Chapter 2

Assessing the Big Five: Applications of 10 psychometric criteria to the development of marker scales

Gerard Saucier
Lewis R. Goldberg

Introduction

A factor is a parsimonious reduction of many observed variables into one hypothetical variable, accomplished within a particular set of data. The Big Five personality factor structure (Goldberg, 1981; Saucier & Goldberg, 1996a, 1996b) involves five orthogonal (i.e., mutually uncorrelated) factors that capture the five largest sources of variance shared by the variables in fairly representative assemblages of personality-attribute descriptors in a number of languages (e.g., English, German, Polish, Czech, Turkish). Whether the Big Five is the optimal cross-culturally generalizable taxonomic structure for human personality is still a matter of controversy (see Saucier & Goldberg, in press), but it is clearly a very useful structure.

Once a useful set of factors like the Big Five is discovered, it is expedient to extend them beyond the particular set of data in which they were first located. How might one do so? One option would be to readminister the entire set of variables that led to the factors and repeat the analysis in new samples. But in the case of factors based on large numbers of variables (like the Big Five), this is quite inconvenient. Instead, it would be desirable to discover a relatively small set of variables that will consistently produce the structure a set of factor "markers."

This chapter describes various marker sets developed by the authors for the Big Five and related structures. We present these marker sets within a broader conceptual framework, reviewing 10 diverse psychometric criteria by which marker sets can be developed and evaluated. Because constructing a set of factor markers is typically an item-reduction exercise (i.e., selecting an optimal set of items from a

Big Five Assessment, edited by B. De Raad & M. Perugini. © 2002, Hogrefe & Huber Publishers.

larger item pool), we focus on the item selection process. The principles and issues we discuss are important to personality-test construction in general, in most cases applying also to scales that are not factor-analytically derived as are Big Five scales.

To conserve space, we will generally provide a summary of our scale development procedures, and then refer the reader to published articles on these marker sets; in the case of our new unpublished marker sets we will provide more detail.

Item phrasing: Two widely recognized basics

Criterion 1: Clearly understandable items

Unless one has an explicit interest in collecting responses to ambiguous stimuli (as in projective instruments), the meaning of an item should be relatively unambiguous to the respondents. Assuming one's stimuli include words, a clear and easy to understand item is one that uses familiar rather than difficult vocabulary, and simple phrases lacking conjunctions which make items "double-barrelled." Indeed, from this standpoint single words (e.g., adjectives) might be considered superior, but they have one important limitation. Single words are often polysemous (i.e., they have multiple meanings) and thus somewhat ambiguous; a good item subdues rather than aggravates this tendency.

Those items that meet various other criteria we describe later, such as being highly associated with other items or having high loadings on factors, tend to be clear and unambiguous ones. At the outset of item selection, however, the investigator may save much time and effort by identifying and eliminating the most unclear and difficult items. In lexical studies that have led to the Big Five structure, this elimination process has been built into the initial process of reducing the number of variables from thousands to hundreds in preparation for data collection.

Criterion 2: Balanced keying

Imagine that all of the items indexing an attribute were formulated so that the keyed response (the one that contributes to a high rather than a low score) involved the same response option (e.g., "True" rather than "False") or were at the same end of a rating scale. In this case, the content of the scale will be inextricably confounded with individuals' preferences to use one or the other end of the rating scale (i.e., response "acquiescence"). In general, each of the scales in an optimal marker set should have an equal number of items representing the presence of an attribute and

either its opposite attribute, if possible, or its absence[1]. Balanced-keyed scales possess one kind of desirable method-heterogeneity (Nunnally & Bernstein, 1994, p. 313). As we shall see, the balanced-keying desideratum presents a challenge in some domains (e.g., Neuroticism) where it is difficult to find or formulate a large number of candidate items representing the lack (or opposite) of that particular attribute.

The importance of balanced keying suggests that optimal measures should be developed not from single items but rather from parcels of items, each parcel consisting of an equal number of items keyed in each direction. Because acquiescence would then not contribute significantly to factors, factor analysis of parcels should be preferable to analyses of single items. One could use the sum of responses to pairs of opposite-keyed items (with scores not reflected) as an index of acquiescence; this index may be a useful covariate, inasmuch as acquiescence variance can affect the factor structure (Hofstee, Ten Berge, & Hendriks, 1998). Without balanced keying, acquiescence is likely to be confounded with item content and with social-desirability responding (Hofstee *et al.*, 1998). Marker sets described by Saucier (2000b), described later in this article, make use of parcels.

Incorporating desirable elements of diverse scale-construction strategies

Goldberg (1972; Hase & Goldberg, 1967) described three general strategies of test construction, labeled Intuitive, Internal, and External. In the Internal (or factor-analytic) strategy, items loading most highly on a factor are selected for the scale measure of the factor. Only the internal structure of the initial item pool determines item selection and keying direction, although the labeling of the scales developed by this strategy rests on the test constructor's personal judgment. Because marker sets are by definition based on factors, it is the Internal strategy that we will emphasize in this chapter. However, the Internal strategy used alone can lead to limitations in the resulting scales. Elements of the two other strategies can make an incremental contribution to marker set construction, as seen in our next two criteria.

[1] At a more technical level, it should be apparent that balanced keying will not guarantee perfect balance between the two types of items, since the correlations among the items of each type will affect the variance associated with each of those half-scales; differences in these correlational patterns can lead to differences in the relative weights of the half-scales in the composite measure. Nonetheless, balanced keying will generally provide at least some rough control of this problem.

Criterion 3: Intuitive fit between item and construct

Goldberg (1972) noted that:

... the very characteristic of both the External and Internal strategies that gives them their power also provides their Achilles' Heel: namely, their dependence upon — and vulnerability to — characteristics of the particular samples used in their construction. The Intuitive strategy, in contrast, is minimally dependent on sample-specific characteristics; only at the stage of scale "purification" (e.g., discarding items with low correlations with scale scores) do sample characteristics have any chance to enter the scale construction process (p. 49-50).

Goldberg (1972) found that Intuitive scales, those developed solely from judgments about the item content, turned out to be of comparable validity to scales developed by other strategies, a finding replicated by Burisch (1978; see Burisch, 1984a). Indeed, Ashton and Goldberg (1973) found that the average psychology student was able to construct scales as reliable and valid as well-known External scales constructed by a far more expensive and time-consuming process.

What gives the Intuitive approach its strength? Ashton and Goldberg (1973) noted that face validity and empirical validity should converge when there are conditions of mutual trust between subjects and investigators, as in typical self-reports under anonymous conditions (though not necessarily when there is something to gain by deception). Under such research conditions, it has long been known that the more directly the content of the items corresponds to the content of the construct, the better is the measure; and alternatively the more "subtle" are the items (in terms of the scoring keys), the less robust are those items across different subject samples and assessment contexts (Goldberg & Slovic, 1967; Jackson, 1971; Norman, 1963).

What's the take-home message for developers of marker scales? There is much to gain by ensuring that the items relate to one's intuitive or theoretical understanding of the content of the dimension in question. Items that do not have this relation are more prone to be reflecting artifacts, or chance characteristics of the sample at hand. Thus, the Intuitive approach provides some assurance against faulty reliance on sample-specific characteristics.

Criterion 4: Suitable bandwidth

Hase and Goldberg (1967) described the External strategy as one in which the items are selected on the basis of their associations with some external criterion (e.g., peer ratings, job performance). In one version of this strategy, the test constructor initially attempts to locate two distinct groups of subjects who differ in some significant manner (e.g., schizophrenics vs. normals, lawyers vs. people in general, males vs. females) or who fall at each of the two poles of a personality trait (as determined, for example, by peer ratings). The test items are then administered to members of

both criterion groups, and those items that differentiate most strongly between the groups are retained for the scale. In the pure form of this strategy, only the empirically discovered discriminating power of the item determines item selection for a scale, and the scale is typically labeled in terms of the criterion groups used. Common characteristics of scales developed from the External strategy are their heterogeneity in content, which results in rather low intercorrelations among the items; to ensure high Alpha coefficients for the resulting scales, the scales must be quite long.

As already noted, two studies (Hase & Goldberg, 1967; Ashton & Goldberg, 1973) failed to find any validity advantages for External scales. However, Goldberg (1972) found that the External strategy appears to produce a broader bandwidth instrument — that is, one valid for a broader array of criteria — though one with slightly lower fidelity for the most predicable criteria. The slight advantage in bandwidth must be due to these less homogeneous scales including some personologically relevant variance that is not included in the scales developed by the Internal and Intuitive strategies. Although we do not typically advocate the use of the External strategy, these findings suggest an important caution for developers of marker sets. To the extent that an investigator seeks to maximize homogeneity, he/she may be unknowingly compromising validity, especially with respect to any additional criteria beyond those that may be originally anticipated (Loevinger, 1954).[2]

One application of the External strategy of scale construction that has been used to develop marker scales relies on the selection of items with particularly strong correlations between self and peer descriptions of the target person. For example, in the development of their Five-Factor Personality Inventory (FFPI) Hendriks, Hofstee, and De Raad (1999) used self-peer agreement as a primary (external) criterion for item selection. To the extent that marker items selected using this criterion are not particularly univocal indicators of the factors they were selected to approximate, this strategy can lead to high inter-scale associations, as is true of the FFPI scales.

