The Alberta Journal of Educational Research *Vol. XLIX, No. 3, Fall 2003, 212-230*

Roderick P. McDonald
University of Illinois

Behavior Domains in Theory and in Practice

The concept of a behavior domain is a reasonable and essential foundation for psychometric work based on true score theory, the linear model of common factor analysis, and the nonlinear models of item response theory. Investigators applying these models to test data generally treat the true scores or factors or traits as abstractive psychological attributes: common properties of the items, possibly with some inconsistency between their practice and their theoretical statements. A countably infinite item domain defines an attribute uniquely, and a function of the domain item scores gives an identified measure of it, to be estimated from a finite set of item scores, with a defined error of measurement. In test development the investigator must consider and justify the assumption that an item domain exists for the specific measurement application and is large enough to be treated as infinite for that application.

In the following, three fundamental matters are considered. First, it will be suggested that the concept of a behavior domain—of a universal set of "items" of a given "kind" to be explicated—is the most reasonable foundation for such standard psychometric methods as the following: (a) generalizability theory, (b) errors of psychological measurement, (c) theory for altering the length of a test, (d) construct validity, (e) alternate test forms, (f) computer adaptive testing, (g) differential item functioning, (h) facet theory, and (i) many more specialized problems. Second, we consider the question: How should psychological attributes such as abilities, personality traits, states such as moods, attitudes, or values be conceptualized, and what does our conceptualization of the attribute imply for the denotation of behavior domains and the determination of an attribute by observations? Third, we are led to examine the practical or substantive implications of the behavior domain concept for the construction of tests, and coming full circle the limitations on behavior domain theory imposed by the practical limitations of test construction.

An adequate account of such large questions cannot be given within the compass of a single article. A number of compromises are adopted including some specialization for concreteness and the deliberate adoption of a not-too-deeply philosophical level of analysis, that is, not too far removed from the practical. Known technical (mathematical) results are summarized as informally as possible, because mathematical results are not central to the questions. It is not expected that the account will appear definitive to all readers or to any reader. It is enough that a neglected topic is opened for further discussion.

As a preliminary it is necessary to establish some terminology and some mild specializations to give focus to the discussion. The terms *behavior domain* and near synonyms, the *universe of content* and the *universe of admissible measurements* seem to have been most extensively used by Guttman (1953a, 1954, 1955,

---

Roderick McDonald is a professor in the UIUC Department of Psychology in the Quantitative Division.

1957, 1959, 1965, 1971) in the contexts of factor analysis and of facet theory and by Cronbach, Gleser, Nanda, and Rajaratnam (1972) in the context of generalizability theory, a variant of Guttman's (1965) facet theory. Guttman used the concept to obtain limiting properties of the common factor model as the number of tests analyzed approaches infinity, where the tests are in some sense drawn from a previously identified infinite set of tests. Cronbach and his colleagues have used the concept of an infinite behavior domain in a linear (ANOVA) model for a finite number of items, raters, and so forth to estimate generalizability to further items, raters, occasions, and so forth, drawn in some sense from a *universe* of such entities. See also Lord and Novick (1968) for a number of applications of a more or less explicit notion of the drawing of items from a domain that might yield a test of infinite length.

These brief remarks must serve as a summary of the literature on behavior domains. I propose to set out a conceptualization of an *item domain* as the basis of the present treatment that is consistent enough with that literature, although in some respects more specific and more limited. A behavior domain based on tests and their scores—item-sums—is already at a second level of abstraction or complexity depending on the homogeneity or heterogeneity of the item sets forming them. These complexities are avoided here by using items as the elements of the domains rather than tests.

Recall that the typical item in an objective test measuring a psychological attribute—trait or state—consists of a stem and a set of options, the choice of which by a respondent is objectively scorable. In self-report personality or attitude or value items, the stem is generally in a language shared by the test constructor and the respondent. (The philosophically inclined reader will recall Wittgenstein on the impossibility of a private language.) Cognitive items and projective test items admit a wider variety of stimulus materials, not strictly requiring a shared language yet still resting in a sense on shared symbolic systems. The chosen response option may be coded to give an item score, for example, 0/1 for a binary item, 1 through k for Likert scoring of k ordered categories, or similarly an integer score for answers to questions on a common stem, for example, a passage of prose or verse. Commonly the scores from the set of items are combined to yield a sum- mean- or formula-score for a "homogeneous" test or possibly to yield a profile of subtest scores.

We also recall that the mathematical concept of a domain is the set of elements on which a mathematical or logical variable is defined. The obvious but perhaps neglected implication of these truisms is that the item stems identify the items as the elements of a set that constitutes a psychological test. Of course, each item-score can be modeled as a random variable defined on a domain consisting of a sample space of possible respondents, but this is not the domain that concerns us in item-domain theory, and the item-scores of a set of items do not themselves constitute the elements of a behavior domain or universe of content or admissible measurements. More generally, in multifaceted designs such as those treated by Cronbach et al. (1972), it would again be the distinct raters, occasions, or situations that define the *universe of admissible measurements*, not the measurements: the quantities themselves, which lack the required property of distinctness. Here we will limit discussion to item(-stem) domains, both for concreteness and for brevity, and avoid alterna-

tive terminologies. The conception of a test as a set of items in the strict sense of set theory forces attention on the common property of the items that defines set membership, or if the items in a given finite test are merely listed, forces attention on the lack of a criterion for set membership.

A trivial formalization of the commonsense notion of shortening a test follows by noting that a shortened test contains a proper subset of the items (the identifying stems and the associated response options) in an initially *realized* set of items. (The term *realized* is used here to mean "made real" instead of saying "written" or "constructed.") Often, but not always, test construction begins with a "large" set of realized items designed to measure a psychological attribute from which a smaller set is chosen either for convenient administration or because some items are not considered satisfactory indicators of the attribute.

Less trivially, a formalization of the commonsense notion of lengthening a test follows by noting two possibilities. First, a lengthened test might be obtained from a large set of realized items—the *item pool or bank*—from which the short test was taken, so that lengthening the test consists merely in adding more realized items from the pool. Given the item parameters, available theory allows optimal choice of items to add. Second, a lengthened test might be contemplated in theory where the short initial test consists of all items so far realized, and, for example, we ask how many more items are needed to attain a certain precision of measurement (reliability). This second case points to some fundamental conceptual problems and thereby to the main argument of this article.

