Correcting fallacies in validity, reliability, and classification

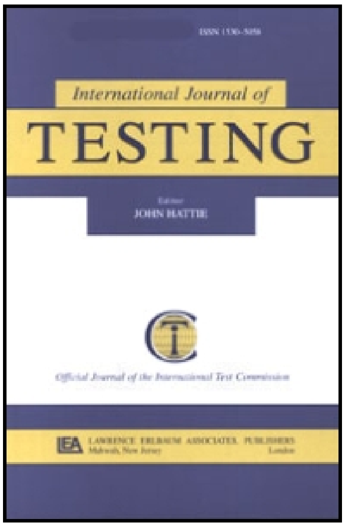
Sijtsma, K.

Published in:
International Journal of Testing

Publication date:
2009

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*(3), 167-194.



International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653658

Correcting Fallacies in Validity, Reliability, and Classification

Klaas Sijtsma [a]
[a] Tilburg University, Tilburg, The Netherlands

To cite this Article Sijtsma, Klaas(2009) 'Correcting Fallacies in Validity, Reliability, and Classification', International Journal of Testing, 9: 3, 167 — 194

To link to this Article: DOI: 10.1080/15305050903106883

URL: http://dx.doi.org/10.1080/15305050903106883

International Journal of Testing, 9: 167–194, 2009
Copyright © Taylor & Francis Group, LLC
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305050903106883

Routledge
Taylor & Francis Group

Correcting Fallacies in Validity, Reliability, and Classification

Klaas Sijtsma
Tilburg University, Tilburg, The Netherlands

This article reviews three topics from test theory that continue to raise discussion and controversy and capture test theorists' and constructors' interest. The first topic concerns the discussion of the methodology of investigating and establishing construct validity; the second topic concerns reliability and its misuse, alternative definitions of reliability, and methods for estimating reliability; and the third topic concerns the relationships between reliability, test length, and the insufficient quality of decision making using short but reliable tests.

INTRODUCTION

Over the past century, starting with the work by Edgeworth, Spearman, and Binet, psychological test theory has shown an impressive growth. This growth is visible in the expanding psychometric theory that guides the construction of tests and also in the huge number of tests constructed using psychometric methods. The classical themes in test theory have been and continue to be validity and reliability. A test is valid if it measures the attribute of interest and reliable if it measures this attribute with high precision.

In this review article, I focus on three different topics. The first topic concerns the status of psychological attributes and the methodology of investigating and establishing construct validity; the second topic concerns reliability and present practices of estimating reliability and alternative approaches to reliability; and the third topic concerns the relationships between reliability, test length, and the insufficient quality of decision making using short but reliable tests. Each topic

I wish to express my gratitude to Denny Borsboom, John Hattie, Rob R. Meijer, Sanneke Schouwstra, and two anonymous reviewers for providing critical comments on a previous draft of this manuscript, and to Wilco H. M. Emons for preparing some of the computations presented in this paper. The opinions expressed in this paper remain mine.

Correspondence should be sent to Klaas Sijtsma, Department of Methodology and Statistics, FSW, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: k.sijtsma@uvt.nl

is important enough to warrant a separate article. However, the topics are connected in more than tangential ways and, moreover, by the continued importance for test theory and test construction. This justifies discussing them in the same article.

The first topic, construct validity, has proven to be difficult perhaps mostly because of the unclear status of the psychological attributes a test purports to measure. Presently, the dominant conceptions seem to be realism, which claims that attributes exist as real entities; constructivism, which considers them as constructions only to be known through samples of behavior; and rationalism, which equates them to what the test measures. The dominant methodology of the past decades, which was the investigation of the attribute's nomological network, has led to some frustration among researchers because in principle this methodology implies an endless array of research projects, which threatens to result in indecisiveness with respect to construct validity. Several suggestions on how to solve this problem have been recently made. I will discuss some of these suggestions and prompt one in particular.

The second topic, reliability assessment, is not so much difficult but instead fascinates because of the persistence with which psychological researchers ignore useful information on reliability use. This information has been around in the psychometric literature for a long time but somehow is not picked up well. For example, even though it has long been known that coefficient alpha is not suited for evaluating a test's internal consistency, textbooks keep promoting this use of alpha. In addition, despite being almost the worst reliability estimate available, alpha remains by far the most used estimate. I will reiterate the arguments against the internal consistency use of alpha and suggest other methods for estimating a test's reliability. In addition, I will briefly discuss reliability estimation by means of generalizability theory and structural equation modeling, which seem to introduce validity issues into reliability assessment. The question is whether one should blend matters of validity and reliability or separate them.

The third topic is the growing popularity of short tests consisting of say, only ten items, for decision making about individuals. Short tests are used because they relieve the burden on patients, little children, and impatient clients and managers that is caused by longer tests, and the short tests' use is justified because their test scores often have surprisingly good reliability provided that only high-quality items are used. However, even with the best items available, using only a small number of items will necessarily result in relatively large error variance at and around the cut-score used for decision making. I will show this and argue that important decisions require long tests.