Representative sampling is an alternative approach that promotes the selection of content with broad bandwidth. Loevinger (1957) suggested that in any measure "the various areas or subareas of content should be represented in proportion to their life-importance" and noted that Cattell, an early advocate of the lexical approach, "assumed that life-importance could be judged by dictionary representation" (p. 659). Representative sampling of items from some domain of content is no different in principle from representative sampling of subjects from some population of interest. In both cases, one must select the strata, regions, or facets that one wants to sample, and then one selects the individual persons or items within each class on a quasi-random basis. For the representative sampling of items, one may attempt to include representatives of as wide a range of variables as one can locate. To the extent to which one can locate a full range of facets in the domain, one can sample broadly,

[2] In this chapter, we focus on bandwidth at the scale level. Broad item-level bandwidth (i.e., the extent to which a single item captures a broad array of content) might lead to an increase rather than a decrease in scale homogeneity. Broad items might be constructed using broad, familiar descriptive concepts (e.g., Is good, Is attractive) or, more problematically given our Criterion 1, by joining several forms of content by conjunctions (e.g., Is kind and generous and humble).

and thus one's measure should be associated with a wider range of potential criterion variables.

Representative sampling is consonant with the goal of "content validation." Content validation is appropriate to situations where two conditions hold: (a) validity depends greatly on the adequacy with which a specified domain of content is sampled and (b) the measure must stand by itself as an adequate measure of what it is supposed to measure, with no ultimate gold-standard criterion ever likely to become available to serve in its validation. Content validation generally involves reference to a standard source or to relatively objective expert views (which might be represented in the literature). One example of an attempt to provide a representative sample of constructs from the scientific literature is the set of six facet scales targeted at each of the five factors included in the NEO-PI-R (Costa & McCrae, 1992).

A study by Peabody (1987) exemplifies a representative-sampling approach. He began with an item pool of 571 personality adjectives derived from previous research (e.g., Goldberg, 1982; Norman, 1967). These terms were reduced systematically to a set of 53 bipolar pairs, which were included as a representative set in the studies by Peabody and Goldberg (1989). Similarly, Saucier (in press) developed 100 representative parcels based on 500 very high frequency adjectival person-descriptors in English. Pairs of terms whose highest correlation (positive or negative) was with each other were supplemented by additional terms as needed to increase Alpha. A marker set derived from Saucier's representative set of parcels is described later in this chapter.

Goldberg's (1990) 133 clusters provides another illustration of Big Five factor markers based on representative sampling. The starting point was a set of 1,431 personality adjectives that Norman in unpublished research had classified into 75 categories. Using criteria of (a) lexicographically documented synonymity and (b) relatively homogeneous social-desirability values, Goldberg (1990, Study 2) reduced the terms to 479, grouped into 133 clusters[3]. In a further study leading to a revised set of 100 clusters, Goldberg used perhaps the most common item-selection criterion — that of internal consistency — our next topic.

[3] The criterion of homogeneous desirability values is related to a criterion of homogeneous response means (similar desirability values tend to lead to similar means), and is highly compatible with the criterion of maximizing internal consistency. This is because variables with similar means have a higher maximum intercorrelation than do variables with differing means. However, this criterion is not harmonious with aspects of modern test theory that critique "parallelism" (e.g., sets of nearly redundant items) and put a premium on scales whose items have a wide range of difficulty levels (with the response-mean parameter being analogous to difficulty level). One could argue, as we do later, that there is advantage in using the short homogeneous parcel (rather than the item) as a basic unit, and aggregating parcels having a range of difficulty levels (response means) into the marker scale.

Criteria consonant with classical test theory

Criterion 5: Maximizing internal consistency

Virtually all psychologists have been taught that a desirable feature of any measure is high reliability (the relative absence of measurement error). Internal consistency is the most commonly employed form of reliability, and it is typically estimated by Coefficient Alpha. Alpha is a function of scale length (the longer the higher) and homogeneity (the average intercorrelation among the items in a scale).[4] Naturally, then, maximizing Coefficient Alpha has been widely used as an item-selection criterion. Using contemporary data-analysis software, one can easily identify and cull out those items whose corrected item-total correlations are sufficiently low that their removal increases the value of Alpha in the sample under study.[5] We will refer to this common strategy as "Alpha-maximizing."

Application: Goldberg's (1990) 100 custers

Goldberg's (1990, Study 3) creation of 100 clusters illustrates the Alpha-maximizing approach. The 133 clusters developed in Goldberg's Study 2 included a

[4] There appear to be two prime reasons for the relative emphasis on internal consistency over retest stability: (a) unlike internal consistency coefficients, retest stability coefficients are influenced by extraneous elements like carryover effects and practice effects as well as the length of time between measurements; (b) the assumption that all personality attributes must ideally be stable over time can be questioned. However, one could conceivably use retest stability as an item selection criterion, particularly if one wished to favor stable over unstable attributes. Obviously, for attributes that are temporary states (e.g., emotions) retest stability would be expected to be moderate. If one focuses on internal consistency, an alternative to Alpha is Omega (Zinbarg, Yovel, Revelle, & McDonald, 2000) which is a function of average general factor loading more than average intercorrelation.

[5] In the simplest IRT (item response theory) model, the one-parameter-logistic (1PL) model, item difficulty levels are allowed to vary, but item-discrimination indices are constrained to be equal. The nearest equivalents in classical measurement approaches to IRT item-discrimination indices are either the corrected item-total correlation or the loading of an item on a factor that represents the attribute. Accordingly, a good scale under the 1PL model has some analogy to a scale in which all the items have similarly high corrected item-total correlations, or similarly high loadings on a common factor. A scale whose items discriminate equally well has some distinct virtues. When these items are subjected to reliability analysis, it will be impossible to improve the internal consistency of the scale by deleting any item. Moreover, if these items were subjected to a factor analysis, they would have the maximum possible tendency to cluster together on a single factor, both in replications as well as in the original sample.

few single-term categories (of uncertain reliability) and a few clusters with low internal consistency coefficients. In new samples of data using the 479 adjectives included in the 133 clusters, Goldberg eliminated the single-term categories, and iteratively eliminated the least homogeneous items from a number of other clusters (i.e., those with the lowest item-total correlations); a few new synonym sets were developed from the items that were no longer included in the remaining clusters (again based mainly on the Alpha-maximizing criterion). The result was a set of 100 clusters based on 339 adjectives.[6]

As would be expected, the Alpha values of the new 100 clusters were higher than those from the initial set of 133 scales. In general, Alpha values based on original responses are higher than those based on ipsatized responses (i.e., cases Z-scored so that each respondent has a mean of 0 and a variance of 1 across all of the items), because individual differences in response biases tend to increase indices of internal consistency. Using ipsatized data, the Alpha values for the 100 clusters averaged .61, as compared to .48 for the initial 133 scales. The items included in these 100 clusters, along with the scale reliabilities, are available in an earlier report (Goldberg, 1990, Table 3).

The 100 clusters have the advantage of a high degree of representative sampling: A wide array of personality attributes is included. Another advantage is that the clusters can be used as lower-level facets in their own right, affording an abundance of information for predictive purposes. One disadvantage is that no cluster included reverse-keyed items. Another is that 339 adjectives must be administered in order to provide scores for the Big Five factors. Moreover, one has no guarantee that the five derived factor scores would be the same in different subject samples, partly because some of the clusters have complex associations with the Big Five factors, and some have only weak relations with any of them. In other words, these are a richly detailed but not an efficient set of factor markers. In subsequent work, Goldberg (1992) developed more efficient (albeit less representative) marker sets by applying two other widely used item-selection criteria, to which we now turn.

Criterion 6: Factor saturation (high loadings on the targeted factor)

This criterion is central to the Internal strategy of scale construction (Goldberg, 1972). The rationale is clear-cut. A set of markers for a factor is designed to represent that factor. What more efficient way to represent the factor than with the items that have the highest loadings on (or extension correlations with) that factor?

High internal consistency is a necessary but not a sufficient indicator of unidimensionality (Zinbarg, Yovel, Revelle, & McDonald, 2000). However, Criteria 5 (Internal consistency) and 6 (Factor saturation) generally tend to converge, most strongly so in the special case of a structure that properly has only one factor. The first unrotated factor has the maximum internal consistency of any possible linear

[6] However, eliminating items based on a internal consistency criterion can capitalize on chance in a way similar to stepwise regression analyses, and thus both procedures share this potential liability.

combination of the items analyzed. Factor scores on this single factor are highly related to a scale derived from the highest loading items, and thus either can be used to index individual differences on that factor. MacDonald (1999) notes that a "(psychometrically) homogeneous test is one whose items measure just one attribute in common — a common factor," a supposition that can be tested by "seeing if the responses to them fit the single factor model" (p. 78). In a simplified scale-construction approach, one might merely utilize Criteria 5 and 6, which are already highly convergent. Our next criteria can each be seen as a way of correcting for the limitations of this simplified approach.

Criterion 7: Factor discrimination (low loadings on other factors)

In a factor structure like the Big Five that includes more than one factor, those variables having high loadings on each factor can be distinguished from one another by the relative magnitude of their loadings on the other factors. Some variables may have high loadings on one or more of the other factors; variables with such high "complexity" with respect to the factors may be viewed as factorial "blends" or as "interstitial" variables. Some variables may have low loadings on all of the other factors that are retained; variables with such extreme "simplicity" with respect to the factors might be thought of as factorially "univocal" or as "factor-pure." This low-divergent-loadings criterion tends to enhance the unidimensionality of the set of variables associated with each factor.

