p < .001$ (all two-tailed).

predicted to have low validity showed significant correlations with acquaintance ratings except Sociality in the German sample, which correlated *negatively* with the ratings. The differences between the predicted-valid and predicted-invalid scales were statistically significant in seven out of ten cases by McNemar's (1969, p. 158) t-test for correlated correlation coefficients. Thus, the relation between itemmetric properties and validity found in Sample 1 was cross-validated at the scale level in Samples 2 and 3.

The finding that specific itemmetric predictors of validity differed across factors led to one further, unplanned analysis. For all factors, trait indicativity entered the regression equation, but for Sociality and Poise, additional

predictors entered the equations. To see how well trait indicativity alone predicted validity for Sociality and Poise, the 40 items with the highest trait indicativity and 40 items with zero or near-zero trait indicativity were identified for the Sociality and Poise factors, and scales constructed from these items were correlated with acquaintance ratings in Samples 2 and 3.

In both samples, scales designed on the basis of trait indicativity alone to predict Sociality and Poise correlated slightly higher with the ratings than the scales built from regression equations using all significant itemmetric predictors. Thus, scales constructed wholly on the basis of trait indicativity — the single itemmetric characteristic deemed important by the performative view — showed better validity than scales employing items predicted to be valid by regression equations with multiple itemmetric predictors.

Discussion

Summary of Findings and Implications for the Constative and Performative Views

Item Content. Item content type turned out to be a weak and inconsistent predictor of item validity. Many of the statistically significant point-biserial correlations were on the order of only about .10. The three more overt content types deemed important by the constative view — specific behaviors, behavioral traits, and biographical facts — were associated with validity only for Sociality. The seven other positive correlations, contrary to the constative view, involved more covert content types. Almost all content effects were overshadowed by other variables in the multiple regressions. Only attitude/opinion content and physiological content were retained in the final regression equation for Poise, both with positive weights.

Item Ambiguity. Item ambiguity, measured by all indices, showed small but consistent positive correlations with item validity across the five factors. This contradicts the constative view and replicates the findings of Elias (1951), Gordon (1953), Isard (1956), and Johnson (1988), all of whom reported a positive relation between item ambiguity and validity. However, rated item ambiguity in the current study entered only one regression equation (for Sociality), and its weight was negative. Ambiguity might therefore be relatively unimportant for four of the Big 5 dimensions and might function as a suppressor variable (Paulhus, Robins, Trzesniewski & Tracy, 2004, this volume) for Sociality.

Perhaps a clearer understanding of ambiguity and validity will require a more fine-grained approach to item ambiguity, similar to the German distinction

between *Verständlichkeit* (comprehensibility) and *Mehrdeutigkeit* (possessing many meanings). Both the constative and performative views suggest that ambiguous (in the sense of incomprehensible) items would lack validity, but for different reasons. Whereas the constative view emphasizes the need for clear behavioral referents that can be compared to one's actual behavior, the performative view suggests that unclear, incomprehensible items cannot plainly indicate traits. On the other hand, the performative view suggests that items possessing several comprehensible meanings can possess validity because respondents can express a personality disposition by the meaning they see in an item.

Social Desirability. Contrary to the predictions of the constative view, social desirability correlated positively with item validity across all five factors, and social desirability extremity correlated with validity for four factors. This finding is consistent with research by Holden and Fekken (1990), who found that extremely desirable or undesirable items were more valid predictors of psychiatric diagnosis than items with average desirability.

The findings for social desirability support the performative view of item responses, which says that, at least under *typical* testing conditions, the social desirability of people's personality inventory item responses corresponds to the social desirability of their behaviors in everyday life. Individual differences in motivation and grasp of constitutive rules may explain who is most likely to increase the desirability of their responses under highly evaluative testing contexts such as personnel selection (Johnson, 1986b, 1987; Wiggins, 1966, 1973).

Trait Indicativity. The data from the present study support the performative view's suggestion that trait indicativity is the primary determinant of item and scale validity. Simple correlations between item validity and trait indicativity were among the highest of all validity/item-characteristic correlations. Trait indicativity entered into the regressions for all five factors. Finally, correlations between acquaintance ratings and scales based solely on items with high trait indicativity were larger than correlations between ratings and scales built from items selected by regression equations with multiple itemmetric predictors.