1. The psychometric properties of the as yet unrealized items are not yet known: hence, for example, the strong assumptions of exchangeability of items in the Spearman-Brown formula for reliability of the lengthened test and in Cronbach's ANOVA treatment of generalizability.
2. More important, it may not be known what prescribes membership in the extended "universal" set of items. That is, we may lack a denotation for membership of the universal set.
3. Given a sufficiently clear denotation, we may not be able to write more than a limited number of further items that belong to the set, or it may not be known whether this is indeed possible. It may be that we do not see how to write just one more item. Consider a knowledge test on the signature keys of the Beethoven symphonies.

The ultimate extension of a test is, of course, to a universal set containing a countable infinity of realizable items—yielding a test of infinite length—of which only a finite number may be realized in practice. Much elegant psychometric theory rests on the concept of a test of infinite length, but for such theory to be applicable, the application must appear at least plausible to the imagination of researchers. This is a substantive, not a mathematical problem, and it cannot always be simply supposed possible. Again, a simple example is a test on the keys of the nine Beethoven symphonies.

Certainly the attainability of a countable infinity of items cannot be directly verified in an application. We might suppose, much as in the asymptotics of sampling theory for an infinite population of subjects, that behavior at infinity can be well approximated by a finite "sample" of items. We might also sup-

pose—a distinct matter—as in the application of sampling theory for infinite populations to sampling from large but finite populations that infinite item-domain theory will sufficiently approximate the behavior of a finite but large item domain. (Perhaps for a knowledge test the 104 Haydn symphonies are a virtual infinity.) Indeed, Guttman (1953a) conjectured that a domain of about 10-15 variables would closely approximate an infinity of them.

If we have a clear denotation for a psychological attribute giving a clear prescription for the item-writer, item realization is still not analogous to the random sampling of individuals from a defined population of interest. It can be conjectured that commonly the best items are written first, so (a) later items will not match early items in their desirable parameters, (b) later items will less clearly belong to the domain, and (c) later items will begin to fall naturally into subsets corresponding to subdomains. Implicit in these remarks is a conception of the homogeneous test, which for the moment we take simply from the Greek root of the word to be one where the items are of the same kind: the *kind* being given by the denotation of the set of item stems, or possibly by intuitive abstraction from the character of the items.

It might seem that here discussion should continue on the nature of psychological attributes—concepts, constructs—and the conditions needed in applications for the denotation of an item domain—a universal set of as yet unrealized items that is possibly countably infinite—before turning to an account of the formal psychometric theory that is enabled by item domains. The advantage of putting the cart before the horse and turning to a sketch of the main chapters of the formal theory of psychological measurement that can be based on the infinite item domain conception is that the psychometric concepts in turn help to formalize the discussion of psychological attributes in theory and in practice. In the following section I consider the main applications of item domains to test theory. Following this is a section about the determinacy of factors in relation to the conceptualization of psychological attributes. And after this we consider what can be said about the relation between the theory and the application of item-domain concepts.

The Domain of Item Domains

It may be claimed that the range of work that requires infinite item domains for its foundation corresponds to a large part of psychometric theory. To illustrate, I return to the list in the opening paragraph.

1. Generalizability theory has its best known development at the hands of Cronbach et al. (1972). Here we limit consideration to a single-facet design in which $j=1,\ldots,m$ items yield scores X_{ji} from $i=1,\ldots,N$ respondents in a G-study (an initial calibration or generalizability study). It is assumed that the items are exchangeable as to their discrimination parameters, and we use an ANOVA model

$$X_{ij} = \mu + D_j + T_i + E_{ji} \ , \tag{1}$$

where μ is a grand mean, D_j represents item difficulty, T_i respondent attribute, and E_{ji} is a residual, consisting of a random interaction of respondent i with item j (conceptually confounded with an undefined *error*, possibly associated with unrealized replications of the observation). Standard ANOVA yields estimates of σ^2_T, the variance of the attribute, and of σ^2_E, the variance of the

interaction term, in a mixed model (i.e., D_j fixed) or in a random model (i.e., D_j random as well as T_i). In the latter case σ_D^2 the variance of random difficulty, is also estimated. These estimates yield generalizability coefficients: either Cronbach's alpha if difficulties are not random, or an alternative that includes random difficulty in the *error* variance. The generalizability in question is from the finite set of realized items to an infinite set of *admissible measurements*, that is, of as yet unrealized items of which the given items are considered a random and representative sample. The generalizability coefficients as estimated are used to conjecture the properties of responses to finite sets of items drawn from the item domain in a D-study (decision study) under the strong assumptions of the model. It does not seem possible on the face of it to weaken the foundations of this theory by forgoing the infinite item domain while keeping its consequences, except merely by allowing the domain to be finite and "sufficiently" large.

In an alternative treatment of generalizability given by McDonald (1978a) we replace Equation 1 by

$$X_{ji} = \mu + \delta_j + \lambda_j F_i + E_{ji} \tag{2a}$$

or

$$X_{ji} = D_j + \lambda_j F_i + E_{ji} \ , \tag{2b}$$

where D_j (here fixed) again represents item difficulty, λ_j is a scaling constant representing the discriminating power of item j, F_i is a measure of the respondent's attribute, and E_{ji} is again a random interaction between the respondent and item j with variance ψ_j^2. With the usual assumptions, this is just the classical Spearman unidimensional factor model. Here F_i is the respondent's common factor score, and the uncorrelated residuals E_{ji}, E_{ki}, $k \neq j$ are components unique to the item scores. A simple formula score for the set of m items is the mean S_i of the item scores, for which we have a true-score model

$$S_i = D_\bullet + T_i + E_{\bullet i} = D_\bullet + \lambda_\bullet F_i + E_{\bullet i} \ , \tag{3}$$

where $D_\bullet$ and $\lambda_\bullet$ are means of the parameters, S_i is the mean of the item scores, and $T_i = \lambda_\bullet F_i$ a rescaled version of the attribute with variance $(\lambda_\bullet)^2$, It then follows (McDonald, 1978, 1985, 1999) that coefficient omega is a classical reliability coefficient and a coefficient of generalizability, where

$$\omega = \frac{(\lambda_\bullet)^2}{(\lambda_\bullet)^2 + (\psi^2)_\bullet} \ , \tag{4}$$

writing also $(\psi^2)_\bullet$ for the mean of the m unique variances. Omega is the squared correlation between the mean score S_i on the m realized items and the domain mean score T_i, which is the limit of S_i as more items are drawn from a prespecified infinite item domain that fits the unidimensional factor model, that is, is psychometrically homogeneous. It may be shown that $\omega > \alpha$, with equality if and only if the realized items have equal loadings. Coefficient omega has three advantages over coefficient alpha.