If one combines Criteria 6 (Factor saturation) and 7 (Factor discrimination), one selects those items that simultaneously load most highly, and most univocally, on a given factor. These are the variables that conform most directly to the factor rotation concept of "simple structure," as they capture the distinct features of the factor most directly. The most efficient set of factor markers are these univocal variables, and they are the ones that are most likely to produce across-sample replicability of the factor structure. That is, factor markers consisting of univocal items can be expected to reproduce the original factor structure in replication studies more surely than those with either low or complex associations with the factors.

We note one caveat about univocal variables, to which we shall return later: Other things being equal, the more univocal are the variables included in a marker set, the more homogeneous and narrow are they likely to be. But, as indicated by our Criterion 4, the bandwidth of one's markers constitutes an important property of any marker set. To the extent to which one desires a broad-bandwidth instrument, perhaps by the representative sampling of lower-level facets, one will usually go beyond extremely univocal variables, which can be homogeneous but narrow in content reference.[7]

[7] Criterion 7 can be related to the Stylistic strategy of scale construction (Hase & Goldberg, 1967) as exemplified in the development of social-desirability scales. Big Five scales tend to have their favorable poles associated, suggesting that the intercorrelations reflect a single desirability factor. Thus, the "low loadings" criterion tends to minimize the influence of this stylistic factor.

Application: Goldberg's (1992) 100 unipolar markers

Although the 100 clusters from Goldberg (1990) have the advantage of providing markers with an unusually broad bandwidth, they require the administration of 339 adjectives, and thus they are hardly a maximally efficient marker set. In constructing some more efficient sets of Big Five markers, Goldberg (1992) took into account all of the criteria that we have heretofore discussed. Among the marker sets described in that article was a set of 100 unipolar factor markers that has since become widely used.

From a pool of 566 reasonably common personality-descriptive adjectives, Goldberg selected 116 terms having high loadings on one factor and, relative to other candidate terms for that factor, relatively low loadings on the other factors (Criteria 6 and 7). The initial set of 116 terms was reduced to 100 using the internal-consistency criterion (Criterion 5), as well as a criterion of replicability in their factor loadings across three samples of subjects. In order to make the marker sets relatively equal in size, 20 terms were selected for each factor.[8] In order to reduce the effect of individual differences in response scale usage, 10 items were selected for the positive and 10 for the negative pole of each factor, with the exception of Factor IV (Emotional Stability) where a dearth of suitable positive items led to a mix of 6 positive and 14 negative items.

Goldberg (1992) showed that each of the five 20-item subsets of these 100 markers, when considered as separate scales, yielded highly reliable scores: The mean (across factors) Alpha coefficients ranged from .85 to .93 depending on the data set, with all of the coefficients above .80 in each data set. Partly because of these favorable psychometric properties, this marker set has been widely used to index the Big Five factors. However, there are at least three potential problems with this marker set, each of which has led to the development of different types of new markers. First of all, many investigators desire to include some markers of each of the Big Five domains in an extensive battery of other measures, but they balk at devoting the testing time needed to administer all 100 items; for these purposes, shorter and thus more efficient marker sets are needed. In addition, although the Big Five factors are orthogonal conceptually (and when operationalized via orthogonal rotations), the five scales scored from the 100 markers are typically at least slightly interrelated; less highly related marker sets would be desirable in some contexts. These problems are addressed by additional criteria for item-selection that we will discuss shortly.

Another problem with Goldberg's (1990) 100 markers involves the nature of the items themselves. Because single trait-descriptive adjectives encode behaviors at such a high level of abstraction, they are often difficult to translate precisely from

[8] Note that if one's goal was an even more complete equality of the scale variances, one might have had to select slightly different numbers of items for each of the five factors.

one language to another. That is, although it is often possible to locate a term in each of the two languages that refers to much the same type of behavior, the two terms may differ in their social-desirability value (Hofstee, 1990). More behaviorally specified item formats (such as the items included in the International Personality Item Pool; Goldberg, 1999) could turn out to be far easier to translate with precision. One way of addressing this problem is illustrated in the following application.

Applying these criteria to an alternative item format

Goldberg (1999) reported the development of an Internet collaboratory for the advancement of personality measurement, based on an item format pioneered by Hendriks, Hofstee, and De Raad (1999). This International Personality Item Pool (IPIP) now includes nearly 2,000 items, each a short verbal statement describing some aspect of one's thoughts, feelings, or behaviors (e.g., Act wild and crazy; Don't care about rules; Sense others' wishes; Have a soft heart). Preliminary personality scales have been developed from the IPIP to measure the 45 bipolar facets from the Abridged Big Five-dimensional Circumplex model (AB5C) of Hofstee, De Raad, and Goldberg (1992). Table 1 lists the number of items keyed in each direction, the mean item intercorrelation, and the Coefficient Alpha reliability estimate for each of these 45 AB5C marker scales; the items included in each scale are listed in Goldberg (1999). Most of these scales include about 10 items, with mean intercorrelations around .25 and Alpha coefficients around .80.

The coordinates for the AB5C model were based on Goldberg's (1992) 100 unipolar markers. Therefore, the 45 bipolar IPIP-AB5C facet scales can be regarded as a translation of an adjectival Big Five marker set into a more behaviorally specified item format, by means of an application of "uniform" sampling. Uniform sampling of a semantic space is a characteristic of the "circumplex" tradition (e.g., Wiggins, 1980) in which the locations of variables in two dimensions are projected onto a circular representation, and then exemplar items are selected at equally spaced locations around the circle. In uniform sampling, regions where variables are densely concentrated are systematically undersampled whereas more sparsely populated interstitial regions are oversampled (Goldberg, 1992), thus contrasting markedly with representative sampling.

As markers of the broad Big Five domains, one could use the five IPIP scales measuring the factor-pure AB5C facets. However, each of these IPIP scales includes only items that are more highly associated with their narrow facet than with any of the other facets, and such items may not necessarily be optimal measures of the five domains alone. Moreover, those scales are targeted at the Big-Five factor structure of phenotypic personality attributes (Saucier & Goldberg, 1996b), not at McCrae and Costa's (1996) Five-Factor Model of personality traits, which differ to some degree in how the factors are conceptualized. Some investigators may prefer to measure the constructs in the latter model rather than (or in addition to) those in the former one. Consequently, we have developed IPIP-based measures of both models.

Table 1. Characteristics of the 45 preliminary IPIP scales targeted at the AB5C facets

AB5C Facet	Provisional Label	No. of Items	Mean Item r	Coef. Alpha
Factor I				
I+/I+ vs. I-/I-	Gregariousness	4 + 6 = 10	.34	.83
I+/II+ vs. I-/II-	Friendliness	5 + 5 = 10	.37	.85
I+/III+ vs. I-/III-	Assertiveness	9 + 3 = 12	.20	.75
I+/IV+ vs. I-/IV-	Poise	5 + 5 = 10	.31	.82
I+/V+ vs. I-/V-	Leadership	5 + 5 = 10	.31	.82
*I+/II- vs. I-/II+	Provocativeness	8 + 3 = 11	.19	.72
I+/III- vs. I-/III+	Self-Disclosure	8 + 2 = 10	.26	.78
I+/IV- vs. I-/IV+	Talkativeness	8 + 2 = 10	.35	.84
*I+/V- vs. I-/V+	Sociability	3 + 7 = 10	.16	.66
Factor II				
II+/II+ vs. II-/II-	Understanding	5 + 5 = 10	.30	.81
II+/I+ vs. II-/I-	Warmth	9 + 2 = 11	.33	.84
*II+/III+ vs. II-/III-	Morality	5 + 7 = 12	.18	.73
II+/IV+ vs. II-/IV-	Pleasantness	6 + 6 = 12	.22	.76
*II+/V+ vs. II-/V-	Empathy	5 + 4 = 9	.20	.70
*II+/I- vs. II-/I+	Cooperation	2 +10 = 12	.18	.73
*II+/III- vs. II-/III+)	Sympathy	6 + 6 = 12	.20	.74
*II+/IV- vs. II-/IV+	Tenderness	9 + 4 = 13	.18	.74
II+/V- vs. II-/V+	Nurturance	6 + 7 = 13	.16	.71
Factor III				
III+/III+ vs.III-/III-	Conscientiousness	6 + 7 = 13	.19	.75
III+/I+ vs. III-/I-	Efficiency	5 + 6 = 11	.30	.83
*III+/II+ vs. III-/II-	Dutifulness	6 + 7 = 13	.21	.78
III+/IV+ vs. III-/IV-	Purposefulness	5 + 7 = 12	.27	.81
III+/V+ vs. III-/V-	Organization	9 + 3 = 12	.23	.78
*III+/I- vs. III-/I+	Cautiousness	5 + 7 = 12	.21	.77
*III+/II- vs. III-/II+	Rationality	8 + 6 = 14	.13	.67
III+/IV- vs. III-/IV+	Perfectionism	7 + 2 = 9	.26	.76
III+/V- vs. III-/V+	Orderliness	7 + 3 = 10	.27	.78
Factor IV				
IV+/IV+ vs. IV-/IV-	Stability	5 + 5 = 10	.37	.86
IV+/I+ vs. IV-/I-	Happiness	5 + 5 = 10	.34	.84
IV+/II+ vs. IV-/II-	Calmness	4 + 6 = 10	.33	.83
IV+/III+ vs. IV-/III-	Moderation	4 + 6 = 10	.24	.76
IV+/V+ vs. IV-/V-	Toughness	4 + 8 = 12	.29	.84
IV+/I- vs. IV-/I+	Impulse Control	2 + 9 = 11	.24	.78
IV+/II- vs. IV-/II+	Imperturbability	2 + 7 = 9	.37	.84
*IV+/III- vs. IV-/III+	Cool-headedness	0 +10 = 10	.21	.73
*IV+/V- vs. IV-/V+	Tranquility	7 + 4 = 11	.22	.76
Factor V				
V+/V+ vs. V-/V-	Intellect	6 + 5 = 11	.27	.81
V+/I+ vs. V-I-	Ingenuity	6 + 3 = 9	.37	.84
*V+/II+ vs. V-/II-	Reflection	8 + 2 = 10	.26	.75
*V+/III+ vs. V-/III-	Competence	8 + 0 = 8	.26	.74
V+/IV+ vs. V-/IV-	Quickness	7 + 3 = 10	.37	.84
*V+/I- vs. V-/I+	Introspection	10 + 2 = 12	.18	.71
V+/II- vs. V-/II+	Creativity	5 + 5 = 10	.30	.81
V+/III- vs. V-/III+	Imagination	5 + 5 = 10	.27	.78
*V+/IV- vs. V-/IV+	Depth	7 + 2 = 9	.27	.77
Mean			**.26**	**.78**

Note: All analyses are based on the responses of 501 adult subjects from the Eugene-Springfield Community Sample; These scales have been augmented with items from other AB5C facets.

