It's *What* You Ask and *How* You Ask It: An Itemmetric Analysis of Personality Questionnaires[1]

ALOIS ANGLEITNER, OLIVER P. JOHN, and FRANZ-JOSEF LÖHR

Introduction

> "The sight of blood no longer excites me."
> "When I was a child, I was an imaginary playmate."
> "I become homicidal when people try to reason with me."
> "It's hard to concentrate in a room full of mice."[2]

Items from personality questionnaires have long been the butt of humorists and jaded graduate students. What makes these satirical items so funny is that they are instantaneously recognizable as having the correct form and yet their content is patently absurd. The fact that these items succeed as jokes suggests that at least for those familiar with personality scales there are some standard forms for items. This chapter describes three sets of formal item characteristics and demonstrates that the psychometric quality of personality items depends not only on content but also on form.

Historically, research on personality questionnaire construction has focused on the principles of scale construction and neglected the properties of the stimuli from which the scales are constructed. Yet, even the most sophisticated techniques for item grouping may never be able to remedy the flaws introduced by the initial item sample. The need for knowledge about the most basic units of questionnaire scales – the individual items – will be our major concern in this chapter; thus, we will primarily address questions related to the initial stages of test construction: How should the initial item pool be constructed? How should the items be written? How do the stimulus characteristics of the text influence the way in which subjects process the item?

For each of these questions, we try to identify sets of item characteristics that might facilitate adequate interpretation of and response to the individual items. We will describe a series of empirical studies in which we provide operational definitions of these item characteristics and evaluate and compare some widely used questionnaires in terms of the characteristics of their items. Finally, we will provide some initial evidence that these item characteristics are important for obtaining reliable and valid responses from subjects.

The empirical findings summarized here were obtained in a research project on the evaluation of personality questionnaires, which was conducted at the universities of Bonn and Bielefeld in West Germany.[3] To sample the domain of personality questionnaires as comprehensively as possible, we included in our studies a wide variety of frequently used instruments of Anglo-

American origin. In particular, we studied those questionnaires of Eysenck for which German editions were available by 1976, namely the MMQ (Eysenck, 1964), MPI (Eysenck, 1959a, 1959b, 1962), and EPI Forms A and B (Eysenck & Eysenck, 1964; Eggert, 1974); the translated parallel Forms A and B of Cattell's 16PF (Cattell & Eber, 1962, 1964); and the MMPI (Hathaway & McKinley, 1943, adapted by Spreen, 1963).[4] In addition, we included some German-developed questionnaires [i.e., Giessen Test (GT) by Beckmann & Richter, 1972; Persönlichkeits-Interessen-Test (PIT) by Mittenecker & Toman, 1951; Freiburger Persönlichkeits-Inventar (FPI) by Fahrenberg & Selg, 1970, and several questionnaires for children]. In this chapter we will summarize the most important findings from this project, focusing primarily on the former questionnaires, since they are of the most general interest. Their scale labels and the abbreviations used here are given in *Appendix A*.

Overview

This chapter consists of four sections, as outlined in Table 1. We will discuss the tasks of the researcher in constructing an item pool (the construction process) and those of the subjects who later respond to the resulting items (the response process). The lower part of Table 1 lists four sets of item characteristics associated with each stage, viz. (a) the logical relation between the trait construct and its indicator (the item); (b) surface elements of the item text; (c) semantic item properties which are assumed to affect the response process; and (d) item statistics which represent the aggregated response patterns of a sample of subjects. Each section of the chapter will focus on one of these four sets of item characteristics.

In the first section, we discuss some of the central issues that arise during the initial stages of the construction of personality questionnaires. In the construct-oriented approach, the test author first makes decisions about which trait constructs are to be measured and then devises new (or uses existing) construct definitions, which specify an appropriate item universe.[5] One aspect of such item universes is the relation between the items and the construct. We will define several types of potential verbal manifestations of traits and present an empirical study that compares the frequencies of these item types across various questionnaires and scales.

In the second section of the chapter, we discuss the next stage of the questionnaire construction process. As Table 1 shows, the researcher should write or select a set of items guided by the previously established construct definitions while keeping in mind certain rules of thumb about item writing. After a short literature review and a theoretical linguistic analysis, we will present some findings from an empirical study of the surface structures of current questionnaire items, including variables such as item length, syntactic characteristics, and response format. As Table 1 shows, the output of this stage of the researcher's activities, namely the string of words that make up the item,

Table 1. Item characteristics associated with different stages of item pool construction and item response processes, and the sections of this chapter where each is discussed

Stages	*Researcher* Constructs item pool			*Subjects* Respond to each item					Aggregated output
	Item generation		Output	Input	Processing stages			Output	
Specific tasks	Decisions about trait and its relation to observable events	Definition of relevant types of trait manifestations in *verbal* form	Writing or seclecting the items	Reading	Encoding	Comparing to stored information about oneself	Utility control during response selection	Response	Item response parameters
Item characteristics	Semantic trait-item relation (prototypicality)	Logical item-trait relations	Surface elements of item text: length, syntax, and response format		Comprehensibility Ambiguity	Abstractness Self-reference	Social norms and values	Endorsement	Item mean, variance, stability, item-test correlation
Section entitled	"Personality Traits and Personality Questionnaire Items"		"Surface Elements of Questionnaire Items: A Linguistic Analysis"		"Item Characteristics and Item Response Processes"			"The Prediction of Psychometric Characteristics"	

serves as the input to the subjects' item response processes. Consequently, before moving on to item selection based on statistical analyses of item responses, the researcher should check whether the response processes taking place are appropriate, that is, whether subjects process (understand and respond to) the item in the way intended by the researcher.

Thus, in the third section of the chapter, we discuss some frequently postulated item response processes and their associated item characteristics. The model of the item response process outlined in Table 1 suggests that items characterized by a high degree of ambiguity and low comprehensibility will be difficult to encode; abstractness and lack of self-reference of the item text can result in errors during the comparison stage; and references to social norms and values might bias response selection during the stage of utility control. All these item characteristics can affect the subject's decision to endorse the item. Thus, inasmuch as they lead to inappropriate processing of the item (i.e., distort the response process), a subject's response cannot be taken as an indication of his or her standing on the trait that the item was designed to assess.

Whereas the item characteristics discussed in the first three parts of the chapter were all derived directly from the texts of the items, the final section relates these logical, surface, and process characteristics to a variety of item parameters that are based on subjects' response patterns (i.e., item statistics). Using a multiple regression design, we investigate the extent to which the reliability and validity of item responses can be predicted from these item characteristics.

Whereas the present classification of item characteristics has been derived from the theoretical model outlined in Table 1, other authors have preferred to group item characteristics according to the methodology used to obtain measures of these variables. For example, Jones (1965) employed a two-fold classification: (a) sample specific (i.e., derived from subjects' ratings or responses) vs (b) intrinsic (i.e., denoting structural aspects of the item). Thus, using Jones' (1965) scheme and focusing on how we assessed these item characteristics, those discussed in the first and third sections would be grouped together as sample specific and intrinsic (based on judges' ratings of aspects of the item text); those discussed in the second section would also be intrinsic but would not be sample specific (scored directly and objectively from the item text); and those discussed in the last section would be sample specific but not intrinsic (derived from subjects' item responses). This classification corresponds directly to the three categories proposed by Goldberg (1968), viz. item ratings, lexicographic indices, and test-retest statistics, and to those used by Payne (1970), who labeled them judgmental, structural, and metric.

Personality Traits and Personality Questionnaire Items

Traits can be understood in Allport's (1937) sense transformational principles specifying those sets of equivalent stimulus conditions that consistently elicit

similar responses from the individual. Alternatively, traits have been conceptualized as summarizing statements about stable trends in the behaviors and experiences of individuals (Hampshire, 1953). These conceptualizations imply two essential tasks for personality assessment. After deciding which behavioral domain is to be assessed, the researchers must first specify the set of stimulus conditions under which the behaviors of interest are likely to occur. Secondly, a set of rules must be identified which specify how various sets of behaviors can be aggregated. Similarly, the personality researcher who chooses to develop a questionnaire for personality assessment has to define the trait construct to be measured and to develop a set of items that inquire about trait-relevant behaviors and situations, the responses to which thus serve as indicators of the construct. Other important issues arising at later stages of the test construction process, for example, how items should be combined into scales, are discussed elsewhere (Burisch, this volume).

The Origin of Personality Questionnaire Items

In an exemplary program of test development that resulted in the publication of the Personality Research Form, Jackson (1967) started with the development of mutually exclusive, specific definitions of each construct, on the basis of reviews of the relevant research literature; then he and his coworkers generated an initial item pool for each construct with the help of a "grid of situations and behavioral sequences" (Jackson, 1970, p. 67). Moreover, for a subset of items, judges were given theoretically derived descriptions of target persons who exemplified the trait; then the judges were asked to rate the probability that the persons would endorse the items, to establish the degree to which the items could be considered relevant to the construct in question.

A more recent proposal for the task of item pool generation has been advanced by Buss and Craik (1983). In their act frequency approach to personality measurement, Buss and Craik (1983) began with a set of traits selected systematically from a structural model of interpersonal behavior (Wiggins, 1979). The task of identifying the corresponding observable behavior referents was delegated to a large sample of subjects who were asked to generate behavioral acts that could be viewed as manifestations of the traits. Another sample of subjects then rated the prototypicality of these acts vis-à-vis the traits. Thus, the assignment of behavioral acts to traits is based on explicit rules (i.e., prototypicality judgments) and on the consensus of a large number of speakers of a given language. These procedures capitalize on people's semantic knowledge about which behaviors are typical or indicative instances of a trait. The contention that the folk knowledge about personality that has been encoded into language may, in some respects, be as relevant to item generation as that of many personality researchers is supported by Buss and Craik's (1980, p. 389) finding that act prototypicality judgments showed considerable agreement between panels of personological "experts" and nonexperts.

The act-frequency approach is of limited usefulness if the object is to generate items for traits for which no names exist in the natural language and in cases where people cannot be expected to have well-elaborated semantic intuitions. These conditions may apply to many of the traits that are of interest to psychologists studying abnormal behavior and psychopathology. How should an item pool be constructed, if procedures similar to those suggested by Buss and Craik (1983) cannot be followed?

More than 25 years ago, Loevinger (1957, p. 659) postulated that "at the very least, the items in the pool should be drawn from an area of content defined more broadly than the trait expected to be measured"; that is, some of the items should assess other related traits against which the trait of interest can be discriminated. Within an item pool for a particular trait, the items should be "chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait." Moreover, the various content domains "should be represented in proportion to their life-importance." These postulates were designed to ensure that evidence for the validity of a test or scale could also be interpreted as evidence for the validity of the underlying construct. Thus, Loevinger's recommendations are based on the assumption that researchers constructing questionnaire scales have already elaborated a nomological network in which the construct they wish to assess is embedded. However, many questionnaires were (and still are) constructed atheoretically, often for applied purposes (Goldberg, 1971). Consequently, Loevinger's (1957) recommendations have had little impact on the general practice of personality questionnaire construction.

In marked contrast to the ideals set out by Buss and Craik (1983), Jackson (1970), and Loevinger (1957), most authors of personality questionnaires have used combinations of the following three strategies to generate their items: (a) Constructing new items but leaving the rationale for item generation unexplicated (e.g., the items in Cattell's 16 PF, whose origin is largely unknown); (b) using experts to supply what they regard as typical trait manifestations (e.g., some MMPI items); and/or (c) copying items from other questionnaires (e.g., many of Guilford's GZTS items).

The pervasiveness of "borrowing" as a strategy for generating item pools has been illustrated by Goldberg (1971, p. 335), who surveyed the history of American personality scales and concluded that "items devised around the turn of the century may have worked their way via Woodworth's Personal Data Sheet, to Thurstone and Thurstone's Personality Schedule, hence to Bernreuter's Personality Inventory, and later to the Minnesota Multiphasic Personality Inventory, where they were borrowed for the California Personality Inventory and then injected into the Omnibus Personality Inventory – only to serve as a source of items for the new Academic Behavior Inventory" (and, we may add, only to be translated and included in some new German personality inventories). Angleitner and Löhr (1980) attempted to assess the amount of item overlap among German personality scales by item wording comparisons.

Quite appropriately, then, Loevinger (1957, p. 658) characterized "the process by which the given investigator or group of investigators constructs or selects items to represent the content" as *idiosyncratic* and therefore *nonreproducible*. A major stumbling block to improved personality scales may arise from these defects in the initial item pools, since the use of highly objective, empirical-statistical methods such as item factor analysis or item discriminance analysis to select items from such pools will not correct for the initial subjectivity; what was left out in the first step cannot be reinserted and then selected to become part of the resulting trait scales. As a consequence, the decision as to which of many possible trait labels will eventually be used as the name for the resulting scale will largely reflect the test author's idiosyncratic preconceptions of the nature of the trait he or she has tried to assess.

What Kind of Items Do We Find in Questionnaires?

A comparison of Buss and Craik's (1983) act frequency items with those of traditional personality questionnaires shows that many act items refer to single instances of fairly specific behaviors, whereas questionnaire items often include intensity and frequency adverbs, thus assessing more generalized aspects of the ways in which individuals experience, and respond to, their environments. Apparently, at least some questionnaire items assess individual differences at the level of behavioral trends or habits. In our sample of questionnaires, the most common frequency adverbs were sometimes *(manchmal)*, often *(oft)*, always *(immer)*, and never *(nie)*. The most frequently used intensity adverbs were very *(sehr)* and much *(viel)* (see Löhr, 1983a).

Most act-frequency items fall at roughly the same level of abstraction (i.e., specific behavioral acts), *because* Buss and Craik (e.g., 1983) made an explicit effort to eliminate "non-act" statements (e.g., trait attributions), general tendency and habit statements (e.g., "She tends to avoid parties"), and statements they considered too vague to constitute observable acts. The fact that only a subset of questionnaire items inquires about the respondents' acts suggests that personality questionnaire authors have not screened their items as carefully as Buss and Craik (1983) have done. In particular, critics suggest that the unsystematic construction of item pools has led to extremely heterogeneous mixtures of item types in contemporary personality questionnaires:

> The item sets of our questionnaires contain an odd mixture of questions of totally divergent logical and empirical status. Questions about concrete biographical data or about behaviors and experiences in specific situations appear next to items that ask the respondents to integrate behaviors and experiences that have arisen in specific or in completely vague situations across time and intensity, or to decide whether they behave in a particular way more frequently, or with higher intensity, than does some other, unspecified person. Besides that, the respondents are asked about their interests, preferences, and attitudes and requested to give evaluations of their own behaviors or those of others. Moreover, they are used as "trait construct-constructors" or "construct-validators" in such questions as "I am a timid, neurotic person." Items referring to "habits," "traits," and "types" stand side by side in one and the same questionnaire; particular behaviors and subaspects of just these behaviors are summed to one total score without hesitation. Some items ask for the frequency of behaviors;

others – listed in the same questionnaire – ask for the intensity of reactions. Yet other items represent weird mixtures of frequency and intensity aspects (e.g., I *sometimes* react *strongly* in certain situations). Furthermore, it is questionable whether responses to such items should be described by a cumulative or an increasing-probability item response model (Janke, 1973, p. 47, translated by the authors).