1. It does not require that the m realized items are representative of as yet unrealized items in the domain.

2. In the course of its computation we test and possibly falsify the assumption yielding alpha that the items have equal loadings.
3. We test and hope to verify that at least the items so far realized form a psychometrically homogeneous set in the sense of fitting a unidimensional model.

Estimating an attribute of an examinee and assigning an error of measurement to the estimate requires inter alia a clear definition of what is estimated. In classical true-score theory we write

$$S_i = T_i + E_i \ , \tag{5}$$

(where again it is convenient to take the item mean score as the observed test score). This yields for a test of known reliability ρ_r , S_i as the unbiased estimator of T_i , and $(1\text{-}\rho_r)\mathrm{Var}\{S\}$ for the error variance, which may be used to put confidence bounds on the estimator. If ρ_r is taken to be coefficient alpha, the implicit model is Equation 1. It is also possible (McDonald, 1970) to take ρ_r to be coefficient omega, in which case the implicit model is Equation 2. In principle this is preferable as it requires only homogeneity. However, other devices yield a reliability coefficient, the commonest being the test-retest method, which correlates replicate measures on the same m items across two occasions. Retest correlations often give useful information about the temporal stability of a test score over a chosen interval from which we might sometimes be willing to draw indirect inferences about the temporal stability of the psychological attribute the items have been chosen to measure. Generally, however, retest reliability bears no relation to the precision of measurement of the attribute that is the objective of the measurement process.

We may, therefore, reasonably take the strong, falsifiable assumptions of Equation 1 or the weaker, falsifiable assumptions of Equation 2 to define true scores. Then the reliability coefficient is the generalizability coefficient, the true score is the domain score, and the error is due to interactions of random respondents with unique properties of the m realized items, which increasingly cancel on average as the number of items grows. It might be too much to claim that the infinite domain score is the only possible definition of true score. It should be clear that any conceptualization of true score justifying putting confidence bounds on that of a given examinee must make it an identified person parameter to be estimated—with error. Guttman (1953b, 1969) has given cogent critiques of undefined true scores.

An obvious limitation of this discussion is that the linear models in Equations 1 and 2 may be used to a reasonable approximation for items whose options allow scoring over a large enough range of integers, as in Likert scales for attitude or self-report personality items, but at best provide a rough approximation to binary item responses. For a modern account of the estimation of an attribute with an appropriate standard error of measurement, we of course use item response theory. Given m item-stems with a binary response (pass/fail, agree/disagree) we fit, say, a normal ogive or logistic model

$$P\{X_{ji} = 1 \mid \theta_i\} = L(a_j + b_j \theta_i) \ , \tag{6}$$

where $L(\bullet)$ is a suitable nonlinear link-function, squeezing the response probability between the required bounds of zero and unity, whereas the *latent* trait

θ_i is defined unboundedly from negative to positive infinity. The principle of local independence, which states that conditional on θ_i the item responses are mutually independent, is a stronger analogue of the assumption of the linear model in Equation 2, that conditional on the common factor the item scores are uncorrelated (McDonald, 1981). In applications, instead of estimating a given examinee's trait value θ_i we might instead estimate her or his true score t_i. The test characteristic curve

$$t_i = \frac{1}{m} \sum P\{X_{ij} = 1 \mid \Theta_i\} \tag{7}$$

gives the true score as a function of the trait. (This is a *relative* true score bounded by zero and unity.)

There appears to be little to choose in applications between the unbounded metric for θ and the bounded metric for t, because, it may be claimed, the psychological attribute is at best defined on an ordinal scale. It is well known—see, for example, Lord (1980)—that the error variance of the observed mean score S about the true score t of a specific examinee is given by

$$\sigma_{S|\Theta} = \frac{1}{m} \sum P\{\theta\}[1 - P\{\theta\}]. \tag{8}$$

For brevity and simplicity I avoid a more developed exposition of information functions and efficient estimators in favor of an account closer to classical test theory (see Lord, 1980, or McDonald, 1999, chap. 13, for a further account).

As in classical test theory, a fundamental question for item response theory is what defines in their respective metrics the trait θ and the true score t for a chosen examinee, that we seek to estimate from his or her pattern of responses. The reader may consider it too much to claim that the infinite item domain score is the only possible realization of true score in an item response model or that θ given by the item domain is the only possible realization of the latent trait: which, we note, is then seen to be not in any sense latent, hidden, or underlying, but simply as the limit of a sequence of observations that is possible in principle, but cannot be completed in practice. For this reason it should always be preferable to speak of common factors, not latent traits, in IRT.) Yet again it should be clear that any alternative treatment of the estimation of true scores or (latent) traits yielding a standard error of measurement must make these quantities identified person parameters to be estimated. On the face of it, without a clearly denoted item domain, identifiability is lacking.

The effect of shortening or lengthening a test on reliability or error of measurement can easily be estimated under the model of Equation 1. By well-known theory this requires a simple application of the Spearman-Brown formula. If we have actually tested this model, verifying that the item covariances are equal or—an equivalent—the item loadings are equal, it is a matter of indifference which items we remove to shorten the test. In adding further items to lengthen it, the use of the Spearman-Brown formula is conjectural as we cannot know that as yet unrealized items will be parallel to the realized set.