The findings concerning trait indicativity in the current study parallel findings from extensive studies of item subtlety presented by Holden (Holden, 1989; Holden & Fekken, 1990; Holden, Fekken, & Jackson, 1985; Holden & Jackson, 1979) and Gynther (see Gynther & Burkhart, 1983, for a review). Subtle items are presumably valid but low in trait indicativity. Both Gynther's and Holden's research — conducted principally on measures of psycho-

pathology — indicate that subtle items are consistently less valid than obvious items, and this supports earlier research on the topic (Duff, 1965; Goldberg & Slovic, 1967). The current research extends these findings to the five factor domains of normal personality.

The centrality of trait indicativity for item and scale validity underscores the importance of the *social significance* of item responses, supporting the performative view of item response dynamics over the constative view. It is insufficient for respondents to disclose accurate, objective, information about themselves with responses that correspond veridically to their actual thoughts, feelings, and behaviors. Responses that veridically disclose facts without clear social implications cannot predict how someone will be regarded by others. In fact, other research (Johnson, 1990) indicates that under some circumstances, stretching the truth through exaggeration creates a more valid impression than disclosing unexaggerated, accurate information.

Caveats

Readers should keep in mind that the findings in this study apply only to the restricted range of itemmetric properties in polished items appearing in a published inventory. Both the restriction of range and less-than-perfect reliabilities of the measures attenuated to correlations between itemmetric properties and item validity.

Second, the study relied on relatively small samples of college students and restricted itself to one particular True-False personality inventory, the CPI. While 194 Form 462 CPI items originated in the MMPI, extending the present findings to the MMPI domain, using larger, more diverse samples and different personality inventories employing other response formats (Likert, paired comparisons, triads, etc.) would show how well the present results generalize. The study also did not measure all possible item properties. Other item properties that might be studied in future research include response concordance for monozygotic versus dizygotic twins, reference to past, present, or future events, and response reaction time.

Next, we must keep in mind that item validity was limited in the present study to correlations between items and acquaintance ratings defined by the five-factor model. If one wished to define item validity in terms other than acquaintance ratings (e.g., life events such as illnesses), we might have found results that differed from those in the present study (e.g., physiological items may be much more valid than other types of items for predicting illnesses).

Finally, the present study is purely correlational, and might be improved by taking a more experimental approach. In a truly experimental study, one

can write items that differ according to a limited number of characteristics of interest, while holding other item properties constant.

Future Directions

The unmistakable importance of trait indicativity for item and scale validity presents a powerful take-home message about writing items for personality scales. Although the traditional, constative criteria for item writing such as clarity, brevity, and simplicity (Hendriks, 1997; Wolfe, 1993) shouldn't be ignored, the crucial characteristic of an item seems to be its ability to clearly indicate the personality trait for which it is being scored.

To determine whether an item response might serve as a good performative for indicating a particular trait, we need to know the constitutive rules (Wiggins, 1974/1997) that map speech acts onto their social meanings. Personality scale authors who use the rational-intuitive method for writing items have been relying upon their personal understanding of constitutive rules for writing items. An alternative strategy would be to present potential items to a panel of judges with instructions to describe the personality of a hypothetical person who has endorsed the item (Johnson, 2002). When judges strongly agree on an item's trait indicativity, the item becomes a good candidate for a scale measuring that trait.

Even though pre-screening items for consensus on trait indicativity should improve the validity of personality scales, we need to keep in mind factors beyond item characteristics that affect validity. Two such factors are characteristics of the person responding and the purpose for which the test scores are to be used (Löhr, 1977). It is possible that these characteristics might interact with the itemmetric characteristics examined in the present study. The adverse impact of an item's syntactic complexity on its validity might be greater for respondents with lower verbal intelligence. Whether validity is affected by item content (e.g., observable behavior versus psychological states) might depend upon the psychological-mindedness of the respondent Burkhart, Gynther, & Christian, 1978; Gough, 1975). The relation between validity and social desirability might be different in highly evaluative testing contexts such as personnel selection. Future research that simultaneously considers the influences of item properties, person variables, and testing context will form a more complete picture of the way in which item characteristics affect item validity.