An Empirical Investigation of Item-Trait Relations

To assess whether Janke's characterization of the item pools of contemporary personality questionnaires is justified, a large set of items taken from various German-language personality questionnaires was classified. A second goal of this study was to find out whether differences in theoretical viewpoint and item selection strategies among test authors are related to differences in the kinds of items chosen as indicators of the trait. Finally, we wanted to investigate whether some types of items have better psychometric characteristics than others. This question is of great importance for the development of new questionnaires and will be addressed in the final section of this chapter.

The Category System. Guided by Janke's (1973 b) examples, Lennertz's (1973, pp. 59–60) description of consistent classes of logical item-trait relations, and our content analyses of some selected item sets, we developed a category system that would permit a systematic description of item-trait relations. An overview of this schema and some examples are given in Table 2. The central categories were: (a) Descriptions of *reactions*, which were further subdivided into *overt*, *covert*, and *bodily* reactions *(symptoms)*; and (b) *trait attributions* which could be either *unmodified* or *modified* by, for example, situational context specifications or frequency qualifiers.

In addition, there were other categories representing (c) *wishes and interests;* (d) *biographical facts;* (e) *attitudes and beliefs;* (f) *reactions of other persons;* and (g) *bizarre item content*. These categories refer predominantly to items whose content is rather indirectly related to personality traits; that is, the evidential link between the item and the trait is more tenuous for these "indirect" item categories than for the first two categories. This distinction might be related to inter-item differences in response stability; as Goldberg (1963, p. 482) has pointed out, "the more an item reflects some aspect of behavior that is directly observable and easily identifiable the more stable the item; on the other hand, the more an item reflects some attitude, value, or other 'internal' state of mind the more unstable is the item."

Theory. The frequencies of these item types in questionnaires may vary with the theoretical orientation of the test author, the scale construction strategy applied, and the trait being assessed. Cattell (1957), for example, interprets his personality factors as *source* traits, which are thought to have a pervasive influence on every aspect of a person's way of experiencing and responding to the environment. Given this theoretical perspective, one might expect to find

Table 2. Item-trait relations: A category system

1. Descriptions of reactions
   This category contains items that are intended to assess
   a) *Overt* reactions of the test person that are – in principle – publicly observable (e.g., "I often go to parties");
   b) *Covert* or internal reactions that are private and generally not observable by others, such as internal sensations, feelings, and cognitions (e.g., "I think a lot about myself") and
   c) *Symptoms*, that is, physical reactions (e.g., "I sweat a lot")
2. Trait attributions
   Items classified in this category refer to dispositions of the test person. The traits attributed to the test person are usually expressed by adjectives and nouns (e.g., "I have good acting abilities"). These items are further subclassified in regard to whether they are unmodified or modified. Frequent modifications are specifications of frequency, duration, and situational contexts
3. Wishes and interests
   These items refer to the intention to engage in particular behaviors or the desire for something (e.g., "Sometimes I would really like to curse"). *Not* included here are those items that mention wishes and desires but specify that they are realized in actual behavior (e.g., "I like to spend my free time by myself")
4. Biographical facts
   This category contains items that focus on aspects of the past (e.g., "I had some trouble with the law when I was younger")
5. Attitudes and beliefs
   The items of this category refer to strongly held beliefs, opinions and attitudes about various kinds of general, social, and personal issues (e.g., "I think the law should be strictly enforced")
6. Others' reactions to the person
   This category subsumes items that describe behaviors and attitudes of other people toward the person (e.g., "At parties, I am seldom the center of attention")
7. Bizarre items
   Most of these items describe clearly unusual, strange, or even abnormal behaviors and experiences (e.g., "Somebody is trying to poison me")

a large number of different types of item-trait relations in Cattell's questionnaires. In contrast, Eysenck (e.g., 1947) had been influenced by Hull's behavioristic learning theory and developed a hierarchical model of personality. *Specific responses* to situational cues, which form the lowest level in his hierarchy, can be mapped into broader categories called *habitual response patterns*, which in turn combine into concepts at the *trait* level, such as sociability, impulsiveness, or activity. Finally, at the *type* level one finds "supraordinate" concepts such as extraversion, which are based on the observed intercorrelations among traits. Thus, since Eysenck's (1947) model was built up entirely from behavioral responses, there was no room for indirect trait indicators; item categories such as wishes and interests or attitudes and beliefs should therefore not be represented in item sets developed by Eysenck and his co-

workers. Moreover, since Eysenck viewed specific responses as too unstable and unreliable to be useful for the assessment of personality characteristics, we expect his item sets to be drawn primarily from the second and third levels of his hierarchy (i.e., habitual responses and trait attributions) and therefore form fairly homogeneous sets of item types.

The test authors' preferences for different strategies of scale construction (see Burisch, this volume) by themselves provide only indirect clues as to the criteria they use in the construction of their initial item pools. For example, externally constructed questionnaires such as the MMPI may show highly heterogeneous mixtures of different types of item-trait relations. Finally, one can speculate that the intuitions of different test authors may converge on the kinds of items that are best suited for the assessment of a particular trait. For example, interpersonal traits such as sociability or dominance, which are often conceptualized in terms of observable social acts (e.g., Wiggins, 1979; Buss & Craik, 1983), might be assessed predominantly via items describing overt reactions.

Item Classification. The items of the German-language personality inventories used for adults were assembled in lists, each of which contained approximately 250 items. Each of 85 advanced psychology students, who were paid for their participation, was randomly assigned to one of the lists, and classified that item set according to the category system described above. Each item was categorized by at least nine judges.

To assess interjudge agreement, the frequencies of the seven main categories were computed for each item, and the category with the highest frequency was used as the key. The mean proportion of items that each judge had classified in agreement with the plurality of judges, averaged across the items of all ten adult questionnaires, was .79, whereas Kappa, a measure corrected for possible inflation by chance, was .61. For further analyses, each item was assigned to the category that had been selected by the majority of the judges. If none of the seven categories was chosen by more than 50% of the judges, the item apparently did not belong to just one category and was therefore left unclassified for the present purposes. Overall, only 8.7% of the items were not assigned to one of the seven main categories. The same set of procedures was used to assign items initially classified as trait attributions or reaction descriptions to the appropriate subcategories.

Results. In Table 3, the relative frequencies of the items in each category are reported for the item pools of the seven Anglo-American questionnaires included in this study; Table 4 presents the results separately for each scale. As Table 3 shows, the indirect item types were relatively infrequent (20%). The vast majority of items (63%) described various kinds of reactions, whereas less than 10% were trait attributions. The most frequent indirect reference categories were *wishes and interests* and *attitudes and beliefs*. The percentage of items referring to reactions of others and bizarre content was small.

Table 3. Relative frequencies of items in each category from the personality scales in seven questionnaires and the total set of 964 items

Questionnaire (no. of items)	Description of reactions				Trait attributions			Indirect references						Left unclassified	Number of main categories used
	1a Overt	1b Covert	1c Symptoms	Not subclassified	2a Unmod.	2b Mod.	Not Subclassified	3 Wishes	4 Biog.	5 Attds.	6 Others	7 Bizarre	Sum 3–7		
MPI (48)	18.8	41.7	2.1	8.2	18.8	4.2	2.0	0	0	0	0	0	0	4.2	2
MMQ (56)	19.6	23.2	17.9	8.9	10.7	7.1	7.2	0	1.8	0	0	0	1.8	3.6	3
EPI-A (57)	42.1	36.8	7.0	3.6	7.0	3.5	0	0	0	0	0	0	0	0	2
EPI-B (57)	24.6	22.8	14.0	12.3	7.0	5.3	1.7	0	1.8	0	0	0	1.8	10.5	3
16PF-A (171)	18.1	21.1	1.8	14.6	4.7	0	1.7	11.1	3.5	11.1	1.8	0	27.5	10.5	6
16PF-B (171)	17.5	24.0	3.5	10.6	1.2	.6	3.5	14.0	2.9	13.5	1.2	0	31.6	7.6	6
MMPI (404)	10.4	27.0	13.1	11.1	3.2	1.0	2.0	4.7	5.7	5.9	2.2	3.0	21.5	10.6	7
All items (964)	16.7	26.3	8.8	11.0	4.8	1.7	2.4	6.4	3.7	6.8	1.5	1.3	19.7	8.7	4.1

Unmod., unmodified; Mod., modified; Biog., biographical facts; Attds., attitudes and beliefs; Others, other's reactions to the subject.

Table 4. Relative frequencies of items in each category for the personality scales in seven Anglo-American questionnaires

Scales (no. of items)		Description of reactions				Trait attributions			Indirect references					Left completely unclassified	Number of main categories used
		1a Overt	1b Covert	1c Symptoms	Not subclassified	2a Unmod.	2b Mod.	Not subclassified	3 Wishes	4 Biog.	5 Attds.	6 Others	7 Bizarre		
MPI	E 24	37.5	12.5	0	8.3	29.2	4.2	0	0	0	0	0	0	8.3	2
	N 24	0	70.8	4.2	8.3	8.3	4.2	4.2	0	0	0	0	0	0	2
MMQ	N 38	5.3	23.7	26.3	5.2	15.9	7.9	10.5	0	2.6	0	0	0	2.6	3
	L 18	50.0	22.2	0	16.7	0	5.6	0	0	0	0	0	0	5.6	2
EPI-A	N 24	8.3	62.5	16.7	0	12.5	0	0	0	0	0	0	0	0	2
	E 24	62.5	16.7	0	8.3	4.2	8.3	0	0	0	0	0	0	0	2
	L 9	77.8	22.2	0	0	0	0	0	0	0	0	0	0	0	1
EPI-B	N 24	8.3	41.7	33.3	4.2	4.2	8.3	0	0	0	0	0	0	0	2
	E 24	37.5	12.5	0	20.8	8.3	4.2	4.2	0	0	0	0	0	12.5	2
	L 9	33.3	0	0	11.1	11.1	0	0	0	11.1	0	0	0	33.3	3
16PF-A	A 10	30.0	0	0	0	0	0	0	60.0	0	10.0	0	0	0	3
	C 13	15.4	30.8	7.7	15.3	0	0	0	7.7	0	0	15.4	0	7.7	3
	E 13	30.8	15.4	0	23.0	15.4	0	7.7	0	0	0	0	0	7.7	2
	F 13	30.8	0	0	0	7.7	0	7.7	7.7	0	7.7	0	0	39.5	4
	G 10	10.0	10.0	0	10.0	10.0	0	0	0	0	30.0	0	0	30.0	3
	H 13	30.8	23.1	0	23.0	15.4	0	0	7.7	0	0	0	0	0	3
	I 10	20.0	0	0	10.0	0	0	0	40.0	20.0	0	0	0	10.0	3
	L 10	20.0	30.0	0	20.0	0	0	0	0	0	20.0	10.0	0	0	3
	M 13	7.7	15.4	0	30.7	0	0	0	15.4	0	15.4	0	0	15.4	3
	N 10	20.0	10.0	0	10.0	0	0	0	0	0	30.0	0	0	30.0	2
	O 13	0	46.2	7.7	0	15.4	0	0	0	7.7	7.7	0	0	15.4	4
	Q1 10	0	36.0	0	10.0	0	0	10.0	10.0	10.0	30.0	0	0	0	5
	Q2 10	10.0	10.0	0	20.0	0	0	0	20.0	20.0	10.0	0	0	0	4
	Q3 10	30.0	10.0	0	40.0	0	0	0	0	0	20.0	0	0	0	2
	Q4 13	15.4	69.2	7.7	7.7	0	0	0	0	0	0	0	0	0	1

16PF-B	A	10	10.0	10.0	0	0	0	0	0	80.0	0	0	0	0	0	2
	C	13	15.4	38.5	0	15.3	0	0	0	15.4	7.7	7.7	0	0	0	4
	E	13	15.4	15.4	0	7.7	7.7	7.7	0	0	15.4	15.4	0	0	15.4	4
	F	13	23.1	15.4	0	7.7	7.7	0	15.4	15.4	0	7.7	0	0	7.7	4
	G	10	40.0	20.0	0	0	0	0	0	0	0	40.0	0	0	0	2
	H	13	30.8	30.8	0	23.0	0	0	7.7	0	0	0	0	0	7.7	2
	I	10	0	40.0	0	10.0	0	0	10.0	40.0	0	0	0	0	0	3
	L	10	10.0	30.0	0	0	0	0	0	0	10.0	30.0	10.0	0	10.0	4
	M	13	15.4	7.7	7.7	0	0	0	7.7	15.4	7.7	30.8	0	0	7.0	5
	N	10	20.0	30.0	0	10.0	0	0	0	20.0	0	20.0	0	0	0	3
	Q	13	7.0	61.5	7.7	7.7	0	0	0	0	0	7.7	0	0	7.7	2
	Q1	10	30.0	0	0	10.0	0	0	0	10.0	0	20.0	0	0	30.0	3
	Q2	10	10.0	10.0	10.0	10.0	0	0	1.0	0	0	20.0	10.0	0	20.0	4
	Q3	10	20.0	10.0	0	30.0	0	0	0	20.0	0	10.0	0	0	10.0	3
	Q4	13	15.4	30.0	23.0	23.0	0	0	0	7.7	0	0	0	0	0	2
MMPI	L	15	46.7	20.0	0	13.3	0	0	6.7	13.3	0	0	0	0	0	3
	F	64	7.8	12.5	9.4	4.7	3.1	1.6	0	1.6	9.4	12.5	3.1	17.2	17.2	7
	K	30	6.7	43.3	0	16.7	0	0	0	6.7	0	16.7	0	0	10.0	3
	Hs	33	0	0	87.9	3.0	3.0	3.0	3.1	0	0	0	0	0	0	2
	D	60	16.7	28.3	20.0	5.0	6.7	5.0	6.6	5.0	0	3.3	0	0	3.3	4
	Hy	60	8.3	23.3	35.0	13.4	1.7	3.3	1.7	3.3	0	8.3	0	0	1.7	4
	Pd	50	8.0	38.0	2.0	8.0	0	2.0	0	4.0	12.0	2.0	8.0	0	16.0	6
	Mf	60	13.3	23.3	1.7	3.4	5.0	0	0	21.7	8.3	6.7	0	0	16.7	5
	Pa	40	2.5	22.5	5.0	12.5	5.0	2.5	0	0	5.0	15.0	5.0	15.0	10.0	6
	Pt	48	4.2	52.1	6.3	14.5	4.2	2.1	8.3	2.1	2.1	0	0	0	4.2	4
	Sc	78	6.4	30.8	12.8	12.8	0	0	2.6	1.3	7.7	0	5.1	3.8	16.7	6
	Ma	46	13.0	39.1	4.3	21.9	2.2	0	0	0	6.5	6.5	2.2	0	4.3	5
	Si	70	17.1	25.7	5.7	17.2	4.3	0	2.8	5.7	5.7	5.7	0	0	10.0	5
	Sd	39	5.1	38.5	17.9	7.7	5.1	5.1	2.6	2.6	2.6	2.6	5.1	0	5.1	6

Abbreviations as for Table 3

The four questionnaires published by Eysenck (i.e., MPI, MMQ, and EPI Forms A and B) were found to consist almost exclusively of items that require the respondent to provide reaction descriptions and trait attributions. Thus, the item types identified in Eysenck's questionnaires formed a homogeneous set consistent with his hierarchical model of personality. The scale-by-scale analyses (see Table 4) showed that Eysenck's Extraversion and Lie items were categorized primarily as *overt* reactions, whereas *covert* reactions and *physical symptoms* were the most frequent item categories found in his Neuroticism scales. A comparison of the distributions of category frequencies in EPI Forms A and B reveals that supposedly parallel scales were not entirely parallel with regard to the types of items used to assess the same construct. This is particularly true of the two Lie scales, which also have very low parallel test reliabilities (see Table 9).