Under the model of Equation 2 it is easy to show that we can choose as a subset giving maximum reliability those with the largest information measured by λ_j^2/ψ_j^2. This model does not help us to go beyond the Spearman-

Brown formula for a conjecture as to the gain in reliability from adding further items, but it does weaken the needed assumption to the requirement that the average loading and average residual variance of the added items be the same as corresponding averages of the given set (McDonald, 1999). There is now considerable theory for shortening tests fitting an item response model, but the problem of adding items does not allow, on the face of it, any new approach.

Fundamental to these models for the effects of shortening or lengthening a test is the notion that the given set of m items, the chosen proper subset, and the extended set all yield estimates of the same quantity: true score or trait. It is then reasonable to regard the true score or trait to be estimated as the score given by the infinite item domain. The alternative to the assumption of an infinite item domain here seems to be a vicious particularism in which we say that adding or subtracting one item changes what is measured.

Construct validity, as (apparently) distinct from content validity, remains perhaps an ill-defined yet influential notion. It roughly amounts to the claim that a test of m items is valid to the extent that a sum or formula score derived from the item responses measures a theoretical concept or *construct*. (The notion of a *construct* belongs to a long-dead fashion in philosophy, but the word has unaccountably survived.) Factor analysis and (occasionally) path analysis seem to be the two main recognized devices for establishing construct validity (Messick, 1989). Here I risk making the claim, for brevity, that in many studies purporting to establish construct validity, the construct is indeed taken to be the common factor of the items. It follows that coefficient omega in Equation 4, the squared correlation between the test score and the common factor, is a measure of construct validity. In this view, at this point we conclude that reliability is generalizability is validity, which makes a considerable simplification in thought (McDonald, 1985).

I now state the central claim of this article: In applications to homogeneous tests the Spearman factor, by Equation 2, corresponds to the attribute the m realized item stems indicate in common, and the responses to them measure in common. The countably infinite set of item stems in the domain substantively gives a unique identity to that attribute, and a function of the scores on that set determines a measure of the attribute uniquely, as the quantity to be estimated from any finite subset.

The list thus begun could continue, to show how the infinite item domain provides a foundation for alternative test forms—disjoint sets of items—measuring "the same attribute"; computer adaptive tests measuring "the same attribute" with individualized subsets of items; detecting items that function differentially in measuring "the same attribute" in distinct populations of interest; a quantitative account of facet theory: the reader may continue the list. It could also be shown at length (but it is hoped that this is self-evident) that those well-known devices for the analysis of item responses that do not constitute falsifiable statistical models—principal components, smallest space analysis, correspondence analysis—cannot provide an alternative technology for problems (a) through (h) or for others of the same kind.

Factor Determinacy and the Conceptualization of Psychological Attributes

It has long been known that for a fixed and finite set of variables, the mathematical equations of the Spearman factor model—and multiple factor counter-

parts—have an infinity of alternative solutions in terms of common-factor and unique-factor scores given the item parameters. This was first noted in the unidimensional case by Wilson (1928), and Spearman (1929) responded in terms of an infinite domain of tests. Kestelman (1952) gave corresponding results for multiple factors—the multidimensional case—and Guttman (1955) showed how to choose maximally distinct alternative mathematical solutions to the model equations. Further discussions of the mathematics of the *factor indeterminacy* question are given by Maraun (1996a, 1996b, 1996c), McDonald (1996a, 1996b), Bartholomew (1996a, 1996b), Mulaik (1996a, 1996b), Rozeboom (1996a, 1996b), Schonemann (1996a, 1996b), and Steiger (1996a, 1996b).

The mathematical results have long been understood. Their possible implications in the use of the common-factor equations to model the behavior of examinees in response to tests or items do not seem to have been unequivocally stated and remain problematic. Some discussion of possible implications for practice can be found in the references cited above, but these still perhaps remain obscured by technicalities.

The algebraic results, which are indeed quite technical, can be sketched as follows. Because in the common factor model the number of common factors plus unique factors—residuals—must exceed the number m of variables X_{ji}, that is, there are more unknowns than knowns in the system of linear equations, then from the values of the observations X_{ji} we can construct infinitely many alternative sets of common factor scores F_i and residuals E_{ji}, which jointly satisfy the equations by the use of arbitrarily generated numbers (Kestelman, 1952). The correlation between maximally dissimilar lists of common factor scores is $2\rho_m^2 - 1$, where ρ_m^2 is the squared multiple correlation between F_i and the m item scores (Guttman, 1955). A necessary condition for the Spearman factor and the residuals to be determined by a finite number of items is that one of them shall have zero residual variance. This condition can be approximated in applications and is a special type of improper solution: an exact Heywood case. The best discussion of this remains that of Thomson (1951). On rewriting an item response model for binary variables as a common-factor model for continuous variables that yield the binary variables by dichotomization, we easily see that item response models give corresponding multiple solutions for θ_i in Equation 6. Further remarks below implicitly apply to item-response models also. Some readers may not be aware that the factor indeterminacy "problem" is equally a problem in the Rasch model. Conventional accounts of this model have not commonly defined the quantities being estimated or shown how these are distinguished from the estimators. For our purpose it will suffice to consider the unidimensional case: the Spearman model given in Equation 2.

The results cited belong to the basic mathematical structure of the common factor model for a fixed set of just m variables. It is not obvious what empirical counterparts they might have when the factor model is used to represent responses of human examinees to item stems written to realize a psychological attribute such as extraversion, social conservatism, or attitude to gun control. Like all mathematical models the factor model has a purely mathematical structure that constitutes the syntactics—the "grammatical" structure—of the

model, and rules of correspondence are needed that constitute its semantics whereby we may say things like "the common factor of these items is anxiety."

The problem is to give a rationale for alternative solutions to the equations of the model when it is applied to an empirical dataset. We may safely dismiss any accounts in the confused and confusing literature on this topic that broadly imply that the existence of alternative solutions to the factor equations makes a problem for applications without showing how the alternative solutions can be obtained from empirical measurements on examinees. The reader is encouraged to apply this test to the collection of references cited on this topic.