Specifically, we have developed both 50-item (10 items per domain) and 100-item (20 items per domain) scales to measure the five domains in each of the two models (Goldberg, 1997).

Over 500 adult participants from a community sample completed both the 240-item NEO-PI-R and an initial set of 1,252 IPIP items (Goldberg, in press); the IPIP items were administered in three separate questionnaires over a three-year period of time, and the NEO inventory was administered on another occasion during the same time period. Each of the participants had previously completed an inventory of 360 trait-descriptive adjectives which included Goldberg's (1992) 100 unipolar Big-Five factor markers. The five orthogonal factor scores from the 100 markers (based on ipsatized data) served as the criteria for the Big-Five constructs, and scores on the five 48-item domain scales from the NEO-PI-R served in that role for the Five-Factor (NEO) model. Responses to all of the IPIP items were first correlated with each of the criterion indices, and the items were then categorized by their highest correlations. Initial scales were developed using the most highly related items in each category, and if necessary these scales were then refined by internal consistency analyses (Criterion 5, maximizing Alpha).

Table 2 presents some characteristics of the new IPIP scales for the Big-Five domains, including the number of positively and negatively keyed items in each scale, its mean item intercorrelation, its Coefficient Alpha reliability estimate, and its correlation with the orthogonal factor scores derived from the Big-Five adjective markers. On average, the shorter scales had a mean item intercorrelation of .34, an Alpha of .84, and a correlation of .67 with the factor markers (.81 when corrected for unreliability); the longer scales had a mean item intercorrelation of .31, an Alpha of

Table 2. Characteristics of the Preliminary IPIP Scales Measuring the Big Five Domains

Big Five Domain	Number of Items	Mean Item Intercorrelation	Coefficient Alpha	Correlation with Markers
Shorter Scales				
I. Extraversion	5 + 5 = 10	.40	.87	.73 [.84]
II. Agreeableness	6 + 4 = 10	.31	.82	.54 [.66]
III. Conscientiousness	6 + 4 = 10	.29	.79	.71 [.90]
IV. Emot. Stability	2 + 8 = 10	.38	.86	.72 [.84]
V. Intellect	7 + 3 = 10	.34	.84	.67 [.80]
Total/Mean	26 +24 = 50	.34	.84	.67 [.81]
Longer Scales				
I. Extraversion	10 +10 = 20	.34	.91	.76 [.84]
II. Agreeableness	14 + 6 = 20	.28	.88	.57 [.65]
III. Conscient.	11 + 9 = 20	.27	.88	.74 [.84]
IV. Emot. Stability	4 +16 = 20	.35	.91	.74 [.81]
V. Intellect	13 + 7 = 20	.32	.90	.69 [.77]
Total/Mean	52 +48 = 100	.31	.90	.70 [.78]

Note: Values in brackets are correlations corrected for unreliability; these may be underestimates, given that the reliabilities of the factor markers were assumed to be the same as those of their corresponding IPIP scales.

.90, and correlated .70 with the markers (.78 when corrected). The items included in each of these new Big-Five scales are provided on the IPIP Website.

Table 3 presents the corresponding values for the new IPIP scales measuring the constructs in the Five-Factor (NEO) model, including the correlations with the 48-item NEO domain scales. On average, the shorter scales had a mean item intercorrelation of .33, an Alpha of .82, a correlation with the NEO domains of .77 (.90 when corrected); the longer scales had a mean item intercorrelation of .30, an Alpha of .89, and a mean correlation of .81 (again .90 when corrected). The items included in each of these new FFM scales are also provided at the IPIP Website.

There are no common items among the scales within each scale set, although all of the items in the 10-item scales are included in their 20-item counterparts. The part-whole correlations between the shorter and the longer scales were: .95, .94, .95, .96, and .96 for the Big Five constructs, and .95, .92, .96, .95, and .96 for the Five-Factor (NEO) constructs, both sets in Big Five order. The average of the intercorrelations among the scales based on the Big-Five constructs, presented in Table 4, were very slightly lower than for those based on the Five-Factor (NEO) constructs. When corrected for attenuation due to the scale unreliabilities, the across-set convergence was essentially perfect (r = 1.00) for the Extraversion, Conscientiousness, and Emotional Stability (Neuroticism) constructs; the corrected correlations for the Agreeableness scales were .79 (.84) and for Intellect/Openness they were .83 (.86).

Table 3. Characteristics of the preliminary IPIP scales measuring the NEO domain constructs

NEO Domain	Number of Items	Mean Item Intercorrelation	Coefficient Alpha	Correlation with NEO
Shorter Scales				
I. Neuroticism	5 + 5 = 10	.37	.86	.82 [.92]
II. Extraversion	5 + 5 = 10	.38	.86	.77 [.88]
III. Openness	5 + 5 = 10	.33	.82	.79 [.91]
IV. Agreeableness	5 + 5 = 10	.27	.77	.70 [.85]
V. Conscientiousness	5 + 5 = 10	.31	.81	.79 [.92]
Total/Mean	25 +25 = 50	.33	.82	.77 [.90]
Longer Scales				
I. Neuroticism	10 +10 = 20	.33	.91	.86 [.93]
II. Extraversion	10 +10 = 20	.35	.91	.79 [.88]
III. Openness	10 +10 = 20	.29	.89	.83 [.92]
IV. Agreeableness	10 +10 = 20	.23	.85	.78 [.90]
V. Conscientiousness	10 +10 = 20	.31	.90	.80 [.88]
Total/Mean	50 +50 = 100	.30	.89	.81 [.90]

Note: Values in brackets are correlations corrected for unreliability. The Coefficient Alpha reliability values for the 48-item NEO domain scales were: N = .93; E = .89; O = .91; A = .89; and C = .91.

Table 4. Correlations among and between the preliminary IPIP scales measuring the domain constructs from the Big Five and the Five Factor (NEO) Models

	I/E	II/A	III/C	IV/N	V/O
I/E	.93 (.96)	.28 (.39)	.07 (.17)	.18 (.27)	.35 (.40)
II/A	.15 (.22)	.63 (.73)	.11 (.17)	.21 (.23)	.17 (.18)
III/C	.24 (.28)	.22 (.21)	.81 (.87)	.15 (.15)	.03 (.07)
IV/N	-.35 (-.38)	-.43 (-.41)	-.36 (-.40)	-.89 (-.93)	.13 (.20)
V/O	.36 (.37)	.11 (.09)	.01 (.05)	-.08 (-.09)	.69 (.77)

Note: Correlations among the IPIP Big Five domain scales are presented above the main diagonal, correlations among the IPIP Five Factor (NEO) domain scales are below the diagonal, and correlations between the corresponding scales in each set are listed in bold in the diagonal. (Correlations based on the 20-item scales are listed in parentheses after the values for the 10-item scales.) Factor I = Extraversion; Factor II = Agreeableness; Factor III = Conscientiousness; Factor IV = Emotional Stability (versus Neuroticism); and Factor V = Intellect/Openness to Experience.

Criterion 8: Scale brevity, or keeping it short and sweet

In measurement, brevity imparts efficiency, and thus brevity is generally desirable (Burisch, 1984b). We noted the value of item brevity with respect to our first criterion, but Criterion 8 addresses scale brevity. For some research, teaching, and assessment purposes, even a 100-item inventory, such as the marker set from Goldberg (1992), is too lengthy. However, because any abbreviated measure almost inevitably suffers from a loss of reliability compared to the full measure, there is a recurring cost involved in the creation of a "short form" of a longer measure. To minimize this cost, one must attempt to conserve internal consistency while culling items. By doing so, however, one could easily precipitate a decline in validity even though Alpha is relatively constant, because the scale is being made overly narrow and homogeneous (Loevinger, 1954). Smith, McCarthy, and Anderson (2000) discuss other potential problems in short-form development, stressing that short forms (a) be developed only on well-validated measures, (b) preserve the content coverage and subfactors of the longer form, (c) protect reliability, (d) demonstrate overlapping variance with the longer form when administered independently, (e) show a factor structure similar to that of the longer form, (f) have demonstrated validity and high correct classification rates in independent samples, and (g) show meaningful savings in time or resources.