Much as the three topics seem different, they also touch and even mingle in interesting ways. Validity enters reliability assessment when coefficient alpha is interpreted as an index for internal consistency and when generalizability theory or structural equation modeling are used to estimate coefficients. Generalizability theory assesses the degree to which test scores can be repeated, for example, across different item formats, assessment modes and time points, and structural equation modeling assesses repeatability, for example, across the common factor shared by all items but not across other systematic score components such as group factors. An important question then becomes whether reliability must be limited to repeatability of test performance except random measurement error (classical true-score theory), or whether considerations regarding generalizability beyond the test (generalizability theory) or wanted and unwanted systematic components of the true score (structural equation modeling) should be allowed to become issues in reliability.

In classification of people in important diagnostic categories, short but reliable tests are seen to have relatively large standard measurement errors and, consequently, wide confidence intervals for true scores, which cover a large segment of the (short) scale. This results in large proportions of classification errors, which reduce the importance of a high reliability and advances the use of confidence intervals for true scores. Large proportions of classification errors render short tests doubtful for individual diagnosis.

A consequence of the problems mentioned is that test construction and test practice are plagued by bad habits. Construct validity is often ascertained by means of highly exploratory research strategies and is in need of more direction; reliability is often estimated using one of the worst methods possible and is given an incorrect interpretation; and due to practical testing demands short tests are becoming increasingly popular, but their use results in many more classification errors than the use of longer tests. Progress in test theory, both in the development of psychometric methods and their practical use in test construction, is slow. Not only are several validity and reliability issues unresolved or at least continue to be at the center of much debate, novel insights also seem to have trouble finding their way to the community of psychological researchers; see a recent discussion on this theme by Borsboom (2006a, 2006b) and Clark (2006), Heiser (2006), Kane (2006), and Sijtsma (2006). The main goal of my review is to critically discuss some present-day ideas and practices and to suggest alternatives.

ATTRIBUTES, CONSTRUCT VALIDITY, AND CHOOSING THE RIGHT METHODOLOGY

I will use the term "attribute" to represent mental properties such as verbal ability, spatial orientation, and anxiety. Three perspectives seem to capture present-day thinking on the status of, say, attribute A, and the way the validity of a measurement instrument for attribute A is ascertained. I will not discuss the different uses of a particular measurement instrument for attribute A and how these different uses are validated; see the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 9; also Messick, 1989) for a discussion on the need of separately validating these different uses.

The Constructivist Viewpoint: Attributes as Constructions

The constructivist viewpoint is that cognitive abilities and personality traits are constructed in an effort to better understand human behavior. This view implies that attributes are products of the psychologist's imagination and that they do not exist in reality (Nunnally, 1978, pp. 94–109). In addition, human behavior is the source of inspiration of theorizing about attributes. Thus, attributes are the offspring of observable behavior and not the other way around: attributes are not the cause of behavior. We are deeply inclined to think about our behavior in terms of cause-effect schemes. Anxious behavior is caused by anxiety and intelligent behavior by intelligence. In the constructivist view, anxiety and intelligence are convenient constructs that serve to delineate sets of behaviors that hang together and are different from other sets of behaviors. Theories about attributes explain which behaviors hang together and why and also how an attribute relates to other attributes.

Like any theory, the theory about an attribute starts out as an idea, which, in its initial state, is just as immaterial and volatile as other ideas. The more thought the researcher gives to the attribute the more likely his ideas about the attribute will transform into a theory that links the attribute to other attributes, speculates about the relationships between them, and relates the attribute to behavioral correlates. When the theory of the attribute has gone through several stages of development, including logical reasoning, critical observation of people's behavior that is assumed to be typical of the attribute of interest, and empirical research on (particular aspects of) the attribute, it may be possible to formulate hypotheses about these behaviors. The hypotheses predict which behaviors hang together as correlates of the attribute of interest, how the behaviors relate to behaviors typical of other attributes, and how the manipulation of typical behaviors can affect other typical behaviors. Such hypotheses can be tested empirically, and the results may lead to the support or the adaptation of the theory of the attribute.

The constructivist approach to construct validation rests on theory testing, and Nunnally (1978, pp. 98–105) gives a precise description of the process of investigation. In their classical approach to construct validation, Cronbach and Meehl (1955) place the test at the heart of a nomological network, which plays the role of the attribute's theory. The nomological network consists of laws that relate attributes to one another and to observable properties. These laws are tested empirically as hypotheses about the relationships of the attribute of interest with the other attributes in the network using test scores to represent each but also by testing group differences with respect to test scores, score change over time, and the internal structure of the items in the test. In this conception of construct validity, attributes are constructs that may be adopted, not demonstrated to be correct, let alone discovered, and the inferences made by means of the test are validated but not the test itself (Cronbach & Meehl, 1955; also AERA, APA, & NCME, 1999).

The *practice* of construct validation often is different from the *ideal* just described in that theory testing regularly is replaced by ad hoc collecting results on some branches in the nomological network, devoid of a guiding theory. Thus, one often sees that test constructors start with a set of items based more on habit and intuition than on established theory guiding the choice of these items and then work backward to find meaning for the test scores. In this process, knowledge of the test score's interpretation accumulates gradually as more relationships with other attributes and background variables such as age and educational level are clarified. Although aspects of this validation strategy are not unlike what one reads in Cronbach and Meehl (1955), replacing theory testing by exploration and ad hoc correlating variables, which happen to be available, may not be the spirit these authors encouraged. It certainly weakens the conclusions compared with those based on a confirmatory approach.