In the present state of knowledge there would appear to be two recognized ways to translate the arbitrariness of the mathematical solutions of the factor equations into alternative meanings of empirical results. The first of these is here labeled the *omitted cause conception;* the second is the *common properties conception.*

In the first, the common factor is conceptualized as an omitted measurement on the respondents: a single observable empirical variable that we have not yet been able to identify and measure, which will when discovered be recognizable as in some sense a common cause of the item responses. An investigator who seriously adopts this conception of a common factor (and its special case, the true score) should not be able to justify using the classical and modern test-theory methods (a) through (h) above, but should instead take the range of solutions of the equations as somehow defining a field of search for a discoverable measure on the examinees whose denotation is independent of the items. The notion is that given a set of unidimensional items, we may eventually discover, and perhaps should immediately try to discover, a measurable variable that has the same correlations with the m item scores as does the common factor and stands in relation to the item responses as an independently identified cause of the responses as effects. In this view we might discover two or more variables with identical profiles of correlations to the item scores and low (possibly negative, if $2\rho_m^2 - 1$ is negative) correlations with each other.

There seems to be no search strategy that could be brought to bear on this problem, and I am aware of no example in the literature of a direct claim by an investigator to have discovered one empirical causal variable—let alone two—that constitutes the identity of a common factor or of a case where the investigator has seriously searched for such a variable. This does not mean, however, that structural analyses of psychological data do not exist that are open to such an interpretation. Nothing in the factor model requires a common factor to be initially unobservable and discoverable in the future. Indeed, any path analysis "without latent variables," that is, "without common factors," in which one externally defined variable is treated as the cause of more than two dependent variables and accordingly omits nondirected paths between the latter, is precisely a case of this kind. (This strains terminology. A model without common factors—latent variables?—may then be a model with observable common factors—latent variables?) We note, however, that any such case can always be interpreted as one where the external variable causes the common factor of the dependent variables, that is, acts to change the level of the psychological attribute common to them. To give a possibly crude illus-

tration, consider the observation that a tranquilizer reduces muscle tension, and suppose that it also reduces state-anxiety measured by a number of self-report items. The reader is invited to decide between (a) saying that muscle tension is a cause or perhaps a physiological correlate of the psychological attribute of anxiety, where *anxiety* is the common factor of the items, and (b) muscle tension is constitutive of anxiety and can now replace the primitive *anxiety* concept. And on what grounds can the choice be made?

We are led inevitably to some of the deeper philosophical questions about the nature of psychological concepts including questions about physicalism or psychophysical dualism. Whether skillfully or clumsily, I propose to sidestep these issues and make the broad and deniable claim that researchers into personality, cognitive abilities, attitudes, and so forth who say and presumably believe that they are investigating as yet unknown neurophysiological entities—"existential concepts" in Feigl's classic terminology—underlying and actually constitutive of common factors such as extraversion, anxiety, agreeableness, economic conservatism, or attitude toward gun control assert a general commitment to physicalism that does not detectably affect their psychological concepts.

A simple criterion we may apply to the behavior of the test constructor to determine the nature of the concept being developed is to ask on what grounds the investigator would add a further item stem to lengthen the test, and continue to measure the same attribute. For example, if we accept that anxiety can be replaced by muscle tension, we will add items like "I need a massage." Suppose also that we are supplied with the observation that quadriplegics can report severe anxiety. We revise our hypothesis to the claim that a tranquilizer directly acts to inhibit a specific cortical activity, which in turn both reduces anxiety and relaxes muscles in the intact person. The respondent cannot generally report the level of the specific cortical process, so this hypothesis gives no rationale for writing an $(m+1)$st item.

It is my deniable but not unreasonable claim, made from wide observations of the literature, that investigators do not operate a common-cause notion in applications of common factor/item response models. Rather, they write or add them to a given set, to be "of the same kind," in the sense that the items share a common property with each other or the given items while also possessing an idiosyncratic characteristic. In Feigl's classical terminology, the items are instances of an abstractive concept. Thus, given the items "I often feel tense and jittery" and "I'm an even-tempered person" (measuring emotional stability), we recognize that a further item such as "I feel I am capable of coping with most of my problems" is of the same kind, whereas an item such as "I really like most people I meet" is not. This recognition is on semantic-psychological grounds, not on the basis of a causal theory. My claim, which is comparable in its originality to the discovery that we speak in prose, can be tested against the reader's own practice or experience. What is surprising is the number of investigators whose quite rational practice is to put items together in terms of their common properties, but whose theoretical remarks appear to deny that this is what they are doing.

Consequently, if as claimed above a Spearman factor of empirical item responses is a common property of the item domain from which they are

drawn, and not a common cause, then alternative common factors must be alternative common properties of alternative item domains: sets whose intersection contains the given finite set of items as a subset. To examine this possibility we generalize a classical conception of "singly conforming tests" due to Thomson (1934) into the concept of *singly conforming extensions.* In exploratory factor analysis an investigator sometimes embeds what are referred to as marker variables for a common factor—variables whose common property is believed to be understood—in an added set of variables the conceptual properties of which are not believed understood a priori. The exploratory procedure consists in seeking to establish that the added variables, which we will refer to as the *extension set,* load on the same factor as the marker set in an analysis of the union of the two sets and that the loadings of the markers are at least approximately invariant. Nothing in this procedure requires that the extension set should be drawn, with the markers, from a universal set with a clear denotation. In such a study the factor would be "interpreted" as the attribute already identified as the common property of the markers.

Thomson (1934) showed that given a unidimensional set of m (marker) variables, it is mathematically possible to find two [extension] variables, either of which jointly fit the unidimensional—Spearman—model with the marker set, but when both are added the $m+2$ variables cannot fit a single-factor model in which the original variables keep their original loadings. Thomson described such inconsistent extensions as *singly conforming tests.* In adding two items or tests separately or jointly to a given set of m unidimensional items there are four distinct cases to consider (McDonald, 1977).