Whereas only two or three categories were found in Eysenck's questionnaires, Cattell used a rather heterogeneous mixture of different types of item-trait relations in both forms of his 16 PF; of the seven main categories, six were represented in Cattell's item sets. Overall, about 60% of these items involved self-descriptions of reactions (trait attributions being almost completely absent), whereas indirect references were much more frequent in Cattell's than in Eysenck's questionnaires. Specifically, the 16 PF factors *A* and *I* (which supposedly assess warmth and tendermindedness) contained high percentages of items expressing wishes and interests. Since the parallel scales in Forms A and B of the 16 PF different considerably in their category frequency distributions, the scale-by-scale analysis presented in Table 4 showed only a few trends for scales with the same label. In both forms, factor scales *C*, *O*, and *Q4*, which often load a second-order factor labeled Anxiety (Amelang, Sommer, & Bartussek, 1971; Gorsuch & Cattell, 1967; Timm, 1968), were characterized mostly by items that describe covert reactions. Similar to Eysenck's extraversion scales, factors *F* and *H* which form the second-order factor Exvia-Invia had many items referring to overt reactions (approximately 30%), but also contained items from several other categories. Finally, factor *G* (superego strength) was assessed predominantly by attitudes and beliefs (as well as some other item types), whereas the items of factor *N* (sophistication) represented a mixture of attitudes and beliefs, overt reactions, and covert reactions.

The MMPI clearly showed the most heterogeneous mixture of item types of all seven questionnaires analyzed. Apparently, Janke's (1973) description of "item hodge-podges" characterizes the MMPI item pool quite appropriately, even if the scales are considered individually. As the last column of Table 4 shows, the typical MMPI scale contained items from *five* different categories, whereas on average only two main item categories were found in the various Eysenck scales. In comparison with Cattell's and Eysenck's questionnaires, the MMPI had more items referring to physical symptoms. Given that the MMPI is primarily a clinical assessment instrument, this is not really surprising; however, although all clinical scales including Social Introversion (Si) had at least a few physical symptom items, most of these items were concentrated in the

Hypochondriasis (Hs) and, to a lesser extent, the Hysteria (Hy) scales. The modal category for all other clinical scales was that called *covert reactions.*

The MMPI was the only questionnaire that had bizarre items; they appeared in those scales where we would expect them – Paranoia (Pa) and Schizophrenia (Sc). And, again as might be expected, it was the validity scale *F* that contained the highest percentage of bizarre items. Finally, the MMPI Lie scale (*L*) was found to be similar to those developed by Eysenck in that the respondent's tendency to lie was assessed predominantly with items describing overt reactions, whereas the items of the correction scale *K* referred primarily to covert reactions.

Discussion. The finding that the items of most MMPI scales form much more heterogeneous sets of item-trait relations than Eysenck's scales might in part be attributed to the fact that the MMPI scales are generally longer. However, since the effect is still observed when the total item pools are compared (see Table 3), it seems possible that differences between scale construction strategies influence the heterogeneity of the item pools found in different questionnaires. In particular, the fact that the empirical external strategy, which was used to construct the MMPI scales, neglects both item content and scale homogeneity may result in a rather diverse set of different item types within the same scale. In contrast, the heavy reliance on item analysis typical of internal scale construction strategies may constrain the number of different item types.

The finding that the rather homogeneous item sets of Eysenck's questionnaires were consistent with his theory demonstrates how helpful the mere existence of a theoretical model can be for the generation of relevant types of trait indicators. Nevertheless, equal weighting of responses to different item types, such as habitual reactions and direct trait attributions, in the calculation of scale scores seems inconsistent with Eysenck's hierarchical model, which implies that habitual responses belong to a lower level of abstraction and are therefore more specific.

Heterogeneity of item types within a questionnaire scale is not necessarily undesirable in itself – as long as the nature of this heterogeneity is theoretically explicated and justified. For example, one may have a theory of assertiveness suggesting that highly assertive people report (a) engaging in observable assertive acts; (b) thinking a lot about acting assertively; and (c) having particular physical symptoms when deprived of the opportunity to act assertively. If, in addition, the theory states that such people (d) describe themselves as assertive, bold, and not timid; (e) express wishes and interest related to assertiveness; (f) report that they stood up for themselves at nursery school; (g) state that they admire assertive people and endorse the belief that assertiveness is necessary to make it in this world; and finally, (i) indicate that others have reacted to their self-assertion in various ways, then (and only then) could the use of eight different item types to assess assertiveness be justified. Using sets of items selected to cover each of these types (or facets – see Kastner, this vol-

ume), the researcher has to show that the theory is correct, that is, that the responses to all these different types of items are manifestations of the same underlying trait.

In all, the distributions of the item categories *within* trait scales are quite informative. The results suggest that scales that assess similar constructs tend to be measured by somewhat similar types of items, regardless of the many other differences among the questionnaires considered. All else being equal, Neuroticism (or Anxiety) scales were found to contain more items describing covert and physical reactions, whereas Extraversion and Lie scales were characterized by higher percentages of items referring to overt reactions. One wonders why these shared intuitions about what types of items are particularly suitable to measure certain kinds of traits have not yet been systematized and incorporated into a priori construct definitions.

One possibility for changing this unsatisfactory state of affairs can be derived from Buss and Craik's (1983) research on act frequencies and from other approaches that similarly view traits as categorical, prototype-based concepts (e.g., Hampson, 1982). Although Buss and Craik (1983) have concentrated their initial efforts on *interpersonal* traits (which, as the present results for the Extraversion scales suggest, tend to be conceptualized in terms of overt behavioral reactions), it is possible to generate prototypical instances of other classes of behavior as well. For example, Ashton and Goldberg (1973) and Jackson (1975) had fairly naive item writers generate items for traits that involve task-performance (achievement), values (tolerance), and self-related cognitions (self-esteem); the resulting scales had criterion validities superior to those of like-named scales from the empirically derived CPI (Gough, 1957).

The trait indicators obtained in such item generation studies could then be classified into different types, using a category system like the one proposed here. Thus, this approach would not only provide preliminary item pools, but also survey the item types relevant to the traits under investigation. In turn, this information can be used to elaborate and refine the construct definition (e.g., in terms of facets) and thereby help define central aspects of the item universe, from which further indicators of the trait construct could then be drawn.

These procedures, although far from perfect, would at least introduce some objectivity into the process of item pool generation. Compared with previous practices, which have often led to idiosyncratic, nonreproducible, and heterogeneous item pools, these procedures have the advantage of being explicit, quantifiable, reproducible, and open to evaluation in terms of intersubjective agreement. Even more important, they point to a way of satisfying the often-repeated postulate that these aspects of substantive validity (Loevinger, 1957) or content validity (Cronbach, 1970, p. 143) must be incorporated as an explicit part of the definition of the trait construct.

Situational Context in Questionnaire Items

We have already noted that many questionnaire items are modified by intensity and frequency qualifiers (Payne, 1970; Pepper, 1981); items requiring trait attributions sometimes specify at least some minimal situational context. Similarly, items such as "I feel anxious *when I speak in front of a group*" and "It takes me time to overcome my shyness *in new situations*" demonstrate that authors of personality inventories apparently felt that some situational context should be mentioned in the item. However, with the exception of a few S-R (stimulus-response) self-report inventories (see Knudson & Golding, 1974), authors of personality questionnaires have ignored the much-debated issue of situational specificity in human behavior (e.g., Mischel & Peake, 1982) when writing items for their scales. Test authors act as if they believe that the psychological meaning of situational characteristics can be captured by a conditional, temporal, or adverbial clause. Can it be taken for granted that a situational context presented in such a verbal form is adequate, representative, or sufficiently typical to permit a behavior-in-context rating? Also, is it sensible to attempt this within questionnaire items that typically consist of only one sentence? How indicative of the postulated trait is a behavioral act within a situation conceptualized in such a contrived way? These questions need to be investigated systematically before we can adequately construct a trait scale.

In recent years, psychologists have paid increasing attention to the characteristics of situations, and some taxonomies of situations (Magnusson, 1981, 1984; Van Heck, 1984) and of types of persons *within* situations (Cantor, 1981; Cantor, Mischel, & Schwartz, 1982 a, 1982 b) have been constructed. Thus, we believe that theory and methodology have advanced to a point where we can find empirical answers to the questions posed above. Before we develop new questionnaires, then, we should investigate the psychological characteristics of the settings in which the behaviors of interest typically occur. Only after that step can we construct items that assess the postulated trait across a representative sample of relevant situations. And, finally, we may need to find ways of describing these situations in a verbal form that will enable respondents to understand the kinds of situations to which the item refers.

Traits in Ordinary and Scientific Language

Whereas the concept of "life-importance" has not been applied to the construction of item pools, it has been considered in decisions about which of the extremely large number of *traits* should be selected for assessment (Goldberg, 1972). In general, sheer historical accident and societal pressures on psychologists to forecast significant personal outcomes have been more important than theoretical models of personality in determining the selection of traits measured by questionnaires (Goldberg, 1971). However, a few attempts have been made to base the decision about which traits to assess on empirical criteria, rather than on purely applied considerations or on subjective "armchair" theories. For example, life importance has been operationalized in terms of the

frequency of a trait descriptor in spoken or written discourse about people (Goldberg, 1982; Gough, 1965), the number of synonymous expressions a language contains for the same trait (Cattell, 1943, 1946), or the universality of equivalent descriptive terms for the same trait across languages and cultures (Goldberg, 1981a; John, Goldberg, & Angleitner, 1984). This kind of reasoning led Cattell (1943, 1946) to search Allport and Odbert's (1936) compilation of English person-descriptive terms (as well as relevant publications of personologists) for trait terms to be included in his "trait sphere," from which he derived some of the dimensions later measured by the 16PF. Similarly, Gough's (1957) construction of the CPI started with an informal analysis of "folk concepts" of personality, which he defined as "variables used for the description and analysis of personality in everyday life and in social interaction" (p. 295). Gough (1965) viewed such folk concepts as emerging directly from interpersonal behavior, and assumed that they have a "kind of immediate meaningfulness" and "universal relevance."

One implication of the view that the trait terms found in many natural languages provide convenient summary labels for individual differences in behavior is that we need a theory about how trait terms are used in everyday life. Moreover, some items in personality questionnaires require subjects to make trait attributions to themselves; responses to such items are taken as evidence for the presence or absence of a trait in a person without much theoretical reflection about the processes by which people arrive at such attributions. Finally, we seem to use trait concepts in both ordinary and scientific language contexts without making a clear distinction between the two.

When we ask our subjects in questionnaire items whether they see themselves as lively, outgoing, impulsive, or sociable, they will unquestionably answer according to their intuitive, everyday-language understanding of these terms. However, when we refer to Extraversion in a talk delivered at a conference, we assume that the audience has a fairly uniform understanding of this term. Its meaning is supposedly embedded in theoretical and empirical relations so that the hearer can distinguish between, for example, Eysenck's and Guilford's concepts of Extraversion when these names are mentioned. But, where is the transition between these two levels of language? Does the very labeling of a factor according to its marker items produce a scientific term? Or is the mere listing of short paraphrases of the high-loading items still part of everyday language, whereas a more abstract factor label, which is related to a relevant theoretical network, implies the existence of a scientific terminology? What status should then be assigned to those trait terms that are administered as rating scales to subjects and their peers to validate questionnaire factors?

Indeed, Cattell (1973) made an explicit effort "to avoid the pollution of meaning in scientific discussion" (p. 54) by inventing a "host of neologic gobbledegook" (Goldberg, 1980) for his structural dimensions. The problem with his approach is that it has not worked. Personality psychologists, like everybody else, communicate via the natural language. Moreover, personality as-

sessment via questionnaires involves verbal communication among the assessor, the respondents, and the user, who may be a counselor or personnel manager with little knowledge of scientific terminology. Thus, to *really* understand what Cattell means by *Premsia, Harria,* and *Zeppia,* we must somehow manage to translate those terms back into their natural equivalents. Indeed, if those new terms turned out to be particularly useful in communicating about individual differences, they would soon become encoded in the natural language, just like Extraversion-Introversion (Goldberg, 1980).

It appears, then, that dealing with the natural languages is inevitable when personality is being assessed with a method so language-bound as a questionnaire. As a consequence, our scientific terms should be translatable into the everyday language of personality. If we want to use ordinary-language utterances as a source of psychological data (Fiske, 1981), we need an in-depth understanding of everyday-language *usage,* including an appreciation of the lack of precision and the context dependency of natural languages, as well as their power in human communication.