Several researchers (e.g., Borsboom, Mellenbergh, & Van Heerden, 2004; Embretson & Gorin, 2001) have noted that investigation of the nomological network does not aim at revealing the processes, cognitive or otherwise, that were stimulated by the items and led to the item scores. They consider knowledge of these processes to be crucial for establishing a test's construct validity and notice that the process of validation through the nomological network leaves open the possibility that these processes remain *terra incognita*. Even though I think it would be possible to include the study of processes generating item scores into the nomological network, Embretson and Gorin (2001) seem to be less hopeful and argue that Cronbach and Meehl's (1955) methodology precludes the possibility to learn about the cognitive theory underlying item performance.

Finally, it may be noted that the constructivist view leaves an undeniably unsatisfactory feeling about the status of attributes; however, this is not to say that the view is wrong. The dissatisfaction originates from the shaky foundation of constructs, which is the mind of the psychologist. This foundation, or perhaps the lack of it, conveys a kind of arbitrariness to the status of cognitive abilities and personality traits. For many people it probably is more convenient to think of attributes as real entities waiting to be discovered. When it comes to the definition and the measurement of constructs, one feels a little like the baron Von Münchhausen, who pulled himself out of a swamp by his bootstraps. The next section discusses the realist conception of attributes.

The Realist Viewpoint: Attributes as Real Entities

The realist view assumes that mental attributes are real entities. I will take the recent and notable approach proposed by Borsboom et al. (2004) as an example.

These authors adopt realism as their point of departure for outlining an approach to validity research. Briefly, they posit that the measurement of an attribute by means of a test is valid if the attribute exists and if variations in the attribute cause variations in the measurement outcomes of the test. The authors claim that the realist viewpoint agrees with how many psychologists would define construct validity, which psychologists say is about the question whether a test measures what it should measure, but is at odds with much practical validity research, which rather seeks to ascertain the *meaning* of the test scores, not *what* the test measures, after the instrument has been constructed and the data have been collected. Borsboom et al. (2004) consider this latter approach typical of the test validation practice that was advocated if not at least stimulated by Cronbach and Meehl (1955).

Given the realist assumption about attributes, Borsboom et al. (2004) argue in favor of validity investigation that tests the psychological theory of the attribute of interest. This investigation starts with knowing what one wants to measure, and then constructing the set of items that elicit the cognitive processes, which produce the responses to the set of items and the differences between people in response propensity to these items. The resulting sets of item scores are to be the topic of meticulous validity research that has the study of the cognitive processes at its focus. This can only be done when a theory of the attribute is available to such detail that the test can be constructed as a manifestation of the structure of this theory and the item scores can be considered as the result of the cognitive processes involved. Advanced psychometric modeling of the item scores collected by means of the test is then used to test the attribute theory. When the psychometric model fits the data, the theory is supported and measurement is valid. Borsboom et al. (2004) notice that a fitting model may not settle all loose ends of a theory, but the principle stands clear.

As an example, Borsboom et al. (2004) mention the measurement of cognitive development by means of the well-known balance scale task (Jansen & Van der Maas, 1997; Siegler, 1981; Van Maanen, Been, & Sijtsma, 1989). Balance scale tasks are based on a theory of developmental phases, each of which is characterized by typical solution rules. As development progresses and children move into the next phase, they adopt new rules for solving balance scale tasks. Assuming that the theory is true and that the test consists of items that elicit responses that are informative about the developmental phases posited by the theory, a child's test score reveals his/her level of cognitive development. Jansen and Van der Maas (1997) used latent class modeling to demonstrate that the data structure predicted by their balance-scale test corresponded with the empirical data structure. Hence, according to the realist viewpoint they demonstrated that their test is valid (give or take a few loose ends). Other examples of test construction and theory testing that are subsumed under this validation strategy concern the measurement of verbal ability (Janssen & De Boeck, 1997), spatial ability (Embretson & Gorin, 2001), and perceptual classification (Raijmakers, Jansen, & Van der Maas, 2004).

Another example used later on in this section comes from the measurement of the developmental attribute of transitive reasoning (Bouwmeester & Sijtsma, 2004). Here, the item set used was the result of longstanding research (e.g., see Verweij, Sijtsma, & Koops, 1999, for an earlier attempt). This research eventually led to the testing of three competing theories of transitive reasoning by fitting three multilevel latent class models to the item response data, each model reflecting the formal structure of one of these theories (Bouwmeester, Vermunt, & Sijtsma, 2007). It was concluded that fuzzy trace theory (Brainerd & Kingma, 1984; Brainerd & Reyna, 2004) better explained the data structure than Piaget's operational reasoning theory (Piaget, 1947) and Trabasso's linear ordering theory (Trabasso, 1977). (To avoid confusion, it may be noted that the word "fuzzy" refers to the level of detail of information processed cognitively, not to mathematical terminology.) Given this result and following Borsboom et al.'s line of reasoning, a test for transitive reasoning is valid when its items elicit responses that are informative about the cognitive processes described by fuzzy trace theory. Hence, by showing that fuzzy trace theory explains the data structure, Bouwmeester et al. (2007) demonstrated construct validity for their transitive reasoning test.