What is the absolute minimum number of items that should constitute a scale? *One* item is certainly too few; internal consistency is not easily estimated and balanced keying is impossible. On the other hand, in unusual cases where the construct being measured is highly familiar (or "schematized") to respondents, unidimensional, and primarily subjective in content, one item could be adequate (Robins *et al*, 2001). Although *two*-item scales have neither of the disadvantages of one-item scales, internal consistency tends to be purchased at the cost of extreme narrowness of breadth. With *three*-item scales, unbalanced keying is again a problem. Thus, *four*-item scales seem to be a practical minimum in most cases. Such mini-scales have been referred to as "testlets," "item parcels," "homogeneous item composites," "factored homogeneous item dimensions," and the like; and they have been used as

the basic building blocks for longer scales by Comrey (1988), Hogan and Hogan (1995), and Saucier (2001; in press).

It may be, however, that the fewer items one selects from a large item-pool, the greater is the likelihood that their selection will have capitalized on chance characteristics of the derivation sample, leading to decreased internal consistency in new samples. Moreover, a high Coefficient Alpha in a four-item scale is typically only possible if the content is highly focused and narrow, so marker sets commonly include more than four items. And, internal consistency is not the only reason for including more items. Nunnally and Bernstein (1994, p. 16) suggest that data from single items are ordinal, but aggregates of these items are more readily treated at the interval level of measurement. A scale including 8 or 10 items is likely to generate scores with a more Gaussian distribution than would a scale consisting of only four items.

Application: Saucier's Big Five Mini-Markers

Saucier (1994) scrutinized the performance of each of Goldberg's (1992) 100 unipolar markers in 12 data sets, searching for those items that loaded most highly on the expected factor in virtually all analyses. After selecting an initial set of eight items for each factor based on this "factor purity" criterion, revisions were made to (a) increase user-friendliness by reducing the number of negations beginning with the prefix "un-," (b) decrease the number of root-negation pairs (e.g., Kind-Unkind) so as to lessen any overnarrowing of content, and (c) increase the correlation of the brief scales with the original 100 unipolar marker scales. After 9 such item-substitutions, the final 40-item set included eight items for each factor. In the case of Factors I (Extraversion), II (Agreeableness), and III (Conscientiousness) there were four items for each pole of the factor. In the case of Factors IV (Emotional Stability) and V (Intellect/Imagination) a dearth of suitable terms led to the selection of six terms at one pole (low Emotional Stability, high Intellect) and two at the other pole.

Internal consistency estimates for the five Mini-Marker scales were provided by Saucier (1994) in four data sets. The 20 Alpha coefficients ranged from .69 to .86, averaging around .80; these coefficients were generally about .07 lower than those for the longer 100 markers set. There are indications, however, that validity is comparable with that for the longer marker set (Dwight, Cummings, & Glenar, 1998). As would be expected from scales with lower internal consistency, Saucier (1994) noted that the Mini-Markers had lower inter-scale correlations than did the 100 unipolar markers from which they were derived. For example, the mean inter-scale correlations for the 100 markers which averaged .19 in raw data and .10 in ipsatized data were reduced in the Mini-Markers to .15 in raw data and .09 in ipsatized data. Are inter-scale correlations of this size acceptable? Could they be reduced further by purposeful scale-construction procedures?

Criterion 9: Mutual orthogonality among marker scales

In his critique of the Big Five model, Block (1995) pointed out that although the model is based on orthogonal factors, the five factors are normally operationalized with scales that are at least somewhat interrelated. In self-ratings, Goldberg's (1992) unipolar markers have intercorrelations as high as .37, and in peer ratings as high as .58.[9] Indeed, Digman (1997) was able to develop second-order factors on the basis of the intercorrelations among the scales within various five-factor marker sets. Even in data sets where the average intercorrelation is low, the correlation between a pair of markers can be quite high, and one such high correlation alone is enough to call into question the assumption of five "orthogonal" factors.

Orthogonal factors are not necessarily better than oblique factors. But orthogonal factors are an advantageous feature of the Big Five model for at least two reasons. First, when one is mapping a domain of variables, as when one is mapping a physical landscape, orthogonal axes provide a superior coordinate system for locating points on the map. Second, as Jackson (1971) noted, "if one wishes to maximize the predictability of a battery, entirely uncorrelated tests would be appropriate" (p. 246). Orthogonal predictors are more efficient in multiple-regression analyses because they minimize multicollinearity and maximize discriminant validity.

It has long been known that marker scales based on orthogonal factors are not necessarily themselves mutually orthogonal (e.g., Cattell & Tsujioka, 1964). Recognition of non-orthogonality in marker scales has prompted some statistical remedies, such as (a) the ipsatization of the original response data, which tends to lower scale intercorrelations, and (b) the use of orthogonal (e.g., varimax) factor scores (Goldberg, 1992). Ipsatizing within sets of items that do not have balanced keying with respect to content can lead to inadvertently discarding content variance. Moreover, the most common form of ipsatizing, the use of standard (Z) scores, controls for between-subject differences in spread (variance) as well as central tendency (mean); while this practice has been explicitly recommended by Goldberg (1990; 1992), it has recently been criticized by Hofstee *et al.* (1998).

The most direct method for assuring orthogonality is to use orthogonal factor scores instead of scale scores. One limitation of this procedure is that the factors derived de novo on each occasion are less uniform across samples than are scale scores. Perhaps as a consequence, most users of Big Five markers use simple (but correlated) scale scores based on raw data, eschewing both types of statistical remedies. Accordingly, a close approximation to orthogonality would be a desirable feature in a Big Five marker set.

---

[9] Nor are high inter-scale correlations confined to lexical studies or adjective stimuli. The Revised NEO Personality Inventory (NEO-PI-R) has domain-scale intercorrelations as high as -.53 in self-ratings (Costa & McCrae, 1992).

Table 5. The orthogonal subset of the 100 unipolar markers (Ortho-40): Reliabilities and interscale correlations

	Derivation Samples			Cross-Validation Sample
	Self	Liked Peer	Pooled Peer	Community Sample
Coefficient Alpha				
I	.84	.86	.86	.81
II	.73	.79	.92	.71
III	.86	.87	.89	.85
IV	.70	.62	.70	.72
V	.71	.72	.83	.74
Interscale Correlations				
I-II	.03	.02	-.10	.03
I-III	-.07	-.12	-.08	.04
I-IV	.00	.02	-.17	.05
I-V	.03	.06	-.13	.10
II-III	.11	.15	.19	.12
II-IV	-.05	.03	.30	.19
II-V	.04	.21	.40	.06
III-IV	-.03	.09	.14	.13
III-V	.08	.06	.28	.06
IV-V	-.06	-.04	.10	-.03
Mean Correlation				
Ortho-40	.01	.05	.09	.07
100 Markers	.13	.24	.27	.25
40 Mini-Markers	.11	.18	.26	.22

Note: Sample sizes: Self = 320; Liked Peer = 316; Pooled Peer = 205; Community Sample = 1,125. All analyses used the original (non-ipsatized) response data. The 40 Ortho items: I = Bold, Extraverted, Talkative, Unrestrained vs. Introverted, Quiet, Reserved, Shy; II = Kind, Sympathetic, Undemanding, Warm vs. Cold, Demanding, Harsh, Unsympathetic; III = Efficient, Neat, Organized, Systematic vs. Careless, Disorganized, Sloppy, Unsystematic; IV = Unenvious, Unexcitable vs. Anxious, Emotional, Fearful, Fretful, Nervous, Touchy; V = Artistic, Complex, Creative, Deep, Introspective, Philosophical vs. Simple, Unreflective.

The application of Criterion 7, emphasizing low divergent loadings, tends to suppress inter-scale associations but not necessarily to remove them. If most of the potential marker items for a factor are associated in the same direction with another factor, simply choosing those items with the lowest divergent loadings will not serve to guarantee unrelated marker sets. To remove the inter-scale correlations, one must select marker items whose correlations with each of the other factors are balanced with respect to sign. Then, because scale scores that are uncorrelated in a derivation sample may not be uncorrelated in a new sample, one must demonstrate that the approximation to orthogonality persists when the scale scores are intercorrelated in a new sample.

Application: The "Ortho-40" markers

Table 5 provides an illustration of the results of using orthogonality as a criterion for item selection. The 100 items in Goldberg's set of unipolar markers were scrutinized in the self- and peer-rating data sets used in Goldberg's (1992) Study 4. Items that contributed most to the positive scale intercorrelations were removed until eight

items remained per scale (with some priority given to maintaining balanced keying). This 40-item subset is labeled the Ortho-40. Coefficient Alpha reliability averages about .10 lower than for the 100-marker scales, and about .03 lower than for the Mini-Marker subset (also based on 40 items). But inter-scale correlations are dramatically lower than for either of the other two sets, on average about .15 lower per pair of scales. The highest inter-scale correlations are in the Pooled Peer sample, where the general evaluation factor has a powerful effect on these coefficients; in this extreme case, whereas one correlation in the 100 Markers reached .58 (Factors II and IV), the highest correlation in the Ortho-40 was .40 (Factors II and V). Overall, the Ortho-40 sacrifices some internal consistency in order to gain greater mutual orthogonality. The Ortho-40 subset demonstrates that the Big Five are not oblique by necessity; if one has a sufficiently large item pool, it should be possible to develop a set of marker scales that are virtually unrelated.

Another illustration of the application of Criterion 9 is provided by Saucier's (2000a) new Modular Markers, which have inter-scale correlations that are comparable to those of the Ortho-40, but with higher reliabilities. However, these Modular Markers were developed using an additional criterion, which must first be introduced.

