Rational/Theoretical Approach to Test Construction

John Ruscio
The College of New Jersey, U.S.A.

One way to develop self-report measures of personality and psychopathology is to use a theory of the target construct to inform the development of items. This is known as the rational/theoretical approach to test construction. Key to this approach is that there must be a rational link between the items' content and the definition and understanding of the construct. Unlike other approaches to test development, the rational/theoretical approach does not entail the collection and analysis of responses to examine the psychometric properties of items, scales, and subscales, nor their ability to differentiate between criterion groups.

The rational/theoretical approach to test construction is a reasonable starting point in the development of many measures, but inadequate by itself. For example, the 93 forced-choice items on the Myers-Briggs Type Indicator (MBTI) and the 16 personality types into which individuals are classified are based on Jungian personality theory. This measure was developed in a rational/theoretical manner, with little or no empirical analysis to refine it. Despite its popularity, the MBTI is of dubious utility and remains highly controversial (Pittenger, 2005). Even over a relatively short test–retest interval (e.g., approximately one month), up to one half of respondents may be classified into a different personality type. At least as worrisome is that individuals frequently report that they disagree with the MBTI classification of their own personalities. Even proponents recommend allowing respondents to review their results to assess their accuracy, which calls into question the validity of the measure itself. These shortcomings might have been avoided had the MBTI's creators performed rigorous empirical analyses to build upon the rational/theoretical foundation.

In contrast, the Millon Clinical Multiaxial Inventory-III (MCMI; Millon, 2008) contains 175 items based on an evolutionary theory and is designed to coordinate with disorders listed in the *DSM*. The items themselves, as well as the 14 personality disorder scales and 10 clinical syndrome scales, were based on theory, but item responses were submitted to factor analyses and compared across criterion groups to refine the items and weight those in the final item pool that contribute to each scale. With each major revision of the diagnostic manual, the MCMI is revised through a new series of empirical analyses. Though not without criticism, MCMI scales exhibit more favorable psychometric properties than do MBTI types, in large part because MCMI scale development extended beyond the rational/theoretical approach to test construction.

Strengths of Rational/Theoretical Test Construction

Three salient characteristics of the rational/theoretical approach to test construction confer important strengths. The first strength is the *simplicity* of the approach. It is an easy method for test developers to put into practice. It requires no special expertise in research methodology, statistical analysis, or psychometrics. The approach demands only a willingness to judge the apparent quality of a measure based on its correspondence with the theoretical definition and understanding of the target construct.

The second strength is the *transparency* of measures constructed using the rational/theoretical approach. If the test developers have done their job conscientiously, the items themselves should represent the target

The Encyclopedia of Clinical Psychology, First Edition. Edited by Robin L. Cautin and Scott O. Lilienfeld.

DOI: 10.1002/9781118625392.wbecp454

construct in a straightforward way and possess strong *face validity.* As a consequence, respondents can see for themselves what is being assessed rather than guess at the purpose of the assessment. This transparency can build trust and motivate conscientious responding.

The third strength is the *intuitive appeal* of the results of a test constructed in a rational/theoretical manner. Whether items are combined to form scales and subscales or individuals are classified into types, the results lend themselves to fairly natural interpretations. The ease of use can increase the understanding and acceptance of test results among users and respondents.

Limitations of Rational/Theoretical Test Construction

Each of the three strengths of the rational/theoretical approach to test construction entails some important limitations as well. The virtue of the approach's simplicity stems from the fact that items are not analyzed empirically after they are developed. Thus, there is also no mechanism for detecting flawed items. Though following the guidelines for item construction in any text on testing and measurement can help to prevent certain foreseeable problems, it remains possible that even an apparently relevant and well-phrased item can function poorly. There is no substitute for rigorous data analysis when determining whether to retain, revise, or remove an item during the test development process. Some common methods for evaluating items include the examination of item-total correlations to ensure that each is positive in sign and nontrivial in magnitude, internal consistency analysis to check whether a set of items forms a homogeneous scale or subscale, and applications of item response theory to assess how effectively items discriminate at varying levels of an underlying trait. The overarching purpose of such analyses is to evaluate and, through an iterative process of test development, improve measurement reliability until it attains an acceptable level. Reliability itself can be evaluated in many ways (e.g., test–retest, alternate or parallel forms, internal consistency), but the core concept is that measured scores should differ due to differences in true scores, with relatively little influence of measurement error. In the absence of rigorous analyses to detect poorly functioning items or to construct homogeneous scales or subscales, it is possible that reliability can be poor when using the rational/theoretical approach to test construction.