Thus, an important topic for further research is the usage of trait terms in self and peer descriptions. Some initial steps toward the development of semantic taxonomies for various types of personality-descriptive terms have been made in some West European languages (for a recent review, see John et al., 1984). When such taxonomies are mapped into recently developed situation taxonomies (e.g., Van Heck, 1984) and complemented with research on episode cognition (Forgas, 1982) and trait attribution (e.g., Kelley & Michela, 1980), our understanding of the differences and similarities between scientific and lay conceptions of traits will be much advanced.

Conclusions

There seems to be growing agreement among personality researchers of varying persuasions that the writing and selection of items should be preceded by an initial construction stage that involves some reflection about the trait construct and about the characteristics of relevant items (Fiske, 1971; Goldberg & Slovic, 1967; Jackson, 1971; Meehl, 1972; Mischel, 1972; Wiggins, 1973). Four issues that arise during this initial phase have direct implications for the practice of test development. First, rather than constructing measures for any odd trait and thereby contributing to the seemingly infinite proliferation of personality scales, the test developer should have a strong rationale for measuring a particular trait. Optimally, this rationale would be based on a structural theory of personality that reflects the "life-importance" of individual differences among people. Second, rather than "borrowing" items from other questionnaires or making up items on the spur of a creative moment, the test developer should begin the item-generation process by developing an explicit definition of the trait construct. This definition should include a description of the convergent and discriminant relations with other constructs and specify relevant types of trait manifestations, such as observable behavioral

acts, trait attributions to oneself, or wishes and interests. Third, to avoid idiosyncratic and nonreproducible item pools (and consequently trait scales), exemplars for each of the theoretically derived item types (or facets) should, whenever possible, be generated by a sizable sample of people; the selection of items from these initial sets should be based on interjudge agreement on the relevance of the item content for the trait (e.g., estimates of endorsement probabilities, prototypicality), and discriminant item validity should already be taken into account at this stage. Finally, since the meaning of social behavior depends to a large extent on the context in which it is performed, the test developer should, when screening the initially generated item material, make a special effort to ensure that the relevant aspects of the situational context are included in the item text and that they are described as explicitly as possible.

Surface Elements of Questionnaire Items: A Linguistic Analysis

Following these decisions about the trait construct and about relevant item types, the next task of the test developer is to "translate" the theoretically elaborated trait manifestations into an appropriate verbal form. But, what is the most appropriate form to make an item a good indicator of the trait construct? How *should* items be written? At this stage of test construction, detailed knowledge about language would be of great help to questionnaire developers.

Responses to personality questionnaires are highly standardized verbal behaviors elicited by verbal stimuli and should, therefore, be of interest to psycholinguists and linguists. However, linguistic analyses of questionnaires and their items have been rare. Despite recent progress in research on text and discourse processing (Givon, 1979), these approaches are not yet sufficiently developed to provide sophisticated theories that could be applied to questionnaires. Indeed, little research by either linguists or personality psychologists has been directed at the investigation of syntactic and semantic characteristics of the sentences that are used in personality questionnaires.

A small number of studies have investigated the effects of syntactic characteristics of the item formulation on item response consistency. For example, Micklin and Durbin (1969) reported that larger differences between the structural properties of the sentences in otherwise parallel attitude scales were associated with less response consistency to such scales. Other studies have compared various response formats differing in the number of possible response alternatives (Jones, 1968), or analyzed how "?" or "middle" responses are used by subjects (Dubois & Burnes, 1975). Goldberg (1978, 1981 b) recently investigated the meaning of "middle" responses in a trait attribution paradigm. He argued that subjects could interpret the middle response option in four rather different ways. In particular, a middle response to an item can stand for (a) a situational attribution (e.g., "my behavior depends on the situation"); (b) an expression of uncertainty (e.g., "I do not know that aspect of myself well enough to make a decision"); (c) ambiguity of the item (e.g., "I am not sure

what this item means"); and (d) neutrality (e.g., "I am average on this characteristic"). However, in the few questionnaires that offer the respondent a third response option (e.g., Cattell's 16PF and Eysenck's MPI), the middle responses are always counted as at least somewhat indicative of the trait and therefore increase the subject's total score.

Subjects frequently experience the dichotomous (sometimes trichotomous) response format as logically inconsistent with the item content. In our studies, subjects complained particularly about items that consisted of several separate clauses conjoined by *and* or *or* [e.g., MMPI item (131): "I like collecting flowers or growing indoor plants"] or contained unusual contextual specifications [e.g., MMPI item (171): "It makes me uncomfortable to put on a stunt at a party even when others are doing the same sort of things"]. Negations can also create considerable confusion, especially when a negated item does not apply, and thus the respondent should answer "false" or "no" (Peterson & Peterson, 1976).

Thus, it seems plausible that the wording of an item, as captured by surface characteristics such as response format, item length, and syntactic complexity of the sentence, should be related to the way the item is processed by the subjects who are trying to respond to it. Yet correlations between syntactic item characteristics and item response parameters are rarely substantial (Goldberg, 1968; Wiggins & Goldberg, 1965). One reason for the lack of more impressive empirical effects may lie in the fact that only a small number of relevant variables (often selected on an ad hoc basis) have been investigated in any one study. To overcome this deficiency, Löhr and Angleitner (1980) studied a more comprehensive set of precisely defined variables across a large sample of questionnaire items. A theoretical framework for the description of the surface structure of items will be introduced here and used to discuss their findings. The relations between item surface characteristics and item response reliability and validity will be presented in the fourth section of this chapter.

Linguistic Description of Item Surface Characteristics

According to psycholinguistic models of reading (see Foss & Hakes, 1978), the ease with which subjects read a sentence depends on the extent to which short-term memory is overloaded, and also on the structure of the sentence. Thus, errors during the initial stages of the item response process (i.e., reading and subsequent encoding) should increase with the length and complexity of the item and lead to low response stability. Findings by Wiggins and Goldberg (1965) support this reasoning; the correlation between item length and response consistency from first to second administration was $-.30$. To replicate and further explore this finding, Löhr and Angleitner (1980) devised three objective measures of item length, specifically, the numbers of words, of letters, and of clauses.

Although Chomsky (e.g., 1965) has stated that his theory of language cannot be directly interpreted as a theory of language *performance*, many models

of comprehension incorporate some aspects of transformational grammar (Foss & Hakes, 1978). According to such models, each transformation in the grammar has an associated "reverse" transformation, that is, a set of operations which "detransform" the surface structure of a sentence into a semantic representation at the deep-structure level. In particular, these reverse transformations apply to the surface structure, which corresponds to the input string of words that make up the sentence. One implication of this type of model is that the comprehension of sentences that undergo more transformations should take longer and result in more errors than those that undergo fewer transformations. For example, a passive sentence should be more difficult to comprehend than the corresponding active one, since the inverse of the passive transformation has to be applied to it. Helferich (this volume) discusses some of the problems inherent in these models.

In addition to the three measures of item length, Löhr and Angleitner (1980) scored the number of passive verbs and syntactic negations as well as the tense (i.e., past, present, future) and the mood (i.e., indicative vs subjunctive) of the verb of the main clause. These measures seem to reflect the transformational complexity of the item's surface structure. The presence of personal pronouns such as *you* or *I* in the item's surface structure is another variable that could be viewed as an aspect of transformational complexity. That is, personality questionnaire items inquire about self-perceptions and retrospective accounts of self-related events, which seem to be represented within a part of semantic memory that is centered around the respondent's self (Hull & Levy, 1979; Rogers, 1977). Items whose surface structures already contain direct references to the respondent's self may require fewer transformations, and may therefore be more easily compared with self-relevant information already stored, than are items that do not contain any personal references.

Two other surface characteristics of questionnaire items that are beyond the reach of psycholinguistic theorizing are of a pragmatic nature, that is, they are specific to the particular format and communicative purpose of personality questionnaires: The respondents are expected to provide the researcher with information about themselves, but in order to do so they have to use the highly restricted and standardized response alternatives specified by the researcher. To elicit the desired information from the respondent, questionnaire authors have used a variety of response formats. Specifically, they have used as sentence types either direct questions or assertions, followed by response formats that differ in the number of response alternatives they permit. Several kinds of response formats can be distinguished; some require the same general response types to all questions (e.g., yes vs no or true vs false) whereas others consist of specifications of the item content (e.g., often–sometimes–never).

Surface Characteristics of Items in Contemporary Questionnaires

In Löhr and Angleitner's (1980) study, two judges scored 1624 items assembled from ten personality inventories on eleven variables, which can be grouped ac-

cording to three aspects of the item's surface structure: Item *length* as indicated by number of words, letters, and clauses; item *complexity* assessed in terms of negations, passive voice, tense, mood, and personal reference; and finally, the item *format*, including the sentence type as well as number and kind of response options. As one should expect of such objectively defined measures, interjudge agreement was very high; overall, the two judges agreed on 98% of their ratings.

The results of this study show that many questionnaire authors do not even follow those guidelines that the relevant textbooks list as rules of thumb for item writing. Lienert (1969, pp. 63–64), for example, recommends keeping items short, and avoiding multiple clauses and negations. Yet, the average item in the set of 964 items compiled from seven Anglo-American questionnaires (see *Appendix A* for a list of the scales analyzed) was more than 12 words long (SD = 6.2) and consisted of two clauses (SD = 1.1). Moreover, almost a quarter of these items contained at least one negation. Double negations and passive verbs, however, were relatively infrequent (2% and 5%, respectively). Use of the past tense and subjunctive were found in 10.5% of the items; 4.5% contained *no* personal reference to the respondent. To evaluate these findings within a broader context of language use, comparison data for sentences found in other sources could be helpful. Thus, it might be interesting to investigate whether the text form "personality questionnaire" differs in some important ways from other text forms such as newspaper articles, autobiographies, and diaries, and whether such differences in surface structure reflect the particular communicative purposes of questionnaires.

There were some systematic differences in the items' surface characteristics among the seven questionnaires of Anglo-American origin included in this project. Cattell's 16PF items consistently had the highest scores on all three measures of item length. Whereas the items of the four Eysenck questionnaires (i.e., MPI, MMQ, EPI-A and -B) and the MMPI items averaged ten words and consisted of less than two clauses, the 16PF items had a mean length of 16 words and 2.5 clauses.

Comparisons of these questionnaires on the syntactic characteristics suggest that there is some correspondence between the logical relation of the item to the trait and the way it is phrased. That is, the surface structure of the item seems to reflect shared conventions about the phrasing of particular types of trait indicators. The most obvious example is the MMPI, which of all seven questionnaires contained the highest percentage of biographical items, and correspondingly had the highest percentage of items with the main verb in the past tense. More importantly, Löhr and Angleitner (1980) found that, compared with the other questionnaires, both forms of the 16PF had the highest percentages of items with subjunctive verb constructions and no direct references to the respondent. This finding reflects the much higher frequencies of inquiries about wishes and interests and about attitudes and beliefs in Cattell's item sets. Apparently, wishes and interests are often expressed as actions whose *hypothetical* nature is indicated by subjunctive verb forms, whereas at-

titudes and beliefs seem to be assessed predominantly in terms of categorical statements of a *general* nature, and therefore they may lack personal references in their surface structures.

Although we need to follow up these suggestive findings more systematically, it is evident that syntactic aspects of questionnaire items cannot be conveniently separated from semantic considerations (Fillmore, 1977). Indeed, our findings illustrate the complex interaction between the characteristics of trait constructs and the syntactic aspects of the items that psychologists construct to measure them. We hope that more rigorous theorizing and research in psycholinguistics will eventually increase our limited understanding of these intricate relations.

Item Characteristics and Item Response Processes

As Table 1 shows, the output of the researcher's item-writing efforts serves as the raw verbal input to the subjects who read the item and try to respond to it. Thus, whereas we took a linguistic perspective in the previous section and described and evaluated the syntactic surface characteristics of the item text with reference to the researcher's task of item *writing*, we will now discuss item characteristics that are relevant to the subject's task in *responding* to the item. Our perspective is that of cognitive psychology, and we will describe the cognitive stages and operations that need to be performed to arrive at a response. For each stage of this response process, we will identify the item characteristics that can be expected to facilitate or complicate the processing of the item, define, and measure them in a rating study, and finally, in the last section, use these measures to predict item response parameters such as item stability and validity.

Stages of the Response Process and Item Characteristics

Several different models have been proposed to describe the processes involved in responding to personality questionnaire items (e.g., Cliff, Bradley, & Girard, 1973; Kuncel, 1973; Nowakowska, 1970; Rogers, 1971, 1974a, 1974b; Schneider-Dueker & Schneider, 1977). Although these authors differ considerably in their methodological approaches (e.g., content analyses of thinking-aloud protocols, reaction time studies, pairwise ratings of the similarity between items), their models show a modest degree of convergence. Most of these models can be classified as three-stage, sequential processing models. When responding to a questionnaire item, the subjects first read the item and form an internal representation of its meaning (encoding stage). Second, they compare that meaning with internally stored information about themselves, and decide whether they agree with the item or not (item-self comparison stage). Before the response is marked on the answer sheet, the latent decision may also be checked for its utility, especially with regard to its congruence with social norms and values (utility-control stage).

Although Table 1 shows these hypothetical stages of the response process as if they were distinct and strictly sequential (see Rogers, 1974a), we do not claim that the operations implied by these processing stages are completely independent of each other or that they could not be executed in an interactive or parallel fashion. Rather, we view this model as an heuristic one that can help identify, for each stage of the response process, those item characteristics that might elicit processing difficulties at that stage. In particular, we will consider the following five questions: (a) Is the item difficult or easy to understand? (b) Is it possible to assign more than one meaning to the item, or is it unambiguous? (c) Does the item involve abstract information that may require additional complex processing or does it require only concrete information that is

Table 5. The rating scales used to assess the five response characteristics

Comprehensibility

1. You understand the item immediately upon first reading
2. After the first reading, you have to read at least parts of the item again
3. After the first reading, you find that at least one "aspect" is still completely unclear
4. You have to consciously apply grammatical rules to understand the item or read it at least three times

Ambiguity

1. The item is completely unambiguous
2. Only if you try to be very nitpicky can you discover a slight ambiguity, which is, however, of no practical consequence
3. If one reads the item carefullly and thinks about the possible answers, it is clear that the item cannot be answered unambiguously
4. The item is obviously ambiguous

Abstractness

1. The item is concrete in every respect
2. The important aspects of the item are worded in a concrete way; the overall impression is concrete
3. The important aspects of the item are worded in an abstract way; the overall impression is abstract
4. The item has almost no concrete reference

Self-Reference

1. The item refers strongly to personal experiences
2. The item refers predominantly to personal experiences. The respondent is the focal point of the statement
3. The respondent is mentioned in the item but is not the focal point of the statement
4. The respondent is either not mentioned at all in the item or "stands completely at the sidelines"

Evaluation

1. This item is as neutral when it is endorsed as when it is not endorsed
2. The item implies social norms; however, the individual's self-concept weighs more heavily in deciding how to respond
3. When responding to the item, one is likely to follow general value standards
4. The item expresses a clear positively or negatively valued standard

readily available and can be easily retrieved from memory? (d) Does the item refer to personally relevant information? (e) Is the item evaluatively neutral or is it socially desirable or undesirable to endorse the item? Corresponding to these five questions, five item characteristics were defined, and verbally anchored four-step rating scales were constructed to measure each of them. These rating scales are presented in Table 5.