One recent research study has examined a characteristic of persons that apparently interacts with an item's trait indicativity. This characteristic, which I dubbed *construal communality* (Johnson, 2002), refers to the extent to which an individual's understanding of constitutive rules corresponds to the

constitutive rules shared by the rest of the linguistic community. In this study, the consensual meaning of item responses was defined by the average trait-indicativity ratings provided by participants who also completed the personality inventory. Construal communality was measured by the distance between an individual's ratings and the average ratings for the entire sample. Validity, defined as correlations between self-report scores and acquaintance ratings, was higher for participants possessing higher construal communality, although significantly so only for Factor I, Sociality.

While construal communality (i.e., knowledge of constitutive rules) might be necessary for producing valid item responses, it may be insufficient. Respondents must actually apply this knowledge to make an item response a performative vehicle rather than a constative description. This is especially true when a constatively accurate response to an item (e.g., "False" to "I read at least ten books a year") fails to indicate a trait that everyone agrees is characteristic of the person ("*X* is intelligent and cultured"). Respondents employ many different cognitive strategies while responding to items (Gordon & Holden, 1996; Holden & Fekken, 1990). Some apparently use a constative strategy such as trying to recall one or more specific behaviors or experiences. Others think in more performative terms, such as their general traits and what others have said about them. Holden and his colleagues focused on the typical cognitive strategies evoked by items and found that items evoking the recollection of specific behaviors were less valid than items that encouraged respondents to think about themselves in terms of general traits, comparisons to others, and statements by others. When respondents are told to use particular strategies, similar results are obtained (Gordon & Holden, 1998). From a person-centered context, the performative view of item responses says that respondents who habitually think about responding in terms of traits are considering the trait indicativity of the item and will therefore tend to provide more valid responses.

Item validity depends not only on item characteristics, the respondent's traits, and the testing context, but also on the factor being measured. There appears to be something especially unique about the first factor of the Big Five. Itemmetric variables predicted validity for Factor I far better (simple r from trait indicativity = .51; Multiple R = .58) than for any other factor. Johnson (2002) found that construal communality was related to validity significantly only for Factor I. John and Robins (1993) found that, out of the five factors, self-peer agreement on personality ratings was highest for Factor I. Reasons for the higher predictability of validity for Factor I items need to be investigated.

Finally, this study has implications for research on the cognitive processes underlying responses to individual personality items (Angleitner,

John, & Löhr, 1986; Bond, 1987; Goldberg, 1963; Jackson, 1986; Kuncel, 1973; Rogers, 1974). Much existing research on cognitive processes mediating item responses has taken a constative view. Hogan (1991, p. 901) describes the constative view as follows: "On this view, a person reads an item on a personality measure (e.g., 'I often have strange and unusual thoughts'), reviews his or her memory, compares the item to the relevant memory trace, decides whether the item matches the memory trace, and then endorses or rejects the statement as being self-descriptive. This is a pretheoretical account of item response dynamics that rests on a naive notion of how memory seems to work." Wiggins (1973, pp. 382-383) compares the unrealistic assumptions of these cognitive models to the untenable tenets of early introspectionism and, like Hogan, describes this approach as "pretheoretical" (p. 401). If we are to borrow ideas for cognitive models of item response dynamics from other disciplines, perhaps we should heed Jerry Wiggins' call to move beyond 19th-century psychophysics and to embrace 20th-century speech act theory (Wiggins, 1974/1997).

References

Allan, K. (1994). Speech act theory — An overview. In R. Asher (Ed.), *Encyclopedia of Language and Linguistics* (Vol. 8, pp. 4127-4138). Oxford: Pergamon Press.

Angleitner, A. (1981). *Teststatistische kennwerte der items aus 10 Deutschsprachigen persönlichkeitsfragenbogen* (Test-statistical, characteristic values of items from ten German language personality questionnaires). Bielefeld: Arbeitsberichte aus dem Projekt Persönlichkeitsfragebogen (Nr. 2), Universität Bielefeld.