The virtue of a measure's transparency also means that responses can be subject to intentional or unintentional biases. For example, individuals differ in their tendency to answer questions in ways that will be evaluated favorably by others, which is known as *social desirability bias.* Overreporting "good" or desirable behavior, and underreporting "bad" or undesirable behavior, can compromise the utility of test scores whether they are used to assess and understand individuals or describe group averages or variability. In contrast, some individuals are more affected by *deviation bias*, a tendency to give unusual or uncommon responses. Such responses can be difficult to distinguish from the valid endorsement of construct-relevant deviant thoughts or behaviors. Another individual difference in response styles is *acquiescence*, which can be conceptualized as a continuum ranging from so-called "yea-sayers" to "nay-sayers." Irrespective of item content, the former tend to respond "yes," "true," or "agree," whereas the latter tend to respond "no," "false," or "disagree." Each of these biases can introduce interpretive problems even though individual respondents may not be acting intentionally to distort test results.

Transparent measures can be especially vulnerable to intentionally deceptive responding. In clinical and forensic assessment contexts, there can be powerful incentives for respondents to "fake good" (e.g., when custody of one's children is at stake) or "fake bad" (e.g., when compensation may be available for disability or following an accidental injury).

Some measures of personality and psychopathology include mechanisms to minimize or at least detect the influence of response biases or deceptive responding. For example, the Minnesota Multiphasic Personality Inventory (MMPI; Greene, 2000) includes items that form a "lie" scale used to detect faking good, an "infrequency" scale used to detect faking bad, a "variable response inconsistency" scale used to detect inconsistent responses to similar (or opposite) pairs of items, a "true response inconsistency" scale to detect responses that tend to be all true (or all false), and other so-called *validity scales.* Even if a test developer is willing to reduce the transparency of a measure to prevent or detect certain kinds of response bias or deceptive responding, many of these tools would be very difficult to incorporate when adhering to a strictly rational/theoretical approach because they require sophisticated strategies of data collection and analysis.

The virtue of test results' intuitive appeal can stem from methods that also compromise a measure's validity. Concerns involving threats to the validity of test results are addressed in the following section.

Situating the Approach Within a Validity Framework

Cronbach and Meehl (1955) provided the classic framework on test validity within which one can examine some additional limitations inherent in exclusive reliance on the rational/theoretical approach to test construction. This classic paper identified three broad kinds of validity: content validity, criterion-related validity, and construct validity. *Content validity* is the extent to which a measure represents all aspects of a construct. Because theoretical understanding guides rational/theoretical test construction, following this approach would be expected to yield measures with strong content validity. Exceptions can involve an overemphasis on features of a construct that make it unique, or on features that it shares with other, related constructs. The potential consequences of misplaced emphasis on subsets of construct-relevant features are discussed below in reference to construct validity. Whereas following the rational/theoretical approach to test construction with proper emphasis on all aspects of the target construct should yield good content validity, even under these conditions the other two kinds of validity may be more questionable.