Comprehensibility. Item comprehensibility is of obvious importance for the response process. The more difficult it is for subjects to comprehend the meaning of the item, the more likely it is that the item will be misunderstood, with the result that the response will not bear on the subject's real position. The importance of this item characteristic is reflected in various models of the response process, such as in Rogers' (1974a) stimulus encoding and Kuncel's (1973) meaning assignment stages. Causes of insufficient comprehensibility are uncommon words, complex sentence structures, unclear and confusing relations among clauses within sentences, and grammatical errors. In an attempt to construct a rating scale anchored in subjects' subjective reactions to items, each of the four steps of the scale were clearly labeled (see Table 5). Subjects were warned that comprehensibility and ambiguity are distinct concepts; an item that is difficult to comprehend may nevertheless have one perfectly clear and unambiguous meaning if enough time is spent trying to work out its meaning. Thus, a high degree of comprehensibility is particularly important to make sure that subjects with little verbal sophistication answer correctly, whereas subjects with high verbal sophistication are more likely to notice and to be bothered by ambiguity.

Ambiguity. Ambiguity has been traditionally defined as doubtfulness or uncertainty about the meaning of a stimulus; more specifically, an item is called ambiguous if its surface structure can be interpreted in more than one way, all of which are linguistically correct. The concept of ambiguity implies intraindividual uncertainty about the meaning of a stimulus, whereas interindividual differences in stimulus interpretation have been referred to as *equivocality* (Goldberg, 1963).

Since ambiguity is generally assumed to influence the interpretation of items, this item characteristic should be relevant at the encoding stage, the stage called stimulus comprehension by Rogers (1974a), intellectual evaluation of question and answer by Nowakowska (1970), and meaning assignment by Kuncel (1973). Item ambiguity can be due to the presence of words or phrases that have several meanings (Pepper, 1981), to relations among sentence clauses that are not unequivocal, or to incompatibilities between item and response format. The latter may stem from negations, or from conjunctions (e.g., *and, or*) in items that contain more than one main clause.

Item ambiguity can have two undesirable effects on response processing. First, the subject may not recognize the ambiguous nature of the item and may "misunderstand" it (i.e., interpret it in a way not intended by the author). A

second possibility is that the subject does recognize the ambiguity but is not sure which of the two (or more) meanings is the intended one, thereby allowing some other variable unrelated to the item content, such as interpretive acquiescence (Messick, 1967), to influence the response.

Abstractness. We expect this item characteristic to influence the processing of items during the item-self comparison stage. Abstractness is related to Nowakowska's (1970) memory variable called "specific past experience," and refers to the extent to which an item inquires about abstract information. In particular, the more abstract the information required, the more likely it is that specific information stored in memory has to be further analyzed (e.g., integrated, supplemented by inferences) during the item-self comparison stage.

Concrete items refer to specific behaviors, and explicitly name relevant conditions, situations, people, or groups; they may even involve facts whose veridicality can be ascertained. In contrast, items that attribute to the respondent a general disposition, attitude, or wish are more abstract; these items require the interpretation of various past events (e.g., behaviors and experiences), their integration across different situations, comparisons with some unspecified standards, and other inferences. In other words, abstract items give the respondent considerable leeway in selecting information pertinent to the item and in integrating it into a response. Thus, the more abstract the item, the less likely it is that different subjects will arrive at a uniform understanding of its meaning.

Nisbett and Ross (1980, pp. 47–51) have reviewed evidence showing that concrete and vivid information is more easily recalled than is abstract information. Bellezza (1984) demonstrated that over a 1-week interval information about concrete nouns was more reliably retrieved from memory than was information about abstract nouns, and Payne (1970) showed that items rated as referring to explicit and clearly observable behaviors elicited more response stability than did items rated low on behavioral specificity. Thus, in addition to interindividual differences in understanding, the abstractness of the item content may lead to intraindividual response inconsistency.

Self-Reference. Self-reference is another item characteristic that is relevant to the item-self comparison stage, because it may influence the ease with which subjects can relate the item content to their own personal experiences or self-concepts. Self-reference is here defined as the degree to which the structure and content of the item taps the respondents' self-concept. An item with a high degree of self-reference permits the subjects to respond on the basis of their own perceptions and experiences. Self-reference is present in an item to the extent that the respondent is directly mentioned in the item, experiences or does something, is the object of an action, and/or is emotionally involved.

Evaluation. Items differ in the extent to which they invoke socioculturally determined values, norms, and standards. The more evaluatively extreme (i.e.,

either desirable or undesirable) the item content, the more likely it is that subjects will select the more socially desirable response. This item characteristic is particularly relevant to the final stage of the response process, namely utility control during response selection (Nowakowska, 1970; Rogers, 1971). Paulhus (this volume) differentiates among various mechanisms that might be involved in this process.

The Rating Study

In all, 1624 personality questionnaire items were rated by 230 paid judges who were advanced students of psychology or German literature. The instructions to the judges contained carefully written definitions of the five characteristics to be rated, several examples intended to anchor the steps of the rating scales, and several practice trials using pretested items (see Langer & Schulz von Thun, 1974). Each judge rated approximately 380 items on only one of the five item characteristics, resulting in 11–12 judges for each dimension and item. The original data were obtained by Löhr (1977) and Keren (1979), who dichotomized the verbally labeled four-step rating scales. The Comprehensibility scale was partitioned into "immediately understandable" (1) vs "not immediately understandable" (2+3+4); on the other four scales, the two higher (1+2) and the two lower (3+4) values were combined (see Table 5). All analyses summarized here were performed on these dichotomous data.

Overall, the mean interjudge agreement was 76%, and the mean value for Kappa was .50. Agreement among judges was consistently lower for the Ambiguity ratings than for the other four dimensions, agreement being highest on the Comprehensibility judgments. On the other hand, intrajudge consistency (assessed across items presented twice) exceeded 80% for Ambiguity, which was at least as stable as the other judgments. Thus, despite Löhr's (1977) and Keren's (1979) efforts to devise clear instructions, the judges seemed to have internally consistent but somewhat idiosyncratic conceptions of ambiguity.

Results. Table 6 presents the percentages of items that were classified by the majority of judges as ambiguous, abstract, value laden, not immediately understandable, and not self-referent; these proportions are presented separately for each of these five dichotomous dimensions and for each of the seven Anglo-American personality questionnaires described in this chapter.

Overall, more than 50% of the items in these personality questionnaires were not immediately understandable upon first reading. Cattell's 16PF was found to contain about twice as many items that were not easy to understand as either the MMPI or the four questionnaires constructed by Eysenck. Considering all seven questionnaires together, more than 25% of their items were classified as ambiguous. The 16PF had the highest percentage of ambiguous items, whereas Eysenck's questionnaires ranged from 10% (MPI) to 35% (EPI-A) ambiguous items. More than 40% of the items in this set had an ab-

Table 6. Item characteristics related to the response process: Percentage of items classified in each category for seven questionnaires. (Compiled from Löhr, in preparation)

Questionnaire	No. of items	Not immediately understandable	Ambigous	Abstract	Not self-referent	Evaluative
MPI	48	41.7	10.4	68.8	12.5	47.9
MMQ	56	35.7	19.6	32.1	12.5	58.9
EPI-A	57	40.4	35.1	45.6	12.3	71.9
EPI-B	57	29.8	17.5	42.1	45.6	50.9
16PF-A	171	72.5	39.2	42.7	31.0	37.4
16PF-B	171	83.0	33.3	37.4	46.8	48.5
MMPI	404	40.6	22.5	43.6	36.1	47.0
All personality	964	52.9	27.1	42.9	33.7	48.1
Interest[a]	168	82.1	15.5	6.0	96.4	2.4
Verbal IQ[b]	26	92.3	3.8	0.0	100.0	0.0

[a] These percentage are based on eight 21-item vocational interest scales included in Mittenecker and Toman's (1951) Persönlichkeits-Interessen-Test and are given here for comparison only.

[b] The items of the 16PF scale B were not included with the other personality scales because that scale was devised to assess intelligence.

stract rather than concrete reference. With the exception of Eysenck's MPI, which had more than 60% abstract items, the questionnaires did not differ much from each other on this dimension. One third of the items lacked reference to the respondent's perceptions and personal experiences; the items of EPI-B and 16PF-B were particularly impersonal, whereas most MPI, MMQ, and EPI-A items did contain direct references to the respondent. Finally, almost 50% of the items were judged as highly value laden; among the seven questionnaires, the percentages ranged from 37% (16PF-A) to 72% (EPI-A).

Table 7 presents the relative frequencies of these item characteristics separately for each scale of the seven questionnaires. Scales assessing the same or similar constructs appear next to each other. The first section of Table 7 permits a comparison among three Extraversion, four Neuroticism, and three Lie scales, all constructed by Eysenck and his coworkers. With regard to the frequencies of difficult and ambiguous items, the three kinds of scales did not differ much from each other. The other three characteristics, however, showed some interesting differences. First, the Neuroticism scales had the most abstract items on average, followed by the Extraversion scales, whereas the Lie scales had a relatively large number of concrete items. Second, all four Neuroticism scales contained more self-referent items than any of the three Lie scales constructed by Eysenck; if the scales within each questionnaire are compared for self-reference, Neuroticism was always highest and the Lie scales were al-

Table 7. Item characteristics related to the response process: Percentage of items classified in each category for all scales. (Based on Löhr, in preparation)

Questionnaire	Scale (no. of items)	Not immediately understandable	Ambiguous	Abstract	Not self-referent	Evaluative
MPI	E (24)	33.3	4.2	62.5	20.8	50.0
EPI-A	E (24)	41.7	33.3	29.2	16.7	58.3
EPI-B	E (24)	45.8	29.2	41.7	58.3	54.2
MPI	N (24)	50.0	16.7	75.0	4.2	45.8
MMQ	N (38)	34.2	13.2	39.5	2.6	47.4
EPI-A	N (24)	37.5	41.7	62.5	0.0	79.2
EPI-B	N (24)	25.0	8.3	45.8	20.8	37.5
MMQ	L (18)	38.9	33.3	16.7	33.3	83.3
EPI-A	L (9)	44.4	22.2	44.4	33.3	88.9
EPI-B	L (9)	0.0	11.1	33.3	77.8	77.8
16PF-A	F (13)	53.8	23.1	38.5	46.2	15.4
16PF-B	F (13)	76.9	15.4	46.2	23.1	46.2
16PF-A	H (13)	61.5	15.4	30.8	7.7	15.4
16PF-B	H (13)	61.5	53.8	38.5	15.4	69.2
16PF-A	E (13)	84.6	46.2	53.8	0.0	69.2
16PF-B	E (13)	84.6	30.8	30.8	69.2	76.9
16PF-A	Q2 (10)	90.0	40.0	50.0	40.0	20.0
16PF-B	Q2 (10)	90.0	30.0	40.0	70.0	30.0
16PF-A	A (10)	50.0	20.0	10.0	60.0	10.0
16PF-B	A (10)	80.0	20.0	0.0	70.0	0.0
16PF-A	I (10)	40.0	30.0	0.0	40.0	10.0
16PF-B	I (10)	60.0	20.0	30.0	50.0	40.0
16PF-A	L (10)	90.0	40.0	50.0	40.0	20.0
16PF-B	L (10)	100.0	30.0	90.0	60.0	70.0
16PF-A	C (13)	61.5	38.5	53.8	15.4	30.8
16PF-B	C (13)	100.0	38.5	53.8	30.8	30.8
16PF-A	O (13)	61.5	69.2	46.2	15.4	53.8
16PF-B	O (13)	92.3	46.2	46.2	15.4	61.5
16PF-A	Q4 (13)	76.9	53.8	53.8	0.0	46.2
16PF-B	Q4 (13)	61.5	7.7	15.4	15.4	38.5
16PF-A	G (10)	90.0	20.0	70.0	50.0	70.0
16PF-B	G (10)	100.0	30.0	30.0	70.0	70.0
16PF-A	Q3 (10)	70.0	30.0	30.0	30.0	70.0
16PF-B	Q3 (10)	90.0	30.0	40.0	40.0	70.0
16PF-A	Q1 (10)	100.0	40.0	50.0	60.0	40.0
16PF-B	Q1 (10)	80.0	50.0	30.0	80.0	30.0
16PF-A	M (13)	76.9	46.2	38.5	38.5	30.8
16PF-B	M (13)	92.3	61.5	53.8	76.9	46.2
16PF-A	N (10)	90.0	70.0	60.0	50.0	60.0
16PF-B	N (10)	80.0	30.0	10.0	40.0	40.0
MMPI	L (15)	33.3	13.3	40.0	46.7	66.7
MMPI	F (64)	39.1	18.8	46.9	39.1	67.2
MMPI	K (30)	43.3	36.7	46.7	33.3	50.0
MMPI	Hs (33)	15.2	18.2	6.1	78.8	0.0
MMPI	D (60)	30.0	23.3	41.7	26.7	43.3

Table 7 (continued)

Questionnaire	Scale (no. of items)	Not immediately understandable	Ambiguous	Abstract	Not self-referent	Evaluative
MMPI	Hy (60)	41.7	25.0	35.0	45.0	31.7
MMPI	Pd (50)	36.0	22.0	62.0	32.0	64.0
MMPI	Mf (60)	30.0	6.7	26.7	36.7	25.0
MMPI	Pa (40)	45.0	32.5	75.0	30.0	50.0
MMPI	Pt (48)	39.6	33.3	58.3	12.5	66.7
MMPI	Sc (78)	43.6	25.6	51.3	23.1	64.1
MMPI	Ma (46)	63.0	32.6	52.2	21.7	52.2
MMPI	Si (70)	51.4	20.0	44.3	38.6	44.3
MMPI	Sd (39)	35.9	17.9	41.0	30.8	61.5

ways lowest. Third, whereas both Neuroticism and Extraversion scales contained approximately equal numbers of evaluative and neutral items, about 80% of the Lie scale items were classified as value laden, a finding which is quite consistent with the intended purpose of such scales.