Angleitner, A., John, O. P., & Löhr, F.-J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 61-108). New York: Springer-Verlag.

Angleitner, A. & Riemann, R. (1991). What can we learn from the discussion of personality questionnaires for the construction of temperament inventories? In J. Strelau & A. Angleitner (Eds.), *Explorations in temperament: International perspectives on theory and measurement* (pp. 191-204). New York: Plenum Press.

Austin, J. L. (1962). *How to do things with words*. Oxford: Oxford University Press.

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

Benton, A. L. (1935). The interpretation of questionnaire items in a personality schedule. *Archives of Psychology*, *90* (No. 190).

Blaney, P. H. (1991). Not personality scales, personality items. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Meehl, Vol. 2, Personality and psychopathology* (pp. 54-70). Minneapolis, MN: University of Minnesota Press.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117,* 187-215.

Block, J. (2001). Millennial contrarianism. *Journal of Research in Personality*, *35*, 98-107.

Bond, J. A. (1987). The process of responding to personality items: Inconsistent responses to repeated presentation of identical items. *Personality and Individual Differences*, *8*, 409-417.

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology*, *58*, 381-396.

Borkenau, P. & Ostendorf, F. (1989). Descriptive consistency and social desirability in self- and peer reports. *European Journal of Personality*, *3*, 31-45.

Briggs, S. R. (1989). The optimal level of measurement for personality constructs. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 246-260). New York: Springer.

Broughton, R. (1984). A prototype strategy for construction of personality scales. *Journal of Personality and Social Psychology*, *47*, 1334-1346.

Buchwald, A. M. (1961). Verbal utterances as data. In H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science: Symposia of scientists and philosophers* (pp. 461-472). New York: Holt, Rinehart and Winston.

Burkhart, B. R., Gynther, M. D., & Christian, W. L. (1978). Psychological mindedness, intelligence, and item subtlety endorsement patterns on the MMPI. *Journal of Clinical Psychology*, *34*, 76-79.

Buss, D. M. & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90*, 105-126.

Costa, P. T., Jr. & McCrae, R. R. (1987). Neuroticism, somatic complaints, and disease: Is the bark worse than the bite? *Journal of Personality*, *55*, 299-316.

Costa, P. T., Jr. & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R™) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Digman, J. M. (1996). The curious history of the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 1-20). New York: Guilford.

Duff, F. L. (1965). Item subtlety in personality inventory scales. *Journal of Consulting Psychology*, *29*, 565-570.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

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

Elias, G. (1951). Self-evaluation questionnaires as projective measures of personality. *Journal of Consulting Psychology*, *15*, 496-500.

Emler, N. P. (1999). Moral character. In V. J. Derlega, B. A. Winstead, & W. H. Jones (Eds.), *Personality: Contemporary theory and research* (2nd ed., pp. 376-404). Chicago: Nelson-Hall.

Fricke, B. G. (1957). A response bias (B) scale for the MMPI. *Journal of Counseling Psychology*, *4*, 149-153.

Funder, D. C. (1991). Global traits: A Neo-Allportian approach to personality. *Psychological Science*, *2*, 31-39.

Funder, D. C. (2001). *The personality puzzle* (2nd ed.). New York: W. W. Norton.

Funder, D. C. & Colvin, C. R. (1997). Congruence of others' and self-judgments of personality. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 617-647). San Diego, CA: Academic Press.

Goldberg, L. R. (1963). A model of item ambiguity in personality assessment. *Educational and Psychological Measurement*, *23*, 467-492.

Goldberg, L. R. (1968). The interrelationships among item characteristics in an adjective checklist: The convergence of different indices of item ambiguity. *Educational and Psychological Measurement*, *28*, 273-296.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26-34.

Goldberg, L. R. & Rorer, L. G. (1964). *Test-retest item statistics for the California Psychological Inventory* (ORI Research Monograph, Vol. 4, No. 1). Eugene, OR: Oregon Research Institute.