Criterion-related validity is the extent to which a measure is empirically related to concrete criteria. This kind of validity is usually evaluated by assessing the relationship between test scores and external criteria such as other measures believed to assess the construct validly or membership in groups that the measure is designed to differentiate. *Concurrent validity* is when the external criteria are assessed at the same point in time, and *predictive validity* is when external criteria are assessed in the future. In both cases, one of the chief concerns is whether the measure is sufficiently reliable to achieve acceptable criterion-related validity. Imperfect measurement reliability will constrain the ability of a measure to achieve acceptable criterion-related validity. As noted earlier, unless reliable measurement was established during an empirically informed, iterative approach to test development, it is possible that reliability may be poor. In contrast to the rational/theoretical approach to test construction, the empirical criterion keying approach exemplified by the MMPI is designed specifically to maximize criterion-related validity. This approach comes with its own costs, such as questionable content validity, the danger of capitalizing on chance in test construction if results are not carefully cross-validated in sufficiently large samples, and the need to periodically reassess item functioning because even the initial item selection was based on strictly empirical criteria rather than on face validity. Nonetheless, this approach highlights the value of incorporating explicit tests of criterion-related validity in test development even if one's initial item pool is rationally/theoretically derived.

Construct validity, the extent to which a test actually measures the construct it is intended to measure, is arguably the most important kind of validity. Evaluating construct validity is a complex endeavor. One commonly used tool in construct validation is the multitrait, multimethod matrix (MTMM matrix; Campbell & Fiske, 1959). Distinct constructs, or traits (e.g., anxiety, depression), are each assessed using different kinds of measures, or methods (e.g., self-report questionnaires, clinical interviews). The MTMM matrix is the matrix of correlations of all measures of all constructs, with an appropriate estimate of each measure's reliability replacing its correlation of 1.00 on the diagonal of this matrix. The strongest support for the construct validity of a measure is to observe (a) high reliability, (b) strong correlations with other measures of the same construct (referred to as *convergent validity*), and (c) weak correlations with measures of other constructs (referred to as *divergent validity* or *discriminant validity*).

In addition to the failure to identify poorly functioning items, which can compromise measurement reliability and criterion-related validity, exclusive reliance on the rational/theoretical approach can allow test developers to pay either too much or too little attention to the similarity of a construct with others. For example, if one focuses too narrowly on unique aspects of a construct when creating items (e.g., symptoms of autonomic arousal that distinguish anxiety from depression but do not adequately characterize the central features of all anxiety disorders), the resulting measure may achieve poor convergent validity with measures that provide more comprehensive coverage of the construct. On the other hand, if one emphasizes aspects of a construct that happen to overlap with other constructs when creating items (e.g., experiences of negative affect that are shared by anxiety and depressive disorders), the resulting measure may achieve poor divergent validity.

A final concern is that when investigators working in different theoretical traditions construct measures in a rational/theoretical manner, it is possible that measures of apparently different constructs may in fact measure a highly similar core construct. The belief that differently named measures necessarily correspond to different constructs is known as the *jangle fallacy.* In a striking example of the pitfalls of the failure to establish constructs' theoretical and empirical uniqueness, Judge, Erez, Bono, and Thoresen (2002) presented evidence that measures of three of the most widely studied traits in psychology (self-esteem, neuroticism, and locus of control) plus a fourth trait (generalized self-efficacy) appear to measure a single higher-order, core construct. Had the creators of the many measures of these differently named, yet theoretically and empirically highly similar, constructs taken greater care to examine evidence of convergent and divergent validity, decades of research might not be spread across multiple, extensive, poorly integrated literatures.

SEE ALSO: Construct Validity; Factor Analysis; Internal Consistency Approach to Test Construction; Item Response Theory, Approach to Test Construction; Millon Inventories (MCMI, MACI, M-PACI, MBMD); Minnesota Multiphasic Personality Inventory (MMPI) Instruments; Multitrait Multimethod Analysis; Reliability

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81 – 105.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281 – 302.

Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Needham Heights, MA: Allyn and Bacon.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83,* 693 – 710.

Millon, T. (2008). The logic and methodology of the Millon Inventories. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *Sage handbook of* *personality theory and assessment. Vol. 2: Personality measurement and assessment* (pp. 663–683). Los Angeles, CA: Sage.

Pittenger, D. J. (2005). Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research, 57,* 210–221.

Further Reading

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.