All but two of the 16PF scales contained more than 50% items classified as not immediately understandable; on several scales [e.g., L (trust vs suspicion)], the percentage of such difficult items reached 90–100. Equally disconcerting was the high number of ambiguous items on the 16PF; the highest percentage of ambiguous items was concentrated on scales M and N, which supposedly assess aspects of cultural sophistication. As with Eysenck's Neuroticism scales, the 16PF scales that measure aspects of anxiety (C, O, Q4) had somewhat more abstract items and many more self-referent items than the other sets of scales. The fewest self-referent items were found for the scales G and Q1 (conscientiousness and traditionalism) and the cultural sophistication scales M and N. Finally, most of Cattell's scales were as value laden as the scales of the other test authors, with the exception of the cluster formed by the A, I, and L scales (warmth and sensitivity).

The distribution of the MMPI items on these characteristics generally mirrored the patterns found for similarly named scales constructed by Eysenck and by Cattell. For example, the MMPI Lie (L) scale had more concrete than abstract items, only about 50% self-referent items (i.e., it was the secondlowest MMPI scale on this characteristic), and a large percentage of value laden items. The Social Introversion (Si) scale was also similar to the Extraversion scales in this sample; its items were relatively unambiguous and self-referent. The Psychasthenia (Pt) and Schizophrenia (Sc) scales resembled the pattern obtained for Eysenck's Neuroticism and Cattell's Anxiety-related scales, namely high percentages of abstract, self-referent, and value laden items. Finally, the Hypochondriasis (Hs) scale should be mentioned for its exceptionally high values on most of these characteristics: 85% of its items were classi-

fied as immediately understandable, 82% as unambiguous, 94% as referring to concrete and specific events, and all 100% of them were classified as evaluatively neutral.

Table 7 also shows that the presumably parallel scales of EPI-A and -B and of 16PF-A and -B did not necessarily include items with similar characteristics; for example, all items of the EPI-B Lie scale were classified as immediately understandable, whereas only 56% of the EPI-A version of the same scale received that classification. Moreover, the Neuroticism scale of EPI-B contained three times as many evaluatively neutral items as did that of EPI-A. Similar examples can be found for a variety of "parallel" 16PF scales. Given such disparities, any claim that these scales are psychometrically equivalent may be hard to justify.

Discussion. These differences even among presumably parallel scales also imply that caution is necessary in the interpretation of the associations between scale content and the processing characteristics of the corresponding items. Nevertheless, we believe that the scale-by-scale analyses show three discernable trends. Neuroticism and related constructs tended to be measured by abstract and overwhelmingly self-referent items; their comprehensibility and ambiguity varied with the test and the test author. Extraversion items tended to be somewhat less abstract and less self-referent. Finally, Lie scale items tended to be relatively unambiguous, more concrete, less highly self-referent, and more value laden than either Neuroticism or Extraversion items.

How should the overall distributions of these characteristics be evaluated? Fortunately, the total battery of questionnaires classified by Löhr (1977) and Keren (1979) included a few vocational interest scales from a German personality and interest test (Mittenecker & Toman, 1951), as well as Cattell's 16PF intelligence scales. For comparison purposes, the percentages of processing characteristics for these two types of items are given in the last two rows of Table 6.

In contrast to the personality trait scales, vocational interest scales and IQ tests do not require the respondent to report or provide information about their past behaviors or personal experiences; instead they ask for a preference for a job and for the solution to a problem, respectively. Thus, the self-reference criterion crucial to the assessment of traits is largely irrelevant for interest and intelligence items. Quite appropriately, then, these items were found to be *uniformly not* self-referent. Moreover, the interest and IQ items, in contrast to the personality items, had extremely homogeneous distributions on the other item characteristics as well. With few exceptions, they were unambiguous and concrete, and did not involve social norms and values (see also Fiske, 1971, p. 222). Given that the IQ items were designed to differentiate subjects with high verbal skills from those with low ones, the finding that almost all IQ items were not easy to understand may actually indicate some construct validity for these classifications. On IQ tests, subjects are expected to try hard to think about each item and figure out the correct meaning; that there generally is only

one correct interpretation is shown by the small percentages of ambiguous IQ (and interest) items. In instructions for personality questionnaires, however, subjects are told *not* to think or deliberate very much. The instructions for Cattell's 16PF, for example, emphasize that one should *"Give the first, natural answer as it comes to you* ... Give the best answer you can at a rate not slower than five or six a minute" (emphasis in original), and those for Eysenck's EPI state that one should "Respond quickly ... we would like to assess your first reaction and not the result of long deliberations." Given this kind of instructional set, the effects of insufficient comprehensibility and ambiguity may be even more pronounced. Thus, in our opinion, the results presented here show that too many items of personality and validity scales have poor information processing characteristics: More than 50% are difficult to understand; 25% are ambiguous; and almost 50% are abstract and value laden.

Conclusions. Over the past decade, there has been increasing consensus in the assessment literature that the content of an item should make theoretical sense (Fiske, 1971; Goldberg & Slovic, 1967; Jackson, 1971; Meehl, 1972; Mischel, 1972; Wiggins, 1973). Thus, if the the goal of test construction is to maximize the content variance in item responses, successful and efficient communication of this content is imperative. This is even more important because responding to the items of paper-and-pencil inventories constitutes a highly constrained communication situation. The respondents are expected to provide information about themselves in a reliable and valid way. To do so, however, they have to use a notoriously fuzzy mode of communication, language, in a highly standardized format, their only means of response being the selection of inflexible alternatives. They cannot ask the test constructor, "What do you mean by that?" Thus, in order to ensure that communication is successful under the impoverished conditions of the testing situation, authors of personality questionnaires must make every effort to limit the effects of factors that are likely to introduce errors into this verbal process.

To summarize, then, items should be easy to understand; otherwise, differences between subjects in verbal sophistication and conscientiousness will introduce errors. Items should be unambiguous; otherwise, it is unclear what the subject responded to and thus what the response means. Items should be concrete; otherwise, a uniform understanding of the item cannot be guaranteed. Perhaps items should be self-referent; otherwise, subjects may be unable to relate the question to their own frames of reference and to provide meaningful answers. Items should be as neutral as possible; otherwise, some subjects are likely to be influenced as much by the value implications of their responses as by the item content. The data presented so far suggests that none of the authors of the personality questionnaires studied here have constructed item pools with all of those characteristics.

The Prediction of Psychometric Characteristics

In the preceding sections of this chapter, we have described and evaluated items on the basis of logical-semantic categories, linguistic surface-structure variables, and cognitive-processing characteristics. In this final part, we turn to those item characteristics that have traditionally been of the most interest to the psychometrically inclined psychologist, item *response* parameters obtained from a group of subjects. We will briefly describe these item statistics and their distributions in a large sample of items taken from personality and validity scales. Whereas the ultimate goal of itemmetric research should be the discovery of the relationships between item properties and *scale* validity, we will restrict ourselves here to the prediction of *item* response parameters. Using multiple regression analyses, we will examine the extent to which item stability and internal validity are influenced by the three kinds of item characteristics that we have discussed in the preceding sections.

Item Response Parameters

Item statistics were obtained by Angleitner (1976, 1981 a), who administered the whole set of questionnaires to a sample of 102 normal adult subjects twice, with a test–retest interval of 2–3 weeks. The actual sample sizes varied from 81 to 99 subjects, depending on the questionnaire. Subjects filled out the questionnaires at home at their own pace. Frequencies of endorsement, frequencies of stable responses, and item-test correlations were used for the analyses described below.

Endorsement frequencies were transformed into adjusted *item means* to obtain an equivalent measure of central tendency independent of differences in response formats and keying among the questionnaires. For Eysenck's MPI, Cattell's 16PF, and the Giessen test (all of which offer more than two response options), the weights of the response options were adjusted to a range from 0 (minimum weight) to 1 (maximum weight), with equidistant intermediate weights. Then, for all reverse-keyed items, the response weights were subtracted from 1. The weights of items with dichotomous response formats and affirmative keying direction remained unaltered. Finally, item means were computed from the new weights and the corresponding endorsement frequencies. *Item variances* were also computed from these adjusted data.

Both test and retest data were transformed in this way, thus permitting the computation of reliability estimates. For the item means, the retest correlation was .98, and for the item variances it was .95. Given these high correlations, only the sample statistics from the first administration will be reported here. Table 8 presents the means and standard deviations (or percentages) for all item properties based on a set of 1051 items. In this set, the items from the interest and intelligence scales were eliminated, as were a large number of MMPI items, because their multiple keying did not permit the calculation of a single estimate of internal validity for each item. Specifically, this item sample con-

Table 8. Means and standard deviations of all item characteristics across 1051 items from personality and validity scales (Löhr, in preparations)

	Mean	SD	Percentages
I. *Logical item-trait relations*			
Description of reactions: Overt			20.7
Covert			27.2
Physical symptom			9.0
Not subclassified			10.3
Trait attribution			11.7
Indirect modes of self-description			14.7
Unclassified			6.4
II. *Item surface structure*			
Number of letters	67.98	32.20	
Number of words	12.40	5.70	
Number of clauses	1.97	1.04	
Negation: Negative phrasing			26.1
Tense: Past			6.4
Voice: Passive			3.8
Mood: Subjunctive			10.1
Impersonal phrasing: No first person pronoun			4.3
Sentence.type: Question			15.4
Number of response alternatives	2.54	.97	
Keying direction: 'Yes'			58.6
III. *Aspects of response process*			
Not immediately understandable			52.0
Ambiguous			29.2
Abstract			47.1
Not self-referent			27.9
Evaluative (not neutral)			54.0
IV. *Psychometric characteristics*			
Item mean[a]	.48	.23	
Item variance[a]	.18	.06	
Item stability	.82	.07	
Item-test correlation[a]	.28	.20	

[a] Data from the first administration.

sisted of the seven questionnaires of Anglo-American origin (MPI, MMQ, EPI-A, and -B, 16PF-A and -B, and the MMPI scales L, D, and Si), plus the personality and validity scales of the German PIT, Giessen Test, and FPI.

As an estimate of the *internal validity* or discriminating power of each item, the point-biserial item-test correlation (part-whole corrected) was used. For all statistical analyses, these correlation coefficients were transformed to Fisher's Z. The reliability of the item validities (first with second administration) was .84.

The proportion of subjects who did not change their response from the first to the second administration was used as an index of *item stability*. Again, the multiple response formats used by three of the questionnaires caused a problem because the probability of changing responses increases with the number of possible response alternatives. Therefore, the response continuum was divided into two halves at its midpoint, regardless of the actual response format, thus redefining item stability with reference to the upper and lower halves of the response continuum. Response options that fell exactly at the midpoint were assigned at random. The stability index for items with multiple response formats was recomputed from the raw data (Löhr, in preparation); for the other items, it was adopted from Angleitner's (1981 a) report. Unlike the other item statistics, a reliability estimate of item stability was not available.

Relations Among These Item Statistics. The frequency distributions of these item statistics were unimodal and reasonably symmetric, with the exception of the item variances, which were extremely skewed. Given the relationship between item mean and variance, and the inconvenient shape of this distribution, item variance did not appear to be a very useful variable for regression analyses. The item mean can influence item stability and validity in the following ways: (a) The more extreme the item mean (i.e., closer to either 0 or 1), the more likely is the item to elicit a high degree of response stability. (b) The more extreme the item mean, the more likely is the item to have low internal validity (i.e., discriminating power). It follows from (a) and (b) that (c) a negative relationship between item stability and validity is to be expected. This phenomenon has been called the psychometric paradox (Goldberg, 1963; Nowakowska, 1970).

In these data, the relationship between item mean and stability was, as expected, curvilinear (U-shaped) and rather strong, the multiple R (using both linear and quadratic terms) being .66. The relationship between item mean and validity, however, was essentially linear and unexpectedly weak ($r = -.12$, multiple $R = .16$), although both the linear and the quadratic components were significant in this large item sample ($n = 1051$). And finally, a weak (again significant) *positive* correlation between item validity and stability ($r = .12$) indicated that the psychometric paradox did not seem to characterize these data. Moreover, an inspection of the bivariate scatterplots suggested no other types of relationships among the item statistics. In summary, the empirical relations among the item statistics suggest that item stability and validity as assessed by the item-test correlation can be treated as practically independent criterion variables. This finding will be important when we interpret the regression analyses.

Item Response Statistics from the Questionnaires in this Sample

Psychometric theory asserts that the consistency of item responses should approach the hypothesized stability of the behaviors being measured. Stability

of response is, therefore, essential at both the item and the scale level. However, as in most studies of item response characteristics, one cannot help but be impressed by the degree of response inconsistency elicited by most personality questionnaire items. As Table 6 shows, almost 20% of the subjects changed their response to the average item over the 2-week test-retest interval. This figure is quite similar to the stability percentages that Goldberg (1963, p. 468) compiled 20 years ago for "some typical early" and for "six more recently developed" personality inventories.

In general, the item validities were relatively low. As Table 8 shows, the average item validity was .28, compared to .36 which is typical for carefully constructed verbal aptitude tests (Green, 1978). Moreover, about 4.8% of the item validities were *negative* in sign, 6.6% were zero, and 13.7% were about 0.1; yet, only one-third of these zero validities can be explained by extremeness of the corresponding item means. In contrast, the verbal aptitude test described by Green (1978) included only two items that had validities lower than .20 (their validities were .16). Obviously, a considerable number of personality questionnaire items have been constructed and selected so poorly that they fall into a range of validity that is characterized by zero discriminating power.

Table 9 presents the mean item stability, mean item-test correlation, the coefficient Alpha, and (when possible) the correlations among parallel scales for each scale of the seven Anglo-American personality questionnaires. The first section of Table 9 permits a comparison among the scales constructed by Eysenck and his coworkers. Overall, the Neuroticism scales were most internally consistent and obtained the highest parallel scale correlations; the Extraversion scales were somewhat less satisfactory. The relatively short Lie scales of the EPI had very low internal consistency values (Alphas of .66 and .61, respectively) and a low parallel test correlation ($r = .51$). In this sample, five of the ten Eysenck scales did not reach his goal (Eysenck, 1953, p. 300) that questionnaire scales should have an Alpha coefficient of at least .80. However, to obtain such high degrees of internal consistency in Neuroticism and Extraversion scales, given the enormous breadth of these constructs, their length would have to be substantially increased. Similarly, the retest reliabilities of Eysenck's scales did not meet his own standard (Eysenck, 1953, p. 301) of .90 (Angleitner, 1976, p. 328). However, since scale stabilities not only reflect response consistency at the item level but also the particular item composition of the scale, we believe that item response stability (corrected for endorsement frequency) may be the more important index. Indeed, item stability should be used as an item selection criterion early in scale construction, and the item stabilities should be reported in the test manual. Despite the generally accepted importance of stability of measurement, in none of today's widely used personality questionnaires has item stability been applied as an item exclusion criterion during the construction stage, not even in Jackson's otherwise exemplary PRF (Jackson, 1970).