Thus, the theory of the attribute is the driving force behind test construction, and the validation of the test resides in showing that the theory adequately predicts the responses to the items in the test. If the theory has been shown to do this, the measurement of the attribute is valid.

The Evolution of Theory and the Role of the Nomological Network

Taken as a forceful plea for theory development as the basis of test construction and measurement, Borsboom et al.'s (2004; also see Borsboom, 2005) view is highly valuable (also, see Embretson, 1983; Embretson & Gorin, 2001). It promotes a mentality of taking the attribute's theory as point of departure for test construction and testing hypotheses to challenge the theory. By doing this, it also takes a powerful stand against exploratory research using preliminary sets of items chosen primarily on the basis of habit and intuition and then claiming that correlations between these items provide evidence of the measurement of an attribute. It is different from approaches like Nunnally's (1978) by asking proof that indeed the test produces measurements that are the result of particular cognitive processes and reflect individual variation regarding these processes. This limitation to cognitive processes also limits the approach, as I will point out shortly.

I also think that Borsboom et al.'s approach does not necessitate considering attributes as real entities waiting to be discovered. The realist assumption seems to ignore how psychological theories about attributes in general originate and develop. Attributes initiate as the result of observing behavior and noting that some behaviors occur together more often than others and seem to have characteristics in common that they do not share with other behaviors. In daily life, such observations are often unstructured if not haphazard but in psychology the observer, once triggered by a particular phenomenon, will look more closely and purposefully. This will lead him to set up a theory and test hypotheses on real data, which in turn will not only affect the theory but also the conception of the attribute.

Thus, following their origin, attributes seem to be constructions that serve to summarize related behaviors, and setting up and testing a theory then amounts to demarcating the set of behaviors. For attributes such as abilities, the set of behaviors may be broadened to include cognitive processes, and in the personality domain the broadening may involve affects and emotions. I am not saying that overt behaviors, processes, affects, and emotions just happen without any causation (e.g., when faced with a tiger a person may show certain fearful behaviors), only that observers see behavior and infer attributes from behavior, not the other way around.

To illustrate this position, I will further discuss the development of transitive reasoning. Initially, transitive reasoning was equated with logical reasoning on the basis of memory of premises (Piaget, 1947): Given the premises that stick A is longer than stick B, and stick B is longer than stick C, the child is capable of transitive reasoning when he can correctly infer from the memorized premise information that stick A is longer than stick C. In the next step of theory development, linear ordering theory (Trabasso, 1977) replaced logical reasoning, typical of Piaget's operational reasoning theory, with the formation of an internal representation of the linear ordering of the objects. For example, in a 5-object task, given an ordered presentation of the premises (e.g., consecutive premises contain longer sticks), an ends-inward internal representation identifies the first presented object A as the shortest and the last presented object E as the longest and interpolates the other objects in between. Given that such internal representations are available, children were assumed to be able to infer the unknown transitive relation by reading the internal representation from memory rather than by using rules of logic.

More recently, fuzzy trace theory has brought the insight that incoming information is encoded and reduced to the essence in so-called traces, which may differ in degree of detail. Verbatim traces contain exact information, and gist traces contain fuzzy, reduced, pattern-like, information that only holds the gist. Different verbatim traces and different gist traces exist next to one another, and incoming information is processed in parallel at different trace levels. Children are inclined to use the least-detailed trace (i.e., the one holding the most global information), which is least-demanding in terms of memory workload, for inferring a transitive relationship. An example of gist trace information is objects become smaller to the left (e.g., in a series of five ordered objects, such that $A < B < C < D < E$), which is easier to use in solving whether A is smaller than E than the verbatim exact premise information. The use of this gist trace renders both logical reasoning on the basis of premises and also memory of the premises superfluous.

This example shows that the theory of transitive reasoning has developed and, with it, the attribute itself. Initially, transitive reasoning was conceived of as a pure logical reasoning ability, which later evolved via the reading of an internal representation into a more complex ability involving memory of encoded logical information, and then finally into the ability of using the least-demanding piece of coded information to solve what formally looks like a logical problem. Thus, even though it still refers to being able to solve the same kinds of tasks, today transitive reasoning is defined to be something else than a few decades ago. Due to progressing insight based on research, the conception of transitive reasoning has changed over time (also, see Borsboom et al., 2004, p. 1063).

Now, the three consecutive theories may be seen as phases one passes before the objective truth is finally uncovered. I think this is the wrong perspective for two related reasons. The first reason is that the perspective leans heavily on the model of physical measurement, which often has been put forward as the role model for psychological measurement (e.g., Michell, 1999, pp. 193–211). However, in physics, scales are the result of strong theories and extremely precise and highly replicable experimentation. In psychology, attributes and theories are much more abstract and vague, and in experiments subjects are difficult to manipulate. As a result, counts of particular kinds of responses are often weakly related, and results cannot be replicated well. Thus, results are not compelling and scales are arbitrary to a high degree. Given this state of affairs, one can only hope to get more grips on human behavior, not to uncover the truth.