Goldberg, L. R. & Slovic, P. (1967). Importance of test item content: An analysis of a corollary of the deviation hypothesis. *Journal of Counseling Psychology*, *14*, 462-472.

Gordon, L. V. (1953). Some interrelationships among personality item characteristics. *Educational and Psychological Measurement*, *13*, 264-272

Gordon, E. D. & Holden, R. R. (1996). Use of item ratings to examine personality test item cognitive response processes. *Personality and Individual Differences*, *21*, 897-905.

Gordon, E. D. & Holden, R. R. (1998). Personality test item validity: Insights from "self" and "other" research and theory. *Personality and Individual Differences*, *25*, 103-117.

Gough, H. G. (1975). *Manual for the California psychological inventory* (rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

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

Gough, H. G. (1991). Some unfinished business. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Meehl, Vol. 2, Personality and psychopathology* (pp. 114-136). Minneapolis, MN: University of Minnesota Press.

Gough, H. G. (1996). *CPI manual: Third edition*. Palo Alto, CA: Consulting Psychologists Press.

Gynther, M. D. & Burkhart, B. R. (1983). Are subtle MMPI items expendable? In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 2, pp. 115-132). Hillsdale, NJ: Erlbaum.

Hanley, C. (1962). The "difficulty" of a personality item. *Educational and Psychological Measurement*, *22*, 577-584.

Hendriks, A. A. J. (1997). *The construction of the Five Factor Personality Inventory*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality*, *8*, 149-162.

Hogan, R. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 873-919). Palo Alto, CA: Consulting Psychologists Press.

Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives*. New York: Guilford Press.

Hogan, R., Carpenter, B. N., Briggs, S. R., & Hansson, R. O. (1985). Personality assessment and personnel selection. In H. J. Bernardin & D. A. Bownas (Eds.), *Personality assessment in organizations* (pp. 21-52). New York: Praeger.

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

Hogan, R., & Johnson, J. A. (1981, September). *The structure of personality*. Paper presented at the 89th Annual Convention of the American Psychological Association, Los Angeles, CA.

Holden, R. R. & Fekken, G. C. (1990). Structured psychopathological test item characteristics and validity. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 35-40.

Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality*, *19*, 386-394.

Holden, R. R. & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology*, *47*, 459-468.

Isard, E. S. (1956). The relationship between item ambiguity and discriminating power in a forced-choice scale. *Journal of Applied Psychology*, *40*, 266-268.

Jackson, D. N. (1986). The process of responding in personality measurement. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 123-142). New York: Springer-Verlag.

John, O. P. & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*, 521-551.

Johnson, J. A. (1981). The "self-disclosure" and "self-presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology*, *40*, 761-769.

Johnson, J. A. (1986a). *Ambiguity, subtlety, and validity of items in the California Psychological Inventory*. A combined version of papers presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada, August, 1984, and the 57th Annual Meeting of the Eastern Psychological Association, New York, April, 1986.

Johnson, J. A. (1986b, August). *Can job applicants dissimulate on personality tests*? Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, DC.

Johnson, J. A. (1987, August). *Dissembling on the Hogan Personality Inventory during simulated personnel selection*. Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY.

Johnson, J. A. (1988, April). *Item ambiguity as a moderator of correlations between observer ratings and self-reports*. Paper presented at the 59th Annual Meeting of the Eastern Psychological Association, Buffalo, NY.

Johnson, J. A. (1990, June). *Unlikely virtues provide multivariate substantive information about personality*. Paper presented at the 2nd Annual Meeting of the American Psychological Society, Dallas, TX.

Johnson, J. A. (1991). *Interpreter's guide to the bipolar adjective rating scales (BARS)*. Unpublished manuscript, Pennsylvania State University, DuBois.

Johnson, J. A. (1997). Seven social performance scales for the California Psychological Inventory. *Human Performance*, *10*, 1-30.

Johnson, J. A. (2002, July). Effect of construal communality on the congruence between self-report and personality impressions. In P. Borkenau & F. M. Spinath (Chairs), *Personality judgments: Theoretical and applied issues*. Invited symposium for the 11th European Conference on Personality, Jena, Germany.