As Table 9 shows, the mean item stabilities of the Eysenck scales generally exceeded 80% response consistency over an interval of 2–3 weeks; the item sta-

Table 9. Mean item stability, item-test correlation, coefficient Alpha, and parallel form reliabilities of seven personality questionnaires ($n = 105$)

Questionnaire	Scale	No. of items	Mean item stability	Mean item-test correlation[a]	Alpha	Parallel form correlation[a]
MPI	E	24	.78	.36	.81	
EPI-A	E	24	.84	.31	.77	.73
EPI-B	E	24	.84	.28	.73	
MPI	N	24	.80	.46	.88	
MMQ	N	38	.89	.36	.86	
EPI-A	N	24	.84	.48	.89	.85
EPI-B	N	24	.84	.39	.84	
MMQ	L	18	.86	.36	.77	
EPI-A	L	9	.82	.35	.66	.51
EPI-B	L	9	.89	.32	.61	
16PF-A	F	13	.77	.32	.70	.61
16PF-B	F	13	.76	.27	.62	
16PF-A	H	13	.76	.33	.70	.71
16PF-B	H	13	.73	.42	.78	
16PF-A	E	13	.72	.14	.40	.52
16PF-B	E	13	.80	.14	.39	
16PF-A	Q2	10	.76	.09	.26	.42
16PF-B	Q2	10	.73	.15	.37	
16PF-A	A	10	.75	.08	.24	.38
16PF-B	A	10	.78	.14	.36	
16PF-A	I	10	.79	.07	.19	.36
16PF-B	I	10	.80	.11	.28	
16PF-A	L	10	.78	.13	.34	.33
16PF-B	L	10	.78	.18	.43	
16PF-A	C	13	.73	.10	.29	.53
16PF-B	C	13	.75	.29	.66	
16PF-A	O	13	.72	.22	.52	.57
16PF-B	O	13	.70	.33	.68	
16PF-A	Q4	13	.73	.35	.73	.64
16PF-B	Q4	13	.71	.30	.67	
16PF-A	G	10	.75	.25	.56	.39
16PF-B	G	10	.77	.12	.33	
16PF-A	Q3	10	.73	.15	.37	.20
16PF-B	Q3	10	.75	.10	.26	
16PF-A	Q1	10	.75	.04	.10	.20
16PF-B	Q1	10	.74	−.02 (!)	−.07 (!)	
16PF-A	M	13	.77	.09	.29	.32
16PF-B	M	13	.76	.14	.39	
16PF-A	N	10	.74	.10	.26	.03 (!)
16PF-B	N	10	.76	−.04 (!)	−.15 (!)	
MMPI	L	15	.88	.28	.67	
MMPI	F	64	.92	.18	.74	
MMPI	K	30	.82	.26	.74	
MMPI	Hs	33	.88	.33	.83	
MMPI	D	60	.85	.16	.66	
MMPI	Hy	60	.85	.11	.56	

Table 9 (continued)

Questionnaire	Scale	No. of items	Mean item stability	Mean item-test correlation[a]	Alpha	Parallel form correlation[a]
MMPI	Pd	50	.84	.20	.71	
MMPI	Mf	60	.86	.18	.72	
MMPI	Pa	40	.87	.11	.44	
MMPI	Pt	48	.87	.37	.89	
MMPI	Sc	78	.89	.25	.86	
MMPI	Ma	46	.83	.11	.51	
MMPI	Si	70	.84	.23	.83	
MMPI	Sd	39	.87	.26	.86	
Means across items and scales						
Eysenck		218	.84	.37	.78	.72
Cattell		342	.75	.18	.40	.43
MMPI		404	.86	.23	.72	–

The item stabilities computed for the 16PF were corrected as described in the text
[a] Data from the first administration

bilities did not differ systematically from scale to scale and were only slightly lower on average than those of the MMPI scales. Overall, the psychometric characteristics of Eysenck's scales were fairly respectable, in particular when they are evaluated with reference to the other questionnaires in this sample, rather than in terms of Eysenck's absolute standards.

In each section of this chapter, in which we have compared the item sets of the seven questionnaires, we had to conclude that the items of Cattell's 16PF were particularly poor. Of all the questionnaires studied here, the 16PF had the longest, most complex, most difficult to understand, and most ambiguous items, and had the highest frequency of indirect types of trait indicators. The psychometric properties of the resulting scales are shown in Table 9; these findings are quite representative of those reported by other researchers (e.g., Bartussek, Weise, & Heinze, 1972). Since Cattell (e.g., 1973) rejects scale homogeneity as a relevant psychometric characteristic, it might be argued that evaluations of the 16PF scales have to focus on stability and parallel form reliability. However, these are also low (for a review, see Angleitner, 1976); Table 9 shows that the correlations between "parallel" scales were particularly low for those scales with extremely low internal consistency, perhaps suggesting that Cattell's scales could use more scale homogeneity.

All MMPI scales had item response stabilities well above .80. The average item-test correlations of the MMPI scales were lower than those of the Eysenck questionnaires, and the Alpha coefficients were, in light of the long scales, not very impressive. However, these results are based on a sample of

normal adults and do not provide a complete appraisal of the MMPI, since that inventory was constructed primarily for clinical populations.

The Predictor Variables and their Interrelations

In the first three sections of this chapter we discussed three sets of stimulus characteristics of personality questionnaire items that might influence psychometric item and scale properties. Table 8 presents an overview of these variables, along with their distribution in the sample of 1051 personality and validity scale items. The empirical relations among these item characteristics, and between them and the four item response parameters, will be described by Löhr (in preparation). A summary of these analyses will now be presented here.

The correlations among the five response process characteristics were generally weak and in all cases much smaller in size than the reliabilities of the judgments, although five of the ten Phi coefficients were statistically significant due to the large sample size. Ambiguity correlated $-.22$ with comprehensibility and .26 with abstractness; although both correlations are in the logical direction, their size indicates that the judges were able to differentiate these three constructs from each other. Thus, we can conclude that these five ratings reflect fairly independent item characteristics.

Also of interest were the relations between these judged item characteristics and the more objectively scored item-surface characteristics decribed in the second section of this chapter. Only a few substantial relationships were found; most of them conformed to our expectations and thus provide some degree of construct validity for the rating-based measures. Specifically, the measures of item length (i.e., the number of letters and of sentence clauses) correlated approximately $-.60$ with the comprehensibility ratings; similarly, the greater the number of response options, the more likely was the item to be judged as difficult to understand (Phi = .30). Moreover, items whose text did not contain any personal pronouns were more likely to be classified as low in self-reference (Phi = .34).

Finally, there were a few substantial relationships between these two sets of item characteristics and the logical categories of item-trait relations (defined in Table 2). In particular, items classified as *indirect* trait indicators (i.e., items bearing an indirect relation to the trait to be assessed, such as wishes or attitudes) were likely to contain subjunctive verb forms (Phi = .48) and no personal pronouns (Phi = .50); with regard to their processing characteristics, such items were likely to be classified as not self-referent (Phi = .37).

Predicting Response Parameters

In order to abstract the general trends from the data, rather than focusing on the results from single variables, Löhr (paper in preparation) computed multiple regressions using each of the psychometric indices as criteria and each of the three sets of item stimulus characteristics as predictors. The findings from these analyses are summarized in Table 10.

Table 10. Multiple correlations between item characteristics and two criteria: Item stability and validity (item-test correlation). (Löhr, in preparation)

Predictors	Item sets considered				
	Total ($n=1051$)	Odd half ($n=526$)	Even half ($n=525$)	1st half[a] ($n=530$)	2nd half[b] ($n=521$)
	Item stability				
Logical relation	.27	.29	.28	.30	.29
Surface text	.44	.43	.47	.48	.46
Response process	.37	.36	.40	.40	.38
All	.51	.51	.55	.56	.54
	Item validity				
Logical relation	.34	.33	.38	.45	.23
Surface text	.40	.42	.40	.46	.43
Response process	.37	.40	.34	.49	.26
All	.50	.53	.52	.58	.48

The item characteristics are listed in Table 8.
[a] Items from the Giessen-Test, MMQ, EPI-A, FPI, and 16PF-B
[b] Items from the MPI, EPI-B, PIT, 16PF-A, and from MMPI scales L, D, and Si

The multiple correlation between item stability and *all* stimulus characteristics was .51. The item internal validities, as assessed by the item-test correlations, were also highly predictable ($R=.50$). Considering the three sets of stimulus characteristics separately, the surface text characteristics that were objectively scored and thus most reliable were the most powerful predictors of response stability ($R=.44$) and item validity ($R=.40$). The five response process variables were also quite predictive of item stability and validity (both $R=.37$); the categorical variables specifying the logical relations between item and trait were found to be least predictive. These coefficients did not shrink when separate subsamples of approximately 525 items each were used. Thus, the multivariate relations presented in Table 10 seem to hold independently of the particular composition of the item sample.

Conclusions. These findings provide impressive empirical support for our general hypothesis: The influence of "formal" item characteristics on the psychometric quality of personality questionnaires is both systematic and substantial. More specifically, our findings suggest that the surface structure of the verbal stimulus, if measured comprehensively, is an essential determinant of the reliability and validity of subjects' item responses. It appears, then, that authors of personality questionnaires would be well advised to spend more time and care in item writing. Second, our findings show that undesirable information processing characteristics of items can be identified with the rating methodology described here. It appears, then, that personality questionnaire authors would be well advised to incorporate criteria such as item comprehensibility, nonambiguity, concreteness, self-reference, and nonevaluative content, into the stan-

dard canon of item selection strategies. Finally, we found that even the logical item-trait relations were related to the item-response parameters, with the more indirect modes of self-description yielding less satisfactory item-response statistics. It appears, then, that personality questionnaire authors would be well advised to think more deeply about the kinds of items that make useful trait indicators.

Although there is clearly plenty of room for improvement, a beginning has already been made. Jackson (1967, 1970) used social desirability and ambiguity as item exclusion criteria during the construction of the initial item pool from which the PRF was developed. Moreover, his construct-oriented approach fostered convergent and discriminant item validity at the very beginning of the questionnaire construction process. More recently, Buss and Craik (1980, 1983) have developed explicit procedures for the generation of trait-relevant items that belong to only one class of logical item types (i.e., overt behavioral acts) and are, in addition, relatively concrete and specific.

Further improvements will not come easily. Meehl (1972) spelled out what will be necessary:

> An empirical showing that an item does not have certain undesirable properties, namely, systematic loadings on nuisance variables, is just as important as an empirical showing that it possesses the desirable properties of criterion discrimination and internal consistency. We must begin routinely to design our test-construction and test-validation research with an eye to negative properties. What makes this such a methodological and logistical pain in the neck is, of course, that the list of undesirable properties is considerably longer than the list of desirable properties! That is to say, given the structure of the human mind (a fact of the world and not a mere weakness of psychometric method or theory), during the course of scale-construction we may have to spend more time, money, brains, and energy measuring things we do not want the item to reflect, than we do measuring things we do want it to reflect (pp. 161–162).

The challenge lies ahead.

Notes

1 Order of authorship is alphabetical; this chapter represents a fully collaborative effort by the three authors, based in part on a paper delivered at the Bielefeld Symposium on Personality Questionnaires, which was held at the Center for Interdisciplinary Research, University of Bielefeld, June 17–20, 1982. Some of the analyses presented in this chapter are based on Franz-Josef Löhr's PhD dissertation (in preparation), a complete report of which may appear elsewhere.
We are grateful to Peter Borkenau, David Buss, William F. Chaplin, Donald Fiske, Sarah Hampson, Willem Hofstee, and Wolf Nowack for their comments on earlier drafts, and especially to Lewis R. Goldberg, the master wordsmith, whose untiring efforts helped improve the readability of this chapter considerably. Our research was supported by the Deutsche Forschungsgemeinschaft (DFG) An 106/1–3. Funds for the support of the second author during the writing of this report were provided in part by Grant MH-39077 from the National Institute of Mental Health (U.S. Public Health Service).

2 With the exception of the fourth statement, these "items" were written by Art Buchwald, the noted American humorist (as cited in Goldberg, 1974).

3 Parts of this project have been described in German by Angleitner (1976, 1981a, 1981b), Angleitner and Löhr (1980), Löhr and Angleitner (1980), and Löhr (1983a, 1983b). These reports are available from A. Angleitner, Fakultät für Psychologie und Sport, Universität Bielefeld, D-4800 Bielefeld 1, West Germany.

4 The German edition of Jackson's (1967, 1970) Personality Research Form (Stumpf et al., 1985) was not available when this project was begun.
5 There seems to be an increasing agreement among test developers of varying persuasions that responses to single items are too fragile to carry the weight of external analyses. Cattell and Comrey, for example, have advocated the factoring of clusters of items; analogously, most proponents of the criterion-group strategy would now probably prefer using homogeneous clusters of items to differentiate between their pairs of criterion samples. Thus, whatever the test developer's particular approach, the need for homogeneous item sets requires an initial construction stage involving some reflection about the trait construct and relevant items.

References

Allport, F.H. (1937). Teleonomic description in the study of personality. *Character and Personality, 6*, 202–214.

Allport, G.W., & Odbert, H.S. (1936). Trait-names: A psycholexical study. *Psychological Monographs, 47* (1, Whole No. 211).

Amelang, M., Sommer, E., & Bartussek, D. (1971). Persönlichkeitsstruktur und Studienrichtung. *Psychologische Beiträge, 13*, 7–15.

Angleitner, A. (1976). *Methodische und theoretische Probleme bei Persönlichkeitsfragebogen unter besonderer Berücksichtigung neuerer deutsch-sprachiger Fragenbogen.* Bonn: Habilitationsschrift, Psychologisches Institut der Universität Bonn.

Angleitner, A. (1981 a). *Teststatistische Kennwerte der Items aus 10 deutschsprachigen Persönlichkeitsfragebogen.* Bielefeld: Arbeitsberichte aus dem Projekt Persönlichkeitsfragebogen (No. 2), Universität Bielefeld.