Because in psychological research so much is elusive, psychological theory testing often stops when psychologists are convinced that they know enough about an attribute and how it must be measured to solve a particular practical problem, but not so much because the truth has been found. For example, hypothetically, research in transitive reasoning may stop when educational psychologists think their test is good enough for diagnosing a child's persistent problems with arithmetic tasks. This attitude known as *practicalism* aims at finding solutions to practical problems, not necessarily to how things are, and is the second reason why theory development does not lead to uncovering the truth. Michell (1999, p. 96) considered practicalism to stand in the way of scientific research beyond the more practical goals. He may be right, but replacing practicalism by a quest for the truth does not alleviate the uncertainty that characterizes psychological research and stands in the way of obtaining firmer results.

One reviewer put forward the thesis that fuzzy traces must have been present in childrens' minds before fuzzy trace theory was posited and found to be the most explanatory of the three competing theories. Based on what I know now, I prefer to think of fuzzy trace theory as a way of acknowledging that the mind likes to take shortcuts (gist information) rather than go all the way (verbatim information), and that traces function as metaphors rather than real things to be discovered eventually. More generally, changes in the conception of transitive reasoning have been fed by inferences from the study of overt behavior, and claims about transitive reasoning as a material cause of this behavior have been unnecessary to come this far.

There are two additional problems. First, Borsboom et al.'s (2004) proposal seems to be too rigorous in its condemnation of the nomological network. Second, it is not suited for the majority of mental attributes.

Regarding the first problem, I agree with Borsboom et al. (2004) that working backward in the nomological network to unravel the meaning of the test score is not a very productive validation strategy. The problem is that this strategy produces large numbers of correlations that often are interpreted ad hoc without necessarily leading to theory formation and the testing of theory. Cronbach (1988) and Kane (2001) made a similar point by labeling this research strategy as the *weak program* of construct validity. Also, it is difficult if not impossible to determine when one is done investigating a construct's nomological network. Thus, construct validation becomes a never-ending process (Messick, 1988) and the question of whether a particular test indeed measures a particular attribute is never answered definitively (Borsboom, 2006a).

However, Borsboom et al.'s (2004) suggestion to equate construct validity to testing a cognitive theory about the attribute does not solve this problem. I assume that this cognitive theory is narrower than its nomological network, which also contains relationships of the attribute with different attributes and group variables that are not needed per se to understand the processes described by the theory but may add to learn as much as possible about the attribute (e.g., Kane, 2001). For example, given that fuzzy trace theory, the least formalized of the three competing theories, best explains the item scores, we may conclude that the ability of transitive reasoning does not rest completely on the application of pure logic or on the bare retrieval of information from memory. Instead, other abilities may be involved and we may learn about this by expanding the theory of transitive reasoning to include relationships with other attributes and background variables; that is, parts of the nomological network of transitive reasoning may be taken into account. Embretson and Gorin (2001) restricted this expansion to the study of the *nomothetic span*, which only includes implications of the theory. Research of the nomothetic span of transitive reasoning may produce results that affect the theory and call for additional testing of particular aspects of the theory. Thus, the theory evolves and the nomological network plays a role in this evolution.

The second problem is that Borsboom et al.'s (2004) proposal seems to be suited particularly for relatively simple, albeit important cognitive abilities that are closely related to some form of logical reasoning. However, most attributes of interest in psychology are broadly defined, formulated in abstract terms and often not very articulated and, consequently, rather uncritical regarding direct behavioral correlates. Examples are creativity, emotional intelligence, leadership, and empathy, as well as many well-accepted personality traits (e.g., Maraun & Peters, 2005). As a result, attribute theory is primitive and tests and questionnaires necessarily may be relatively blunt instruments. Here, theory testing is exploratory, suggesting hypotheses rather than testing them (also, see Kane, 2001), but may lead to a better understanding of the attribute, which in turn may lead to better items, and so on. The level of articulation is lower than that assumed by Borsboom et al. (2004) and leads to weaker grips on the quality of measurement of the attribute. With such attributes one needs all the information one can collect, and the attributes' nomological network is a good place to start looking.

The Rationalist Viewpoint: Attributes as Expert Agreement

As an aside, it is interesting to note that discussions on validity are current also in marketing research. Rossiter (2002) recently challenged Churchill's (1979) classical psychometric approach (strongly inspired by Nunnally, 1978) to developing measures of marketing constructs (e.g., product perception, consumer satisfaction) by putting forward an approach in which expert agreement determines what the test measures and how it should be measured. Expert agreement means that expert raters determine whether the measure is valid. Empirical research determining relations in a nomological network from collected data or the testing of the theory behind the item scores by means of latent variable models is deemed unnecessary. Thus, content validity based on expert judgment takes over from construct validity based on empirical research, a proposal that not only test constructors would consider rather militant but also raised resistance among marketers. For example, Diamantopoulos (2005; also, Finn & Kayande, 2005) criticized Rossiter for adopting a rationalist approach to validity—extremely stated, an approach that considers empirical research unnecessary—and condemning construct validation by means of studying the relations with other attributes as in a nomological network.

For psychological measurement, Kane (2001) suggested restricting expert judgment to so-called observable attributes, such as simple arithmetic items and geometric analogy items, which he claimed "are defined in terms of a universe of possible responses or performances, on some range of tasks under some range of conditions" (ibid., p. 332). The definition need not involve theory. In his view, evidence supporting the content validity of observable attributes need not only come from experts but may also come from experience and knowledge accumulated with the definition, relevance, and representation of particular domains and the statistical analyses of test data. Opposite observable attributes, Kane (2001) distinguished *theoretical constructs*, which involve theory testing in the context of a nomological network as outlined here.