Considerations congruent with newer forms of measurement theory

None of the criteria offered so far are inconsistent with classical test theory (McDonald, 1999). However, from the standpoint of item response theory (e.g., Embretson & Reise, 2000) these criteria, which tend toward maximizing Alpha and homogenizing item difficulties, could lead to scales with a tendency to "parallelism"[10] Strictly parallel items have the same difficulty levels (e.g., mean response) and discrimination (e.g., item-total correlation) parameters; redundant items thus tend to be parallel. A set of relatively redundant items will have a high degree of internal consistency. But a set of such items is problematic because it is likely to be overly narrow, which may decrease validity (see Criterion 4 regarding bandwidth). And it may distinguish well among individuals at one level of the broader construct but not at other levels. For example, a marker scale for Extraversion formed from the three items "Talks too much", "Can't stop talking", and "Chatters away even if no one is listening" might effectively distinguish extreme extraverts from both moderate extraverts and introverts, but would probably do a poor job of distinguishing between the latter two groups. Nonetheless, this set of items should exhibit substantial internal consistency. A peaked, or kurtotic, test maximizes reliability (Lord, 1952), and can be expected to show high levels of consistency across samples in exploratory factor analyses.

[10] Criterion 4, stressing broad bandwidth, is the most likely exception, since a measure of a broad attribute is unlikely to result from a set of redundant items.

Our final criterion is derived from aspects of item response theory (IRT), which has been widely applied to measures of ability and aptitude but has not yet had a major impact on personality measurement. Unfortunately, IRT analyses require relatively large samples and seem better suited to relatively specific, homogeneous content than the heterogeneous constructs of the sort personality psychologists have emphasized (Nunnally & Bernstein, 1994, p. 434). However, without adopting a full-scale IRT approach, one can still borrow at least one important IRT scale-construction criterion.[11]

Criterion 10: Equidiscrimination (discriminating at diverse levels)

A contribution of IRT is its emphasis upon selecting items with a spread of difficulty levels in order to discriminate among (i.e., effectively differentiate) individuals at a variety of levels of the attribute. If one wanted to measure individual differences in the ability to solve arithmetic problems, one would not restrict one's questions to a single level of difficulty (e.g., only addition of single-digit integers, or alternatively only multiplication of twelve-digit numbers). Instead, one would include items covering a range of difficulty levels to allow the measure to discriminate very high ability from moderate ability, and very low ability from mere mediocrity. Tests that include a wide range of item difficulty levels provide more information, and thus have broader bandwidth.

In personality measurement, item difficulty levels index the "difficulty" respondents are likely to have in admitting to, ascribing, or agreeing with the content of the item, as indicated by inter-item variations in the response means. Items that are easy to endorse will tend to discriminate well only between those who are very low and moderately low on the attribute, whereas items that are difficult to endorse will tend to discriminate well only between those who are very high and moderately high on the attribute. Items with more intermediate response-means are prone to discriminate well in the middle of the attribute distribution, but not at either extreme. In most cases, classical item-selection procedures lead to a bias toward selecting items of intermediate difficulty (Nunnally & Bernstein, 1994, p. 329).

One would expect that *any* item selection procedure that works to diversify the content of the selected items will work against parallelism, and in favor of discrimination at diverse levels of the attribute. Thus, five-factor measures like the NEO-PI-R domain scales (Costa & McCrae, 1992) or Johnson's (2000) IPIP-NEO short-form, which build up the score for a factor from subscales with diverse content, are

[11] Minimizing differential item functioning (DIF; also sometimes referred to as item bias) is another scale-construction criterion prominent in IRT deserving of more attention and study with respect to personality measurement. DIF exists whenever two items differ between groups in their parameters (e.g., discrimination, difficulty level). As Nunnally and Bernstein (1994) advise, "one should choose items whose parameters are most similar across groups, whether these parameters are defined classically or through IRT. This is especially true when the groups differ in gender or ethnicity" (p. 417). One might divide one's data into subsamples based on gender or ethnicity and (a) eliminate items with relatively poor discrimination in any subsample, or (b) retain those items that show the smallest differences in parameters between subsamples.

unlikely to be characterized by parallelism. However, the relative discriminatory power of these measures at differing levels of the attribute remains to be demonstrated.

A measure that discriminates well across levels of the latent attribute is most needed when important practical decisions are made about people based on their scores on a measure, and thus highly reliable distinctions at all levels of the attribute are necessary; this criterion would be particularly valuable for any measure that is used in a wide variety of selection situations. To the extent that factor markers are used only for locating other variables, rather than locating individuals, such a criterion may be less necessary. Nonetheless, because markers scales that were originally developed for the purpose of locating variables (e.g., Goldberg, 1992) have then become widely used as measures of individual differences, it may be sensible to incorporate this criterion into marker construction from the onset.

To develop measures that discriminate well at various levels of the latent attribute, Nunnally (1967) proposed a simple item-selection procedure for what he called the equidiscriminating (EQD) test. An EQD measure can be constructed by selecting items based on their characteristics at multiple cutoff levels. On the basis of the frequency distribution of the underlying attribute (e.g., the factor scores for a broad factor), one selects cutpoints between fractions of the distribution. For example, one can select one-third of the items to differentiate the top 25 per cent of the sample from the bottom 75 per cent, another third to differentiate the top half of the sample from the lower half, and a final third to discriminate the bottom 25 per cent of the sample from the top 75 per cent (Nunnally & Bernstein, 1994, p. 330). Then, one selects some items that served best to differentiate individuals above each cutpoint from those below it. There are other ways to reach the same equidiscriminating end result, of course, including procedures specific to IRT, but Nunnally's procedure may be the simplest to implement.

Application: Modular markers

Saucier (in press) created a new set of marker scales for the Big Five, as well as scales for broader structures of one and two factors based on studies of natural-language descriptors. The label "Modular Markers" for these scales is based on the flexible use of item parcels serving in marker sets for the development of scales at more than one hierarchical level. These new scales were constructed so as to simultaneously achieve three major objectives — relative orthogonality (Criterion 9), higher internal consistency (Criterion 5) than was obtained with the Ortho-40, and improved equidiscrimination (Criterion 10) than previous marker sets.

The initial item pool consisted of 100 representative parcels, plus 21 supplementary item parcels, also of two to four adjectives (Saucier, in press). For each of the Big Five factors, the distribution of factor scores based on analyses of personality-

Table 6. The factor loadings of the 32 parcels in the set of Modular Markers

	II	III	I	IV-	V
Kindness	.77*	.26	.03	.00	-.22
Warmth	.73*	.20	.30	.08	.17
Sympathy	.73*	.21	.11	.18	.30
Agreeableness	.65*	.08	-.08	-.06	.26
Sensitivity	.60*	.16	.09	.43	.26
Toughness	-.64*	.04	.09	.10	.06
Slyness	-.55*	-.09	.13	.09	.16
Criticalness	-.47*	.05	.05	.25	.17
Demandingness	-.47*	.19	.22	.31	.11
Efficiency	.23	.79*	.03	-.11	.10
Organization	.07	.77*	-.04	-.05	-.01
Perfectionism	-.04	.71*	-.07	.12	.19
Decisiveness	.00	.55*	.21	-.31	.15
Caution	.21	.50*	-.31	.19	.10
Ambition	.08	.39*	.32	.03	.18
Forgetfulness	-.02	-.57*	-.01	.26	.01
Talkativeness	-.06	-.08	.70*	.11	-.13
Sociability	.35	.20	.66*	-.05	-.07
Assertiveness	-.37	.27	.62*	-.03	.20
Spontaneity	.18	-.16	.51*	.26	.31
Adventurousness	-.03	-.05	.47*	-.03	.34
Restraint	.10	.15	-.71*	.10	.07
Shyness	.11	-.08	-.66*	.23	.11
Fretfulness	-.16	-.20	-.22	.65*	-.08
Anxiety	-.20	-.12	-.01	.63*	-.02
Emotional Excitability	.18	-.06	.39	.59*	.08
Jealousy/Envy	-.34	-.20	.02	.55*	-.11
Hyperdevotedness	.13	.12	-.14	.48*	.09
Analytical Inquiry	.01	.15	-.02	.05	.81*
Reflectiveness	.20	.11	-.16	.07	.65*
Intellectuality	.11	.32	.09	-.12	.52*
Unconventionality	-.24	-.41	.21	-.03	.41*

Note: N = 1,620. Coefficients are varimax-rotated factor loadings; I = Extraversion (Dynamism); II = Agreeableness (Altruism vs. Antagonism); III = Conscientiousness (Self-Regulation); IV = Emotional Stability (reflected: Anxiety); V = (Autonomous) Intellect;* = Highest loading for each variable.

descriptive adjectives in 14 data sets (Saucier, 2001) was dichotomized around cutpoints at the 16.67, 33.33, 50, 66.67, and 83.33 percentiles of the distribution.[12] For each factor, each of the 121 candidate parcels was correlated with each dichotomy.

The three highest-correlating parcels for each dichotomy were retained as part of the initial version of the marker scale. This initial version was revised so as to further reduce scale intercorrelations and also to better maximize correlations with the criterion factor scores.

[12] The 33.33 and 66.67 cutpoints did not have incremental usefulness beyond the other three cutpoints, and thus it was not necessary to use them in this instance.