Angleitner, A. (1981 b). *Teststatistische Kennwerte der Items aus 8 deutschsprachigen Persönlichkeitsfragenbogen für Kinder und Jugendliche.* Bielefeld: Arbeitsberichte aus dem Projekt Persönlichkeitsfragebogen (No. 3), Universität Bielefeld.

Angleitner, A., & Löhr, F.J. (1980). Itemüberlappung zwischen Persönlichkeitsfragebogen als Problem für Validitätsschätzungen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 1*, 127–136.

Ashton, S.G., & Goldberg, L.R. (1973). In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. *Journal of Research in Personality, 7*, 1–20.

Bartussek, D., Weise, G., & Heinze, B. (1972). Reliabilität und faktorielle Validität des deutschen 16 PF-Tests von Cattell. *Arbeiten aus dem psychologischen Institut der Universität Hamburg, 19.*

Beckmann, D., & Richter, H.E. (1972). *Giessen-Test (GT). Ein Test für Individual- und Gruppendiagnostik. Handbuch.* Bern: Huber.

Bellezza, F.S. (1984). Reliability of retrieval from semantic memory: Noun meanings. *Bulletin of the Psychonomic Society, 22*, 377–380.

Buss, D.M., & Craik, K.H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. *Journal of Personality, 48*, 379–392.

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

Cantor, N. (1981). Perceptions of situations: Situations prototypes and person-situation prototypes. In Magnusson, D. (Ed.), *Toward a psychology of situations: An interactional perspective.* Hillsdale: Erlbaum.

Cantor, N., Mischel, W., & Schwartz, J.C. (1981). Social knowledge: Structure, content, use and abuse. In Hastorf, A., & Isen, A. (Eds.), *Cognitive social psychology.* New York: Elsevier/North Holland.

Cantor, N., Mischel, W., & Schwartz, J.C. (1982). A prototype analysis of psychological situations. *Cognitive Psychology, 14*, 45–77.

Cattell, R.B. (1943). The description of personality: 2. Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 476–507.
Cattell, R.B. (1946). *Description and measurement of personality*. New York: World Book.
Cattell, R.B. (1957). *Personality and motivation structure and measurement*. New York: World Book.
Cattell, R.B. (1973). *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.
Cattell, R.B., & Eber, H.W. (1962). *Specimen set for the Sixteen Personality Factor Questionnaire "16 PF", experimental edition der deutschen Form A und B*. Champaign: Institute for Personality and Ability Testing.
Cattell, R.B., & Eber, H.W. (1964). *The Sixteen Personality Factor Questionnaire*. Champaign: Institute for Personality and Ability Testing.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: M.I.T. Press.
Cliff, N., Bradley, P., & Girard, R. (1973). The investigation of cognitive models for inventory response. *Multivariate Behavioral Research, 8*, 407–425.
Cronbach, L.J. (1970). *Essentials of psychological testing*. New York: Harper & Row.
Dubois, B., & Burnes, J.A. (1975). An analysis of the meaning of the question mark response-category in attitude scales. *Educational and Psychological Measurement, 35*, 869–884.
Edwards, A.L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.
Eggert, D. (1974). *Eysenck-Persönlichkeits-Inventar (EPI). Handanweisung für die Durchführung und Auswertung*. Göttingen: Hogrefe.
Eysenck, H.J. (1947). *Dimensions of personality*. London: Routledge & Kegan Paul.
Eysenck, H.J. (1953). Fragebogen als Meßmittel der Persönlichkeit: Eine experimentelle Untersuchung. *Zeitschrift für experimentelle und angewandte Psychologie, 1*, 191–335.
Eysenck, H.J. (1959a). *Das "Maudsley Personality Inventory" (MPI)*. Göttingen: Hogrefe.
Eysenck, H.J. (1959b). Der Maudsley Personality Inventory als Bestimmer der neurotischen Tendenz und Extraversion. *Zeitschrift für experimentelle und angewandte Psychologie, 6*, 167–190.
Eysenck, H.J. (1962). *The Maudsley Personality Inventory*. (Manual prepared by R.R. Knapp). San Diego: Educational and Industrial Testing Service.
Eysenck, H.J. (1964). *MMQ*. Göttingen: Hogrefe.
Eysenck, H.J., & Eysenck, S. (1964). *Manual of the Eysenck Personality Inventory*. London: University of London Press.
Fahrenberg, J., & Selg, H. (1970). *Das Freiburger Persönlichkeitsinventar (FPI) – Handanweisung*. Göttingen: Hogrefe.
Fillmore, C.J. (1977). The case for case reopened. In Cole, P., & Sadock, J. (Eds.), *Grammatical relations* (pp. 59–81). New York: Academic.
Fiske, D.W. (1971). *Measuring the concepts of personality*. Chicago: Aldine.
Fiske, D.W. (1981). *New directions for methodology of social and behavioral science: Problems with language imprecision*. San Francisco: Jossey-Bass.
Forgas, J.P. (1982). Episode cognition: Internal representations of interaction routines. In Berkowitz, L. (Ed.), *Advances in experimental social psychology* (Vol. 15). New York: Academic.
Foss, D.J., Hakes, D.J., & Hakes, D.T. (1978). *Psycholinguistics*. Englewood Cliffs: Prentice Hall.
Givon, T. (1979). *Syntax and semantics 12: Discourse and syntax*. New York: Academic.
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. (1971). A historical survey of personality scales and inventories. In McReynolds, P. (Ed.), *Advances in psychological assessment* (Vol. 2, pp. 293–336). Palo Alto: Science and Behavior.

Goldberg, L.R. (1972). Some recent trends in personality assessment. *Journal of Personality Assessment, 36*, 547–560.

Goldberg, L.R. (1974). Objective diagnostic tests and measures. *Annual Review of Psychology, 25*, 343–366.

Goldberg, L.R. (1978). Differential attribution of trait-descriptive terms to oneself as compared to well-liked, neutral, and disliked others: A psychometric analysis. *Journal of Personality and Social Psychology, 36*, 1012–1028.

Goldberg, L.R. (1980, May). Some ruminations about the structure of individual differences: Developing a common lexicon for the major characteristics of human personality. *Meetings of the Western Psychological Association*, Honolulu, Hawaii.

Goldberg, L.R. (1981 a). Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology, 41(2)*, 141–165.

Goldberg, L.R. (1981 b). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality and Social Psychology, 41(3)*, 517–552.

Goldberg, L.R. (1982). *From Ace to Zombie:* Some explorations in the language of personality. In Spielberger, L.D., & Butcher, J.N. (Eds.), *Advances in personality assessment* (Vol. 1). Hillsdale: Erlbaum.

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.

Gorsuch, R.G., & Cattell, R.B. (1967). Second stratum personality factors defined in the questionnaire realm by the 16PF. *Multivariate Behavioral Research, 2*, 211–224.

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

Gough, H.G. (1965). Conceptual analysis of psychological test scores and other diagnostic variables. *Journal of Abnormal Psychology, 70*, 294–302.

Green, B.F., Jr. (1978). In defense of measurement. *American Psychologist, 33*, 664–670.

Hampshire, S. (1953). Dispositions. *Analysis, 14*, 5–11.

Hampson, S.E. (1982). Person memory: A semantic category model of personality traits. *British Journal of Psychology, 73*, 1–11.

Hathaway, S.R., & McKinley, J.C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory.* New York: Psychological Corporation.

Hull, J.G., & Levy, A.S. (1979). The organizational functions of the self: An alternative to the Duval and Wicklund model of self-awareness. *Journal of Personality and Social Psychology, 37*, 756–768.

Jackson, D.N. (1967). *Personality research form. Manual.* Goshen: Research Psychologists.

Jackson, D.N. (1970). A sequential system for personality scale development. In Spielberger, C.D. (Ed.), *Current topis in clinical and community psychology* (Vol. 2, pp. 61–96). New York: Academic.

Jackson, D.N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review, 78*, 229–248.

Jackson, D.N. (1975). The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational and Psychological Measurement, 35*, 361–370.

Janke, W. (1973). Das Dilemma von Persönlichkeitsfragebogen. Einleitung des Symposiums über Konstruktion von Fragebogen. In Reinhart, G. (Ed.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970.* Göttingen: Hogrefe.

John, O.P., Goldberg, L.R., & Angleitner, A. (1984). Better than the alphabet: Taxonomies of personality-descriptive terms in English, Dutch, and German. In Bonarius, H.M., Van Heck, G., & Smid, N. (Eds.), *Personality psychology in Europe: Theoretical and empirical developments.* Lisse: Swets & Zeitlinger.

Jones, R.R. (1965). The relationships between item properties and scale reliability. *Oregon Research Institute Research Monograph, 5* (No. 1).

Jones, R.R. (1968, August). Differences in response consistency and subjects' preferences for three personality inventory formats. *Proceedings of the 76 th Annual Convention of the American Psychological Association*, San Francisco.

Kelley, H.H., & Michela, G.L. (1980). Attribution theory and research. *Annual Review of Psychology, 34*, 457–504.

Keren, A. (1979). *Inhaltlich-semantische Analyse der Items aus fünf deutschsprachigen Persönlichkeitsfragebogen.* Universität Bonn: Masters Thesis.

Knudson, R.M., & Golding, S.L. (1974). Comparative validity of traditional versus S-R format inventories of interpersonal behavior. *Journal of Research in Personality, 8*, 111–127.

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

Langer, I., & Schulz von Thun, F. (1974). Messung komplexer Merkmale in Psychologie und Pädagogik – Ratingverfahren. Munich: Reinhardt.

Lennertz, E. (1973). Thesen zur Itemsammlung bei Persönlichkeitsfragebogen. In Reinert, G. (Ed.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1973.* Göttingen: Hogrefe.

Lienert, G.A. (1969). *Testaufbau und Testanalyse (3rd ed.).* Weinheim: Beltz.

Löhr, F.J. (1977). *Formale und inhaltliche Merkmale von Fragebogenitems – Zum Einfluß von Itemeigenschaften auf den Beantwortungsprozeß bei deutschen Persönlichkeitsfragebögen.* Universität Bonn: Masters Thesis.

Löhr, F.J. (1983a). *Häufigkeits- und Intensitätsadverbien in den Items von Persönlichkeitsfragebogen.* Bielefeld: Arbeitsberichte aus dem Projekt Persönlichkeitsfragebogen (No. 5), Universität Bielefeld.

Löhr, F.J. (1983b). *Sprachlogische und empirische Beziehungen in Persönlichkeitsfragebogen für Kinder. Eine empirische Untersuchung zur Bezugsebene der Items.* Bielefeld: Arbeitsberichte aus dem Projekt Persönlichkeitsfragebogen (No. 6), Universität Bielefeld.

Löhr, F.J. (In preparation). *Die Bedeutung von Itemeigenschaften für den Prozeß und das Ergebnis der Beantwortung von Persönlichkeitsfragebogen.* Bielefeld: Dissertation, Fakultät für Psychologie und Sportwissenschaft der Universität Bielefeld.

Löhr, F.J., & Angleitner, A. (1980). Eine Untersuchung zu sprachlichen Formulierungen der Items in deutschen Persönlichkeitsfragebogen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 1*, 217–235.

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

Magnusson, D. (1981). *Toward a psychology of situations: An interactional perspective.* Hillsdale: Erlbaum.

Magnusson, D. (1984). Persons in situations: Some comments on a current issue. In Bonarius, H.M., Van Heck, G., & Smid, N. (Eds.), *Personality psychology in Europe: Theoretical and empirical developments.* Lisse: Swets & Zeitlinger.

Meehl, P.E. (1972). Reactions, reflections, projections. In Butcher, J.N. (Ed.), *Objective personality assessment: Changing perspectives* (pp. 131–189). New York: Academic.

Messick, S. (1967). The psychology of acquiescence. In Berg, I.A. (Ed.), *Response set in personality assessment.* Chicago: Aldine.

Micklin, M., & Durbin, M. (1969). Syntactic dimensions of attitude scaling techniques: Sources of variation and bias. *Sociometry, 32*, 194–206.

Mischel, W. (1972). Direct vs. indirect personality assessment: Evidence and implications. *Journal of Consulting and Clinical Psychology, 38*, 319–324.

Mischel, W., & Peake, P.K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730–755.

Mittenecker, E., & Toman, W. (1951). Der P.I.-Test. *Beihefte zur Wiener Zeitschrift für Philosophie, Psychologie, Pädagogik, 1.*

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs: Prentice-Hall.

Nowakowska, M. (1970). A model of answering to a questionnaire item. *Acta Psychologica, 34*, 420–439.

Payne, F.D. (1970). *Structured personality inventory items: Test-retest response consistency and item properties.* Urbana-Champaign: Ph.D. Thesis, University of Illinois.

Pepper, S. (1981). Problems in the quantification of frequency expressions. In Fiske, D. (Ed.), *New directions for methodology of social and behavioral science: Problems with language imprecision.* (No. 9). San Francisco: Jossey-Bass.

Peterson, C.C., & Peterson, J.L. (1976). Linguistic determinants of the difficulty of true-false test items. *Educational and Psychological Measurement, 36*, 161–169.

Rogers, T.B. (1971). The process of responding to personality items: Some issues, a theory and some research. *Multivariate Behavioral Research Monographs, 6(2).*

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

Rogers, T.B. (1974b). An analysis of two central stages underlying responding to personality items: The self-referent decision and response selections. *Journal of Research in Personality, 8*, 128–138.

Rogers, T.B. (1977). Self-reference in memory: Recognition of personality items. *Journal of Research in Personality, 11*, 295–305.

Schneider-Dueker, M., & Schneider, F. (1977). Untersuchungen zum Beantwortungsprozeß bei psychodiagnostischen Fragebogen. *Zeitschrift für experimentelle und angewandte Psychologie, 24*, 282–302.

Spreen, O. (1963). *MMPI: Saarbrücken. Handbuch zur deutschen Ausgabe des Minnesota Multiphasic Personality Inventory.* Bern: Huber.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D.N., & Beloch-Till, H. (1985). *Deutsche Personality Research Form (PRF).* Göttingen: Hogrefe.

Timm, U. (1968). Reliabilität und Faktorenstruktur von Cattells 16PF-Test bei einer deutschen Stichprobe. *Zeitschrift für experimentelle und angewandte Psychologie, 15*, 354–373.

Van Heck, G.L. (1984). The construction of a general taxonomy of situations. In Bonarius, H.M., Van Heck, G., & Smid, N. (Eds.), *Personality psychology in Europe: Theoretical and empirical developments.* Lisse: Swets & Zeitlinger.

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

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

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