We rewrite Equation 2 with all variables standardized, as

$$Y_{0j} = \lambda_{0j} F + \sqrt{(1-\lambda_{0j}^2)}\, E_{0j} \ , j = 1, \ldots, m \ , \tag{9}$$

for a marker set, identified by the added zero subscript. We also have two extension variables Y_1 and Y_2. We write the model for each of the extension variables, when each is added separately, as

$$Y_1 = \rho_1 F + \sqrt{(1-\rho_1^2)}\, E_1 \tag{10a}$$

and

$$Y_2 = \rho_2 F + \sqrt{(1-\rho_2^2)}\, E_2 \ . \tag{10b}$$

(In the Spearman model, with all variables standardized, the loadings are correlations of the observed variables with the factor, as well as being regression weights. Below we consider a situation where the *extension variables* are item composites that estimate the factor. The notation helps cover this.) The four cases are as follows.

Case 1. In this case it turns out that in the joint analysis the m+2 variables fit the unidimensional model because the correlation between Y1 and Y2 happens to be

$$\rho_{12} = \rho_1 \rho_2 \ , \tag{11}$$

(or in a sampling study this holds to an acceptable approximation). In this case the extension variables are jointly conforming to the Spearman model.

Case 2. This is the main case of singly conforming variables treated by Thomson (1951). Here we find that

$$Y_1 = \lambda_1 F \pm \sqrt{(1- \rho_1^2)}\, E \tag{12a}$$

and

$$Y_2 = \lambda_2 F \pm \sqrt{(1- \rho_2^2)}\, E \ . \tag{12b}$$

That is, Y_1 and Y_2 have identical unique components E, so their correlation,

$$\rho_{12} = \rho_1\rho_2 \pm \sqrt{(1 - \rho_1^2)(1 - \rho_2^2)} \ . \tag{13}$$

To borrow a numerical example from Thomson (1951), if $\rho_1 = .8$ and $\rho_2 = .4$, then $\rho_1\rho_2 = .32$, but we might find that

$$\rho_{12} = .32 + \sqrt{(1 - .8^2)(1 - .4^2)} = .87$$

or

$$\rho_{12} = .32 - \sqrt{(1 - .8^2)(1 - .4^2)} = -.23 \ .$$

As Thomson (1951) showed, two such singly conforming variables taken by themselves determine the common factor of the m marker variables precisely. Like the exact Heywood case, such a perfect pair of determiners corresponds to an infinitesimal set of values of the parameters of the model (formally, to a set of points in the parameter space of measure zero), and as Thomson noted, like an exact Heywood case, it stops the process of lengthening the test, as any further variable conforming to the model will "prove the impossibility of its own existence" (p. 232) by making the correlation matrix impossible (not positive definite). Also, in an application it requires the singly conforming items to share their specific properties and errors of replication, which might seem problematic. Even giving the same item twice may not achieve this.

Case 3. Here we find we can write equations

$$Y_1 = \lambda_1 F + \gamma_1 G + D_1 \tag{14a}$$

and

$$Y_2 = \lambda_2 F + \gamma_2 G + D_2. \tag{14b}$$

That is, Y_1 and Y_2 share a part of their separate unique components, plus redefined unique parts D_1 and D_2. In other words, when the extension variables are jointly added to the markers, they define a second common factor and redefine their unique components. For this to occur there is a condition on the correlation ρ_{12}, namely that

$$\rho_1\rho_2 - \sqrt{(1 - \rho_1^2)(1 - \rho_2^2)} < \rho_{12} < \rho_1\rho_2 + \sqrt{(1 - \rho_1^2)(1 - \rho_2^2)} \ . \tag{15}$$

Thus in Thomson's example we must have

$$-.23 < \rho_{12} < .87 \ ,$$

for the joint factor structure to be completed without obtaining an improper solution.

Case 4. We find that the condition of Equation 15 is violated, and we cannot explain ρ_{12} with an added common factor, leaving the loadings of the markers

unaltered. In this case, for the m+2 variables to have a possible (positive definite) correlation matrix, there are still bounds on ρ_{12}, but these are

$$\rho_m^2 \rho_1\rho_2 - \sqrt{(1 - \rho_m^2 \rho_1^2)(1 - \rho_m^2 \rho_2^2)} < \rho_{12} < \rho_m^2 \rho_1 \rho_2 + \sqrt{(1 - \rho_m^2 \rho_2^2)} \ . \tag{16}$$

This is a specialization of a result in McDonald (1977). If the bounds in Equation 15 are not also satisfied, we have a Heywood case, with a negative residual variance in the joint analysis of the m+2 items. Here ρ_m^2 is the squared multiple correlation between the marker variables and the factor F, given by

$$\rho_m^2 = \frac{\sum [\lambda_j^2/(1 - \lambda_j^2)]}{1 + \sum [\lambda_j^2/(1 - \lambda_j^2)]} \ . \tag{17}$$

Suppose, for example, that we have four marker variables each with loading .5. Then, by Equation 17, $\rho_m^2 = .75$, so the bounds are $-.576 < \rho_{12} < .942$, wider than for Case 3.

We see, then, that it is possible to add an (m+1)st item to m marker items for a single factor, in possibly inconsistent ways, to extend the set. Only a joint analysis would show which of four distinct cases occurs.

Now suppose that instead of finding singly conforming item scores Y_1, Y_2, we find two extension item sets $\{Y_{11},\ldots,Y_{1M}\}$, and $\{Y_{21},\ldots,Y_{2M}\}$, each of which fits the Spearman model when analyzed alone, and they respectively have squared multiple correlations ρ_1^2 and ρ_2^2 with their common factors. Suppose we compute the regression estimators of the common factors of the two extension item sets. These are just composite scores which we can then regard as simple extension variables, replacing Y_1 and Y_2 in Equation 10, and they yield one of the cases considered above.

Further, we can suppose each extension set can be taken in the limit to form a test of infinite length. In such a limit we have $\rho_1^2 = \rho_2^2 = 1$. It follows in Cases 1-3 that the correlation between the separate estimators—composites of the extension sets—becomes unity, but in Case 4 the limit gives $2\rho_m^2 - 1 < \rho_{12} < 1$, recognizable as corresponding to Guttman's bound on the correlation between alternative solutions. The joint analysis of the m items with the two tests of infinite length then yields a negative unique variance.