Table 7. The Modular Markers: Reliabilities and Interscale Correlations

	Derivation Samples			Cross-Validation Sample
	Self	Liked Peer	Pooled Peer	Community Sample
Coefficient Alpha				
I	.88	.89	.91	.84
II	.82	.86	.94	.83
III	.85	.88	.91	.86
IV	.79	.75	.80	.82
V	.77	.75	.87	.82
Interscale Correlations				
I-II	-.11	-.07	-.02	-.01
I-III	.11	.04	-.03	.22
I-IV	-.06	-.05	-.10	.08
I-V	.13	.26	.15	.20
II-III	.01	.12	-.01	.05
II-IV	-.09	.04	.28	.28
II-V	.02	.00	.21	-.21
III-IV	.01	.13	.09	.23
III-V	.04	.02	.23	.03
IV-V	.00	.00	.18	-.02
Mean Correlation				
Modular Markers	.01	.05	.10	.08
100 Markers	.13	.24	.27	.25
40 Mini-Markers	.11	.18	.26	.22
Ortho-40 mean	.01	.05	.09	.07

Note: Sample sizes: Self = 320; Liked Peer = 316; Pooled Peer = 205; Community Sample = 592; All analyses used the original (non-ipsatized) response data.

The end result was the set of parcels presented in the Appendix: 7 parcels (20 items) for Extraversion, 9 parcels (22 items) for Agreeableness, 7 parcels (18 items) for Conscientiousness, 5 parcels (16 items) for Emotional Stability, and 4 parcels (14 items) for Intellect. Factor analyses of the 32 parcels (from 90 adjectives) making up the Big Five marker set indicated that the parcels reproduced the desired factors quite faithfully with either varimax or quartimax rotated solutions. The varimax solution for a combined sample of 1,620 ratings is presented in Table 6.

Table 7 provides Big Five Modular Marker scale intercorrelations, using original (non-ipsatized) responses in five samples; the comparable values for Goldberg's (1992) 100 Markers are also provided. The 100 Markers have roughly the same level of average inter-scale correlations as do most previous Big Five marker sets (e.g., Benet-Martinez & John, 1998; Costa & McCrae, 1992) — about .20. In contrast, the Modular Markers have an average inter-scale correlation of only about .05, similar to that of the Ortho-40 set presented earlier. The highest single inter-scale correlation found in any sample was only .28 (compared to .40 for the Ortho-40). However, the Alpha reliability coefficients of the Modular Marker scales are higher than those for the Ortho-40 by about .05 on average, as one would expect given their greater length (90 items instead of 40). These comparisons suggest that the Modular Markers may be slightly superior to the Ortho-40 as a set of mutually orthogonal marker scales.

The Modular Markers, with 90 adjectives, are of roughly comparable length to

Table 8. The Mini-Modular markers (3M40): Reliabilities and interscale correlations

	Derivation Samples			Cross-Validation Sample
	Self	Liked Peer	Pooled Peer	Community Sample
Coefficient Alpha				
I	.82	.84	.85	.77
II	.71	.76	.89	.71
III	.76	.75	.84	.76
IV	.67	.63	.71	.72
V	.67	.64	.80	.73
Interscale Correlations				
I-II	.02	.05	.01	.09
I-III	.03	-.04	-.06	.19
I-IV	-.05	-.09	-.17	.06
I-V	.09	.15	.07	.14
II-III	.01	.08	-.04	.10
II-IV	-.05	.04	.26	.24
II-V	.00	.10	.24	-.10
III-IV	.03	.09	.06	.18
III-V	-.02	-.02	.08	-.04
IV-V	-.02	.06	.17	.08
Mean Correlation				
3M40	.01	.04	.06	.10
Ortho-40	.01	.05	.09	.07
100 Markers	.13	.24	.27	.25
40 Mini-Markers	.11	.18	.26	.22

Note: Sample sizes: Self = 320; Liked Peer = 316; Pooled Peer = 205; Community Sample = 592 for the 3M40 scales and 1,125 for the other marker sets. All analyses used the original (non-ipsatized) response data. The 3M40 items: I = Assertive, Playful, Sociable, Talkative vs. Quiet, Reserved, Shy, Withdrawn; II = Kind, Sentimental, Sympathetic, Tolerant vs. Cold, Critical, Demanding, Harsh; III = Cautious, Efficient, Meticulous, Organized, Perfectionistic vs. Absent-minded, Disorganized, Indecisive; IV = Unenvious, Unexcitable vs. Anxious, Emotional, Fearful, Fretful, High-strung, Nervous; V = Complex, Intellectual, Nonconforming, Philosophical, Unconventional vs. Conventional, Unintellectual, Unreflective.

the 100 Markers, but they were developed using different criteria, reflecting differing priorities. The 100 Markers were constructed with an emphasis on Criteria 5 through 7 (Alpha maximization, Factor saturation, and Discrimination), and as would be expected their reliabilities are slightly higher than for their Modular Markers counterparts, which were developed with more emphasis on Criteria 9 and 10. However, the use of a representative set of item parcels at the first stage of scale construction gives the Modular Markers some kinship to Goldberg's (1990) 133 and 100 clusters which we described earlier, with more emphasis on Criterion 4 than was true for the 100 Markers. Many of the parcels in the Modular Markers have balanced keying, which was true for none of Goldberg's (1990) clusters.

What if one were to apply the brevity criterion (Criterion 8) to the Modular Markers, and seek an abbreviated set? Table 8 provides internal consistency estimates and inter-scale correlations for a set of 40 Mini-Modular-Markers (3M40). This reduced set of adjectives was developed by selecting from the 90 Modular Markers a subset of items that (a) retained the highest-loading items with (b) about equal numbers having positive and negative loadings on each of the other factors, (c) while maintaining a spread of response means on each scale, with some secondary

attention also to (d) maintaining balanced keying, (e) representing as many of the 32 parcels as feasible, and (f) excluding items where doing so increased the internal consistency of the scale. Bases (a) through (f) correspond to our Criteria 6, 9, 10, 2, 4, and 5, respectively.

Compared to the full set of 90 Modular Marker items, inter-scale correlations for the 3M40 are about the same on average. But internal consistency is lower (almost .10 per scale on average) than for the longer marker set. Compared to the Ortho-40 described earlier, the inter-scale correlations are similar, but the Alpha coefficients for the 3M40 scales are slightly lower (generally by less than .05). The lower internal consistency is due to the higher degree of representative sampling in the 3M40 scales. Although the two marker sets have nearly identical items for Emotional Stability, on the other factors the 3M40 scales appear to be broader in content reference, primarily because the item pool in the Modular Markers has more breadth than that found in the 100 Markers on which the Ortho-40 was based. For example, the 3M40's scale for Intellect has "unconventionality" content that is lacking in the Ortho-40 version (as well as the 100 Markers). Representative sampling does not maximize Coefficient Alpha, although it may heighten validity with respect to a broad array of criteria. Indeed, Saucier (in press) reported that the Ortho-40, Modular Markers, and 3M40 Big Five marker sets demonstrated validities as high as the 100 unipolar markers of Goldberg (1992) and the NEO-FFI (Costa & McCrae, 1992) even though their Alpha coefficients were generally lower.

Integrating diverse psychometric criteria for item selection

Scale construction can serve any of many possible masters, but these masters can lead us in divergent directions. One might attempt to create marker scales based on all of the 10 criteria we have discussed, without realizing the extent to which some of these criteria are in conflict with each other. For example, maximizing the Coefficient Alpha of a scale can be done at the expense of (a) maximizing the spread of item difficulties and (b) brevity. If such Alpha-maximization involves narrowing the content of the scale, validity could be attenuated over what it might otherwise be. If one seeks a representative sampling of variables in one's marker set, one is unlikely to achieve relatively orthogonal markers for orthogonal factors; and likewise, if one achieves orthogonal markers, it is probably at the expense of representative sampling. Uniform sampling, such as that used in the development of circumplex scales (e.g., Saucier, Ostendorf, & Peabody, 2001; Wiggins, 1980), will also tend to conflict with representativeness, not to mention brevity.

Thus, in most cases it will not be practical to apply all the criteria we have described to the construction of a single scale. We suggest that, instead, the criteria be integrated into a measurement paradigm in which each of the criteria is applied somewhere, but not necessarily everywhere. For example, one might build an initial set based on representative sampling of the domain, then select markers as a subset of this representative sample. One might utilize Alpha-maximizing approaches in

creating item parcels but temper this standard with the "discrimination at diverse levels" criterion in aggregating the parcels into measures of broader attributes, whose mutual orthogonality could be systematically maximized if factor-orthogonality is important. Although these procedures are more complex, the simultaneous consideration of diverse psychometric goals should lead to higher quality measures than might otherwise be achieved.

Recommendations

We have presented a variety of English-language marker sets targeted at the Big Five. These marker sets differ with respect to the original item pool as well as the criteria used in constructing them. Which is the best marker set? With respect to predictive validity, Saucier (in press) compared all of the adjectival marker sets we have presented (except the 100 clusters) and found no meaningful differences; surprisingly, the 40-item marker sets appeared to have validities equivalent to those with more than twice as many items. The 100 unipolar markers and the Mini-Markers are especially geared toward factorial replicability — generating an intended structure in exploratory factor analysis of the constituent items. Due to their length, the 100 markers, and then the Modular Markers, typically have the highest Alpha coefficients, and thus would provide the most precise differentiation of individuals. On the other hand, the Mini-Markers, Ortho-40, and 3M40 all require less than half as much subject time as these more reliable marker sets. Finally, if one wishes to have more mutually orthogonal scale scores, one would choose the Modular Markers, Ortho-40, or 3M40.

From another perspective, the 100 unipolar markers combine high Alpha coefficients with factorial replicability. The Mini-Markers combine factorial replicability with brevity. The Modular Markers have more breadth and relative mutual orthogonality, although Alphas are not quite as high as those for the 100 Markers. Both the Ortho-40 and 3M40 combine brevity and mutual orthogonality; based on the way that the item pool from which each was derived, the Ortho-40 is likely to have more factorial replicability and the 3M40 more breadth.