CONCLUSION

It may be noted that Kane's suggestions seem to be somewhere halfway between Borsboom et al.'s and Rossiter's. However, Borsboom et al. would regard all attributes as *theoretical constructs* whereas Rossiter would equate all of them with *observable attributes*, which can be validated solely through expert judgment (for the sake of argument, I am ignoring all kinds of nuances in their opinions and Kane's.)

Borsboom et al.'s (2004; also, see Borsboom, 2006a) and Rossiter's (2005; also see Bergkvist & Rossiter, 2007) approaches to the measurement of attributes and the foundation of attribute measurement are extremely different except for one aspect: They share a strong discontentment with the nomological network as a basis for validity research (unlike Kane) and come up with a realistic solution (attributes exist and exercise causation) and a rationalist solution (an attribute's condition is decided by experts).

Regarding Rossiter's position, I think that for psychological measurement, rater agreement (i.e., an impoverished version of content validity) cannot replace the empirical study of construct validity; at best, it may be part of it. As concerns Borsboom et al.'s position, I think that the proposal to place psychological theory construction at the heart of construct validity is a very fruitful idea, but not without studying the relations with other important attributes. For attributes for which sound (cognitive) theory is available, this could mean studying the attribute's nomothetic span (Embretson & Gorin, 2001). For the multitude of weakly defined attributes this probably means taking one's refuge in the nomological network (i.e., Cronbach's weak program of construct validation). Taking (part of) the nomological network into account helps to further develop the attribute's theory. One does not preclude the other, and together they stand stronger (e.g., Kane, 2006).

RELIABILITY AND BAD HABITS, REMEDIES, ALTERNATIVES

Typically, discussions about reliability are more mathematical than discussions about validity. The reason is that reliability formalizes one particular technical property of the test score whereas validity involves the assessment of what the test measures, either by means of the exploration of the nomological network or by means of theory testing. Unlike reliability estimation, validity research revolves around the use of substantive knowledge about the attribute under consideration and decision making with respect to which hypotheses to test, how to test them, how to decide whether they received support or not, and how to continue validity research in either case. Consequently, validity research produces many statistical results whereas the final assessment may take the form of a verbal summary resulting in judgment. Reliability is a much narrower concept and has been found to be easier to quantify. This does not mean that it is undisputed. Before I discuss controversial issues, I introduce some notation.

Let the score on item j be denoted by random variable X_j, and let the test contain J items. An individual's test performance often is expressed by the test score, which is defined as $X_+ = \sum_{j=1}^{J} X_j$. I will use the classical definition of reliability proposed by Lord and Novick (1968). The basis of classical test theory as defined by Lord and Novick (1968, pp. 29–30) and also present in the stochastic subject formulation of item response theory (Holland, 1990), is an individual's *propensity distribution*. This is an individual's distribution of test scores that would result from the endless administration of the same test under the same circumstances, such that different test scores are the result of independent repetitions of the administration procedure. The correlation between only two of these repetitions in the population of interest is the test-score reliability, $\rho_{XX'}$. The variation in an individual's propensity distribution is interpreted as measurement error variance. Thus, a test score's reliability answers the question: What would I find when I could do it all over again? In practical research, only one test score is available for each individual; hence, the reliability cannot be estimated as the correlation between two independent repetitions. Often, approximations based on one test score are used.

Reliability and Internal Consistency

The most common approximation to test-score reliability is coefficient alpha (Cronbach, 1951). Alpha provides an underestimate, often referred to as a lower bound, to the reliability, $\rho_{XX'}$, so that $alpha \leq \rho_{XX'}$. Equality is attained only under unrealistic conditions so that in practice strict inequality holds: $alpha < \rho_{XX'}$. Rather than using alpha purely as a lower bound, test constructors typically use alpha as a measure of the test's internal consistency. The psychometric literature does not define this concept well (Sijtsma, 2009a). For example, Cronbach (1951, p. 320) noted that an internally consistent test is "psychologically interpretable" although this does not mean "that all items be factorially similar" (also, see Cronbach, 2004, pp. 394, 397–398). He also used internal consistency and homogeneity synonymously (but see Cronbach, 2004, p. 403), whereas Schmitt (1996) distinguished the two concepts and claimed that internal consistency refers to the interrelatedness of the items in a set and homogeneity to the unidimensionality of the items in a set. Sijtsma (2009a) provided a more elaborate discussion of internal consistency and concluded that the common denominator of all definitions is something like "the degree to which the items in the test are associated due to what they share in common." Thus, one could argue that alpha is not only interpreted as a lower bound to the test-score reliability but also as an aspect of the test's construct validity, expressing the degree to which the data are unidimensional or 1-factorial. One would expect that one statistic could not express two concepts, reliability and validity, that are so much different, but this truism has not withheld test constructors and test users from simply doing this.