The Case 4 limit appears to supply a possible empirical realization of distinct factor scores with low correlations. Such singly conforming extensions might be approximately realized in applications with large though finite numbers of extension items. An analysis of the union of the marker set and the two extension sets would then yield a Heywood case. We can then equally say that two extensions have been found that might be their common factor, or that no such extension exists. And if in developing the extensions we reach a point where Case 2 holds, no further items can be found to continue the sequences. The argument outlined above is a nontechnical account of that given by McDonald (1977). Closely equivalent treatments are in Mulaik and McDonald (1978) and McDonald and Mulaik (1979). See also McDonald (1978b) for a discussion of the practical consequences of using marker variables and extension sets. This may be summed up as a recommendation against such methods.

Mathematically, Case 3, and therefore possibly Case 4, can be realized by scores from tests of infinite length that are singly conforming when their total

scores are joined to the item scores of the markers. In Case 3 the correlation between scores on the tests of infinite length is unity. In Case 4 it is subject only to the Guttman lower bound. However, there is still need for more detail on the possibility of producing such sequences in empirical work. Seeing how this might be done deliberately may help show if it could happen inadvertently. First, we need to examine the extension process mathematically at the level of item scores and then, by a "thought-experiment," imagine the design followed by corresponding item stems.

It does not seem mathematically possible to create infinite sequences in which each item score in either sequence is singly conforming, that is, continues jointly unidimensional with the marker item scores to yield these tests of infinite length, because to do this each added pair must define a fresh common factor in the union of the sets, hence the two separate sequences do not define a common factor measured by their composite test score.

Instead of attempting to find singly conforming pairs of item scores, we are forced back to the possibility of creating singly conforming test scores with low or minimum correlations. To create Case 3 empirically (and hence make Case 4 possible), we might try to find sets of items measuring a second factor with opposite polarity (positive and negative loadings in the respective sequences, as revealed in a joint analysis). If the joint analysis of the m marker item scores and the two test scores gives a negative residual variance, we have managed to create Case 4. Analysis of the union of the marker set and either extension set at the item-score level, however, will reveal the second factor in the joint structure. That is, the test scores will be singly conforming, but their component item scores will not.

To create such a pair of singly conforming tests, we need actual item stems that are factorially complex and pair up to measure one factor in the same direction and a second factor in opposite directions. For example, we add to a set of marker items for anxiety, say, "I feel anxious when I am alone," or "I feel anxious when I am in a crowd." The hope is that a second dimension of extraversion is being measured in opposed directions. Such a procedure, if successful, would create an empirical Case 3 and thus make possible Case 4. Analysis at the level of item stems would reveal the second factor in either extension and allow it to be interpreted as extraversion. This does not tell us how to design a Case 4—how to create a joint Heywood case deliberately—but only how to make one possible. My rather unsatisfying conclusion is that I do not see how to create Case 4 singly conforming tests deliberately (and so, conversely, to avoid them) in the design of item contents, so I let this statement stand as a challenge to the researchers. Valuable knowledge could be gained from a demonstration of how to do this. On the face of it the problem of finding singly conforming tests whose correlation approaches Guttman's lower bound reduces to the problem of the Heywood case, and we do not yet have full substantive understanding of Heywood cases.

Theory Versus Practice?

First let us review in summary form the points made above.

1. The formal concept of an item domain—a universal set containing a countable infinity of item stems and their response options—provides a most

reasonable foundation for a large part of the theory of the construction and scoring of objective tests measuring psychological attributes.

2. As is well known, the mathematical equations of the single-factor model for a fixed and finite set of items yield arbitrary values for the common and unique scores given the values of the manifest variables unless we have an exact Heywood case. Maximally dissimilar values of *the* general factor have a correlation $2\rho_m^2 - 1$ where $2\rho_m^2$ is its squared multiple correlation with the item scores. Not well known are the implications of this multiplicity of solutions for the construction and application of tests.
3. One possible empirical counterpart of alternative common factors would be alternative possible common causes of the set of item responses. Specific examples of such common causes in the literature seem sufficiently rare to allow a denial that investigators commonly operate this conception with any seriousness. Rather, it appears to represent a commitment to some general philosophical principles, for example, physicalism, that have no detailed implications for the actual practice of writing and lengthening a set of test items and generally would be contradicted by their practice.
4. In the general practice of the psychometric analysis of item scores, the common factors (and in particular the true scores) of items are literally a common property of the items as responded to by the examinees: an abstraction either determining a priori the writing of the item stems or a post facto abstraction from the psychological-semantic characteristics of the items as specific examples of the generic concept they are designed to measure. This is not a mathematical assertion, but a broad claim about the behavior of researchers. Readers may discover its limitations by testing it against their experience of their field.
5. A second possible empirical counterpart of alternative common factors would be scores on alternative, sufficiently long, singly conforming tests, each of which fits the single-factor model with a set of marker variables, while jointly they give a Heywood case. Analysis at the item level will reveal a factorial complexity that shows that the items are not themselves singly conforming.

This completes the summary. Some final speculations now follow. Consider the empirical status of the item-domain conception. It is surely immediately clear that we can invent examples of item sets the denotations of which limit them to a small number of possible distinct elements, although the examples that easily come to mind are knowledge tests with a flavor of triviality about them, such as—noted above—knowing the signature keys of the (m=9) Beethoven symphonies. In principle, vocabulary and arithmetic tests offer domains with clear denotations that few would hesitate to regard as approaching the infinitely large, although rational and empirical analysis of their facets will show that the entire domains are multidimensional. Knowledge of the keys of the (m=104) Haydn symphonies—also noted above—falls somewhere between these extremes. Readers are invited to consult their experience to see how they would settle the *likely* order of magnitude of a denoted domain of item stems and to see if they agree that this will vary widely from case to case.