These all appear to be good marker sets, but there is no single "best" one. Instead, what is best depends on how the investigator weights and values the various scale-development criteria. This is consonant with the overarching theme in our chapter: Trade-offs arise in the scale construction process that usually prevent one from generating a single perfect scale for a construct. We do tend to favor, however, scales that were developed taking a larger number of important criteria into account (such as the Modular Markers and its short form, the 3M40), on the grounds that these scales are less likely to have an "Achilles heel"; they are more balanced with respect to their virtues. Similarly, we encourage other investigators to take a broader view of scale construction, and to integrate a diverse range of useful criteria into the scale-construction process.

Author's note

Work on this article was supported by Grant MH-49227 from the National Institute of Mental Health, U.S. Public Health Service. The authors are enormously indebted to Michael Ashton, Kimberly Anne Barchard, Ira Bernstein, Michael Browne, Matthias Burisch, A. Timothy Church, Robyn M. Dawes, Herbert Eber, Candan Ertubey, David Evans, Willem K. B. Hofstee, Eric Knowles, John Loehlin, Susan D. Long, Roderick P. McDonald, Richard Robins, Oya Somer, Lynne Steinberg, Krista Trobst, Erika Westling, Jerry S. Wiggins, and Richard Zinbarg for their thoughtful comments and suggestions.

References

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.

Benet-Martínez, V., & John, O.P. (1998). *Los Cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729-750.

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

Burisch, M. (1978). Construction strategies for multiscale personality inventories. *Applied Psychological Measurement, 2*, 97-111.

Burisch, M. (1984a). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39*, 214-227.

Burisch, M. (1984b). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18*, 81-98.

Cattell, R.B., & Tsujioka, B (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement, 24*, 3-30.

Comrey, A.L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*, 754-761.

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

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*, 1246-1256.

Dwight, S.A., Cummings, K.M., & Glenar, J.L. (1998). Comparison of criterion-related validity coefficients for the Mini-Markers and for Goldberg's markers of the Big Five personality factors. *Journal of Personality Assessment, 70*, 541-550.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Goldberg, L.R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monograph, 7*, No. 72-2.

Goldberg, L.R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141-165). Beverly Hills, CA: Sage.

Goldberg, L.R. (1982). From Ace to Zombie: Some explorations in the language of personality. In C.D. Spielberger, & J. N. Butcher (Eds.), *Advances in Personality Assessment* (Vol. 1: pp. 203-234). Hillsdale, NJ.: Erlbaum.

Goldberg, L. R. (1990). An alternative "Description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Goldberg, L.R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42.

Goldberg, L.R. (1997). *The Development of Five-Factor Domain Scales from the IPIP Item Pool*. Unpublished manuscript. Oregon Research Institute; Eugene, OR 97403; USA.

Goldberg, L.R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, Vol. 7* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L.R. (in press). The comparative validity of adult personality inventories: Applications of a consumer-testing framework. In S. R. Briggs, J. M. Cheek, & E. M. Donahue (Eds.), *Handbook of Adult Personality Inventories*. New York: Plenum.

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

Hase, H.D., & Goldberg, L.R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67*, 231-248.

Hendriks, A.A.J., Hofstee, W.K.B., & De Raad, B. (1999). The Five-Factor Personality Inventory (FFPI). *Personality and Individual Differences, 27*, 307-325.

Hofstee, W.K.B. (1990). The use of everyday personality language for scientific purposes. *European Journal of Personality, 4*, 77-88.

Hofstee, W.K.B., Ten Berge, J.M.F., & Hendriks, A.A.J. (1998). How to score questionnaires. *Personality and Individual Differences, 25*, 897-909.

Hofstee, W.K.B., De Raad, B., & Goldberg, L.R. (1992). Integration of the Big-Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63*, 146-163.

Hogan, R., & Hogan, J. (1995). *Manual for the Hogan Personality Inventory*. Tulsa, OK: Hogan Assessment Systems.

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

Johnson, J.A. (2000). *Developing a short form of the IPIP-NEO: A report to HGW Consulting*. Unpublished manuscript. Department of Psychology, University of Pennsylvania, DuBois PA.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493-504.

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

Lord, F.M. (1952). The relationship of the reliability of multiple choice items to the distribution of item difficulties. *Psychometrika, 18*, 181-194.

McCrae, R.R., & Costa, P.T., Jr. (1998). Toward a new generation of personality theories: Theoretical contexts for the Five-Factor Model. In J. S. Wiggins (Ed.). *The Five-Factor Model of Personality: Theoretical Perspectives* (pp. 51-87). New York: Guilford.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Norman, W.T. (1963). Relative importance of test item content. *Journal of Consulting Psychology, 27*, 166-174.

Norman, W.T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population*. Department of Psychology, University of Michigan, Ann Arbor, MI.

Nunnally, J.C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Peabody, D. (1987). Selecting representative trait adjectives. *Journal of Personality and Social Psychology, 52*, 59-71.

Peabody, D., & Goldberg, L.R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology, 57*, 552-567.

Robins, R.W., Hendin, H.M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin, 27*, 151-161.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*, 506-516.

Saucier, G. (2001). *What is more replicable than the Big Five? Broader factors in English-language personality adjectives*. Manuscript submitted for publication.

Saucier, G. (in press). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality*.

Saucier, G., & Goldberg. L.R. (1996a). Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality, 10*, 61-77.

Saucier, G., & Goldberg, L.R. (1996b). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.). *The five-factor model of personality: Theoretical perspectives* (pp. 21-50). New York: Guilford.

Saucier, G., & Goldberg, L.R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality, 69*, 847-879.

Saucier, G., Ostendorf, F., & Peabody, D. (2001). The non-evaluative circumplex of personality adjectives. *Journal of Personality, 69*, 537-582.

Smith, G.T., McCarthy, D.M., & Anderson, K.G. (2000). On the sins of short-form development. *Psychological Assessment, 12*, 102-111.

Wiggins, J.S. (1980). Circumplex models of interpersonal behavior. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1, pp. 265-294). Beverly Hills, CA: Sage.

Zinbarg, R.E., Yovel, I., Revelle, W., & McDonald, R.P. (2000). *Beyond Alpha: Coefficients of generalizability and the internal structure of tests*. Unpublished manuscript. Department of Psychology; Northwestern University (Evanston, IL 60208-2710; USA).

Appendix. The 32 parcels included in the Big Five modular markers

Spontaneity:	Impulsive, Spontaneous, Playful (.65, .59)
Talkativeness:	Talkative, (-) Quiet (.61,.77)
Sociability:	Sociable, (-) Unsociable, (-) Withdrawn (.73$_b$, .78)
Assertiveness:	Dominant, Assertive, Forceful,(-) Timid (.68$_a$, .75)
Adventurousness:	Daring, Adventurous, (-) Unadventurous (.77$_a$, .78)
Shyness:	Shy, Bashful (.83, .81)
Restraint:	Inhibited, Reserved, Restrained (.52$_c$, .62)
Warmth:	Warm, (-) Cold (.64, .73)
Sympathy:	Sympathetic, Compassionate (.75,.78)
Sensitivity:	Sensitive, Sentimental (.48, .66)
Kindness:	(-) Cruel, Kind (.56, .66)
Agreeableness:	Agreeable, Tolerant, Lenient (.52$_a$, .65)
Toughness:	Rough, Tough, Stern, Harsh (.58$_b$, .73)
Criticalness:	Critical, (-) Uncritical (.44$_a$, .63)
Demandingness:	Demanding, (-) Undemanding (.53$_a$, .70)
Slyness:	Sly, Cunning, Shrewd (.59$_c$, .69)
Organization:	Organized, (-) Disorganized (.80, .82)
Caution:	Careful, Cautious (.70,.77)
Ambition:	Ambitious, (-) Unambitious (.78, .71)
Decisiveness:	Decisive, (-) Indecisive (.58$_b$, .66)
Efficiency:	Efficient, (-) Inefficient, (-) Careless (.70, .69)
Perfectionism:	Perfectionistic, Exacting, Meticulous, Precise (.74$_c$, .75)
Forgetfulness:	Forgetful, Absent-minded, Scatterbrained (.73$_b$, .76)
Jealousy/Envy:	Jealous, Possessive, Envious, (-) Unenvious (.67, .76)
Emotional Excitability:	Excitable, Emotional, (-) Unexcitable (.63$_a$, .72)
Anxiety:	Anxious, Nervous, High-strung (.73$_a$, .63)
Fretfulness:	Fretful, Fearful (.43$_a$, .54)
Hyperdevotedness:	Overloyal, Overprotective, Overconscientious, Oversentimental (.70$_g$, .61$_d$)
Intellectuality:	Intellectual, (-) Unintellectual (.70, .71)
Analytical Inquiry:	Philosophical, Deep, Complex, Analytical (.67$_b$, .70)
Reflectiveness:	Introspective, Contemplative, (-) Unreflective (.57$_a$, .73)
Unconventionality:	(-)Traditional, (-) Conventional, Unconventional, Nonconforming, Rebellious (.76, .74)

Note: ESPS = Eugene-Springfield Community Sample combined with college peer-rating sample, N = 901; ABCD - Combined college-student samples A, B, C, and D, N = 1,028. Coefficients in parentheses are, respectively, coefficient alpha in ESPS and ABCD; subscript letters indicate sample size for all items in parcel, a – N = 694, b – N = 596, c – N = 592, d – N = 841, e – N = 823.