Sijtsma (2009a) argued that alpha is not a measure of internal consistency; see Bentler (2009), Green and Yang (2009a), and Revelle and Zinbarg (2009) for more discussion. He constructed data sets that are 1-factorial, 2-factorial, and

3-factorial, that each had the same alpha value. Thus, tests of highly different factorial composition can have the same alpha. Reversely, without further knowledge about the test's factorial composition any realistic alpha value may indicate a 1-factorial data structure but just as well structures typical of two or more factors, even when the factors are uncorrelated. This boils down to statements like: Knowing that $alpha = .7$, holds no information about the factorial composition of a test, and neither does knowing that $alpha = .2$ or $alpha = .9$. This is not a revolutionary novel insight (e.g., Cortina, 1993; Schmitt, 1996) but a reiteration of arguments needed to fight old and highly resistant abuse of coefficient alpha. Internal consistency, no matter how one defines it precisely, needs to be evaluated by methods such as factor analyses and item response theory (e.g., Hattie, 1985), and is best seen as part of construct validity research.

Other Estimates of Test-Score Reliability

Coefficient alpha has persisted in being by far the most popular estimate of test-score reliability, and has done so in the face of conflicting evidence about its appropriateness as a reliability estimate. Already since 1945, a coefficient known as lambda2 (Guttman, 1945) is known that is related to alpha and reliability as: $alpha \leq lambda2 \leq \rho_{XX'}$. Thus, if $alpha = .8$ and $\rho_{XX'} = .9$, lambda2 has a value between .8 and .9. Experience has shown that lambda2 usually is within .01 of alpha so that the gain of using lambda2 is small. However, from a rational point of view there is no reason to consistently report the smallest lower bound; yet this is common practice. Lambda2 is available in SPSS, as the second estimate given under the option called *Guttman* (alpha is the third estimate).

Greater lower bounds than lambda2 exist. For example, Guttman (1945) proposed six different methods including alpha and lambda2 that are all available in SPSS, and Ten Berge and Zegers (1978) proposed a series of progressively increasing lower bounds, of which alpha and lambda2 are the first two members. The problem of finding the greatest lower bound (glb) has been solved some 30 years ago (Bentler & Woodward, 1980; Ten Berge, Snijders, & Zegers, 1981). The glb can be considerably higher in real data than alpha and lambda2 (e.g., up to .07; see Sijtsma, 2009a, for a real-data example) but thus far suffers from bias problems. In particular, when the number of items in the test exceeds 10 and the sample size is smaller than 1,000, the glb tends to be inflated. This is a statistical problem in need of a solution, especially as the glb by definition comes closest to the test's true reliability (also, see Ten Berge & Sočan, 2004). Ten Berge and Kiers (2003) and Bentler (2009) provided computer programs for estimating the glb. Clearly, there is much to be gained here, meanwhile acknowledging that a small lower bound such as alpha does not really harm anyone but also does little to provide the test constructor with a realistic reliability estimate.

Alternatives to Reliability

Three interesting alternatives to classical reliability are the generalizability coefficient, reliability based on structural equation modeling, and the test information function.

Generalizability Theory. Depending on the domain of measurements to which one wishes to generalize conclusions based on test scores, the generalizability coefficient (Brennan, 2001a; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) expresses repeatability of test scores across different test versions, different time points, different item formats, and so on. Thus, the concept of generalizability expresses the idea that we may not only be interested knowing what would happen when the test could be repeated under exactly the same circumstances, but acknowledges that these circumstances could be varied in interesting ways.

Thus, the generalizability coefficient broadens the concept of repeatability from independent repetitions to almost any set of conditions relevant to the use of a particular test. This is interesting, but one should realize that by doing this, certain aspects of the test's validity may be introduced into the reliability concept, and a sensible question is whether one should do this or keep validity and reliability sharply separated. For example, generalizability across time refers to the degree to which the attribute is liable to change over time. This change may be due to training or learning processes, and may become manifest as a quantitative score shift on the same scale. But change may also be due to spontaneous developmental processes and may become manifest qualitatively, as in the disappearance of particular processes and the emergence of novel processes, apparent from a change in the items' factorial composition (Embretson, 1991). Another example is generalizability across test versions using different item formats. This refers to the exchangeability of different item formats used for measuring the same attribute and may be interpreted as an instance of convergent validity.

My intention is not to question the usefulness of the generalizability concept but only to note that, although similar at the surface, it may be basically different from reliability (however, see Brennan, 2001b, for a different view). For example, if I am interested to know my body weight, I would like the scale I use at home to be reliable in the sense that it does not fluctuate more than, say, one ounce when I step onto it a couple of times in a row. For my purpose, which is to know my body weight today using my own scale, it does not help to know that the correlation is .8 between a set of measurements collected today and another set collected six months ago. Body weight changes over time, which is a characteristic of body weight, not of the scale, and so I would not expect a perfect correlation anyway. Also, it does not help to know that my scale results correlate .7 with the results obtained by means of another scale as long as I do not know anything about possible mechanical differences between the scales or whether a scale is technically flawed.

Of course, it goes without dispute that generalizability theory answers highly useful questions in, for example, separating test-score variance due to test form from variance due to individual variation in the attribute of interest. My point in this discussion is that one should be careful when introducing validity issues into reliability matters so as to prevent confusion among the people who work with them in practice. The issue of distinguishing validity from reliability is an old one; see Campbell and Fiske (1959).