As implied above, a rather weak test we might apply is whether given m items, we can see how to write an $(m+1)$st item "of the same kind." This is at least a test that we do indeed possess an abstract concept. It is then tempting, although not clearly safe, to suppose that this process can be iterated. That it can be done at all is a weak but important requirement. That it can be iterated indefinitely is a desirable, but strong requirement. We might estimate on a good empirical basis, or in imagination, the number M of items that could be written and tested, given time—the finite universal set from which we suppose our m items are drawn—and use the Spearman-Brown formula to establish the reliability ρ_M^2 of the M items. We might then consider taking $2\,\rho_M^2 - 1$ for the minimum correlation between possible factors of the universal set. But if there are not, *ex hypothesi*, going to be any more, there is no empirical meaning to such alternative factors as further limiting extensions.

The primary objective of this article is to set out relationships between item domains, some standard methods in psychological measurement, and the *determinacy* of psychological attributes. It might be claimed that this objective has been attained at least in sketchy outline. The balance of argument suggests that the justification, in applications, of common-factor or item-response models requires the idealization represented by a denotable infinite item domain. It is not enough to engage in wishful thinking and suppose that such a domain always exists. It can be suggested that (a) researchers should try to make the a priori denotation of an attribute—trait, state, attitude—as precise as possible, implying a clear prescription of exemplar items; (b) they should try to work with attributes for which a "large" number of indicator items can be conceived in principle; and (c) they should avoid exploratory methods of analysis and test extension as far as possible. This is a counsel of perfection and possibly too limiting. The perhaps unsatisfying last word for now is that the infinite item domain will approximate some applications well enough, others not well enough, and some not at all. On the face of it in the latter case we seem to lack an acceptable treatment of the problem of error in the measurement of a psychological attribute. The first question the test constructor must ask is whether a conceptual item domain exists for the attribute to be measured, and if it can reasonably be thought of as *large*. An honest recognition is needed that the answer may be No.

References

Bartholomew, D.J. (1996a). Comment on: Metaphor taken as math: Indeterminacy in the factor model. *Multivariate Behavioral Research, 31*, 551-554.

Bartholomew, D.J. (1996b). Response to Dr. Maraun's first reply to discussion of his paper. *Multivariate Behavioral Research, 31*, 631-636.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Guttman, L. (1953a). Image theory for the structure of quantitative variates. *Psychometrika, 18*, 277-296.

Guttman, L. (1953b). A special review of Harold Gulliksen's "Theory of mental tests." *Psychometrika, 18*, 123-130.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology, 8*, 65-81.

Guttman, L. (1957). Simple proofs of relations between the communality problem and multiple correlation. *Psychometrika, 22*, 147-157.

Guttman, L. (1959). Introduction to facet design and analysis. *Proceedings of the fifteenth international congress of psychology* (pp. 130-132). Amsterdam: North-Holland.

Guttman, L. (1965). The structure of interrelations among intelligence tests. In C.W. Harris (Ed.), *Proceedings of the 1964 invitational conference on testing problems* (pp. 25-36). Princeton, NJ: Educational Testing Service.

Guttman, L. (1969). Review of Lord and Novick's statistical theories of mental test scores. *Psychometrika, 34*, 398-404.

Guttman, L. (1971). Measurement as structure. *Psychometrika, 36*, 329-347.

Kestelman, H. (1952). The fundamental equation of factor analysis. *British Journal of Psychology, Statistical Section, 5*, 1-6.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores, with contributions by Alan Birnbaum*. Reading, MA: Addison-Wesley.

Maraun, M.D. (1996a). Metaphor taken as math: Indeterminacy in the factor analysis model. *Multivariate Behavioral Research, 31*, 517-538.

Maraun, M.D. (1996b). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research, 31*, 603-616.

Maraun, M.D. (1996c). The claims of factor analysis. *Multivariate Behavioral Research, 31*, 673-689.

McDonald, R.P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1-21.

McDonald, R.P. (1977). The indeterminacy of components and the definition of common factors. *British Journal of Mathematical and Statistical Psychology, 30*, 165-176.

McDonald, R.P. (1978a). Generalizability in factorable domains: "Domain validity and generalizability." *Educational and Psychological Measurement, 38*, 75-79.

McDonald, R.P. (1978b). Some checking procedures for extension analysis. *Multivariate Behavioral Research, 13*, 319-325.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R.P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

McDonald, R.P. (1996a). Latent traits and the possibility of motion. *Multivariate Behavioral Research, 31*, 593-601.

McDonald, R.P. (1996b). Consensus emergens: A matter of interpretation. *Multivariate Behavioral Research, 31*, 663-672.

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

McDonald, R.P., & Mulaik, S.A. (1979). Determinacy of common factors. *Psychological Bulletin, 86*, 297-306.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mulaik, S.A. (1996a). On Maraun's deconstruction of factor indeterminacy with constructed factors. *Multivariate Behavioral Research, 31*, 579-592.

Mulaik, S.A. (1996b). Factor analysis is not just a model. *Pure Mathematics, 31*, 655-661.

Mulaik, S.A., & McDonald, R.P. (1978). The effect of additional variables on factor indeterminacy in models with a single factor. *Psychometrika, 43*, 177-192.

Rozeboom, W.W. (1996a). What might common factors be? *Multivariate Behavioral Research, 31*, 555-570.

Rozeboom, W.W. (1996b). Factor-indeterminacy issues are not linguistic confusions. *Multivariate Behavioral Research, 31*, 637-650.

Schonemann, P.H. (1996a). The psychopathology of factor indeterminacy. *Multivariate Behavioral Research, 31*, 571-577.

Schonemann, P.H. (1996b). Syllogisms of factor indeterminacy. *Multivariate Behavioral Research, 31*, 651-654.

Spearman, C. (1929). The uniqueness of g. *Journal of Educational Psychology, 20*, 212-216.

Steiger, J.H. (1996a). Dispelling some myths about factor indeterminacy. *Multivariate Behavioral Research, 31*, 539-550.

Steiger, J.H. (1996b). Coming full circle in the history of factor indeterminacy. *Multivariate Behavioral Research, 31*, 617-630.

Thomson, G.H. (1934). On measuring *g* and *s* by tests that break the hierarchy. *British Journal of Psychology, 25*, 204-210.

Thomson, G.H. (1951). *The factorial analysis of human ability*. London: University of London Press.

Wilson, E.B. (1928). Review of "The abilities of man, their nature and measurement" by C. Spearman. *Science, 67*, 244-248.