Johnson, J. A. & Ostendorf, F. (1993). Clarification of the Five Factor Model with the Abridged Big-Five Dimensional Circumplex. *Journal of Personality and Social Psychology*, *65*, 563-576.

Keren, A. (1979). *Inhaltlich-semantische analyse der items aus fünf deutschsprachigen persönlichkeitsfragebogen* (Content-semantic analyses of items from five German language personality questionnaires). Unpublished master's thesis, Universität Bonn.

Knowles, E. E. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*, 312-320.

Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement*, *33*, 545-563.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports 3* (Monograph Supplement 9), 635-694.

Löhr, F.-J. (1977). *Formale und inhaltliche Merkmale von Fragebogenitems — Zum Einfluß von Itemeigenschaften auf den Beantwortungsprozess bei deutschen Persönlichkeitsfragebogen* (Formal and content characteristics of questionnaire items: On the influence of item characteristics on the response process in German personality questionnaires). Unpublished master's thesis, Universität Bonn.

Löhr, F. J. & Schütte, W. (1980). *Zur linguistischen Analyse von Fragebogenitems - eine Pilot-Studie*. (Regarding linguistic analyses of questionnaire items: A pilot study). Arbeitsberichte aus dem Projekt Persönlichkeitsfragebogen (No. 1). Universität Bielefeld.

McCrae, R. R., Costa, P. T., Jr., & Piedmont, R. L. (1993). Folk concepts, natural language, and psychological constructs: The California Psychological Inventory and the five-factor model. *Journal of Personality*, *61*, 1-26.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

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

Nicholson, R. A. & Hogan, R. (1990). The construct validity of social desirability. *American Psychologist*, *45*, 290-292.

Ostendorf, F. & Angleitner, A. (1992). On the generality and comprehensiveness of the five-factor model of personality: Evidence for five robust factors in questionnaire data. In G. V. Caprara & G. L. van Heck (Eds.), *Modern personality psychology*: *Critical reviews and new directions* (pp. 73-109). New York: Harvester Wheatsheaf.

Paulhus, D. L., Robins, R. W., Trzesniewski, K. H., & Tracy, J. L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research*, *39(2)*, 303-328.

Rogers, T. B. (1974). An analysis of the stages underlying the process of responding to personality items. *Acta Psychologia*, *38*, 205-213.

Schroeder, D. H. & Costa, P. T., Jr. (1984). Influence of life event stress on physical illness: Substantive effects or methodological flaws? *Journal of Personality and Social Psychology*, *46*, 853-863.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, England: Cambridge University Press.

Simpson, R. (1944). The specific meanings of certain items indicating differing degrees of frequency. *Quarterly Journal of Speech*, 30, 328-330.

van Oort, R. (1997, January). Performative-constative revisited: The genetics of Austin's theory of speech acts. *Anthropoetics: The Journal of Generative Anthropology*, *2(2)*. Retrieved June 10, 2002, from http://www.humnet.ucla.edu/humnet/anthropoetics/Ap0202/Vano.htm.

Werner, P. D. & Pervin, L. A. (1986). The content of personality inventory items. *Journal of Personality and Social Psychology*, *51*, 622-628.

Wiggins, J. S. (1966). Social desirability and "faking good" well. *Educational and Psychological Measurement, 26*, 329-341.

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

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms. *Journal of Personality and Social Psychology*, *37*, 395-412.

Wiggins, J. S. (1997). In defense of traits. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 95-115). San Diego, CA: Academic Press. (Originally presented as an invited address to the Ninth Annual Symposium on Recent Developments in the Use of the MMPI, held in Los Angeles on February 28, 1974.)

Wiggins, J. S. & Goldberg, L. R. (1965). Interrelationship among MMPI item characteristics. *Educational and Psychological Measurement*, *25*, 381-397.

Wolfe, R. N. (1993). A commonsense approach to personality measurement. In K. H. Craik, R. Hogan, & R. N. Wolfe (Eds.), *Fifty years of personality psychology* (pp. 269-290). New York: Plenum.