Structural Equation Modeling. Reliability estimation on the basis of structural equation modeling amounts to estimating the common factor for the set of items, and possible group factors that capture variance shared only by subsets of items but not by all items in the test (Bentler, 2009; Green & Yang, 2009b; Raykov & Shrout, 2002). This provides the possibility to estimate reliability as the ratio of common variance (which cannot exceed the true-score variance) over total variance, which by definition cannot exceed the classical reliability. Another possibility is to separate variance hypothesized to be due to response styles from the common variance or to separate variance due to the nesting of items within, for example, text fragments (as in text comprehension tests) or variance due to learning processes that are active during testing and affect item performance more as items are administered later in the test (Green & Yang, 2009a) from the common variance. A key question to be answered by the researcher is which variance components have to be included in the numerator of the reliability coefficient and which components may be seen as error.

Thus, in this approach concerns about the composition of the test-score variance have become important in determining the test-score components across which repeatability of test performance should be considered. In other words, repeatability is not limited to the test score per se but is applied to desirable test-score components, thus excluding unwanted components, which are not only random (measurement error, as in true-score reliability) but also may be systematic (e.g., due to response styles). Again one allows validity considerations to enter the reliability concept, but it is an unresolved issue whether the resulting blur is a good idea or one, which confuses test users.

Test Information Function. The test information function (e.g., Embretson & Reise, 2000, pp. 183–186) expresses the accuracy by which the latent variable is estimated using maximum likelihood methods as a function of the scale for the attribute. The latent variable often is interpreted to represent a trait or an ability, which drives examinees' responses to the items in the test. In item response theory, the inverse of the test information function gives the standard error of the maximum likelihood estimate of the latent variable conditional on the true value of the latent variable. This conditional standard error provides an answer to the question: What would happen when I would draw another examinee sample of equal size from the population of interest and estimate this latent variable value again? This looks like reliability but it is different because, in order to be the same concept, we would also have to consider drawing scores from an individual's propensity distribution after a new sample has been drawn; see Holland (1990) for a discussion of this stochastic subject interpretation, and Borsboom (2005, chap. 3) for arguments in favor of a random sampling interpretation without additional sampling from propensity distributions.

Conclusion. Because of its greater flexibility allowing many highly interesting applications in a more general latent variable framework, item response theory has gained greater popularity than generalizability theory. As a result, the information function has become a more popular tool than the generalizability coefficient. What both theories have had working against them for a long time compared with classical test theory and factor analysis is their lower degree of accessibility for researchers who have not been trained as applied statisticians. Thus, the greater accessibility of classical test theory has led many test constructors and researchers to stay with test-score reliability as an index of how accurate their test measures. Reliability estimation within the framework of structural equation modeling is relatively new and has not been applied much so far. The use of true-score reliability in psychological test construction (but also elsewhere, as in marketing and health-related quality of life measurement) still by far outnumbers the use of the other methods. Because of its continued popularity, I will use test-score reliability estimated as the proportion of true-score variance, $\rho_{XX'}$, in what follows.

What Does a Particular Reliability Value Mean?

One often reads in test manuals or papers reporting on test construction or the use of tests in research that test-score reliability, usually in the form of lower bound coefficient alpha, equalled, say, .8. The author then goes on to say that this is sufficiently high for his purposes. What does such a reliability value imply for the magnitude of score differences to be significant, or for a score to be significantly different from a given cut-score? Or, even more important, for the consistency by which individuals are classified in the right treatment category? Such questions are not new but they are often left unanswered, probably because the answers are difficult. These questions are even more relevant in the light of present-day developments in test construction and test use toward shorter tests that maintain the same reliability level than their longer counterparts. The goal of the next section is to argue that short tests having high reliability cannot be trusted for classifying individuals, which often is the primary goal of individual testing. This conclusion not only questions the usefulness of short scales but also emphasizes the limited usefulness of the reliability concept.

TEST LENGTH, MEASUREMENT INACCURACY, CLASSIFICATION CONSISTENCY

It is well known that short tests tend to have lower reliability than long tests. Yet the present era shows a persistent development toward the construction and the use of short tests in individual testing. The idea is to use only the best items for measurement, and by doing this, reduce administration time. The best items measure as accurate as possible in the region of the scale where accuracy is needed, such as a cut-score, and not necessarily somewhere else. These items have been selected such that they also cover the essence of the attribute, and not unique factors that would attenuate a test's unidimensionality. A small elite of items is thought to do the measurement job as well as a larger set that also contains many weaker items.

What is a long test? An example is the NEO big five inventory (Costa & McCrae, 1992), which consists of five subtests for measuring the big five personality traits, each subtest containing 48 items, thus adding to 240 items in total. Another example is the WISC (originally, Wechsler, 1949), a child intelligence test battery, which consists of 15 subtests for measuring different aspect of intelligence, making up several hundreds of items in total, the exact number depending on the test version under consideration. In contrast with these lengthy tests are the Mini Mental State Examination (Folstein, Folstein, & McHugh, 1975), which contains 11 items, the test for Pathological Dissociative Experiences (Waller, Putnam, & Carlson, 1996) consisting of 8 items, the test for Alcohol Drinking Behaviors (Koppes et al., 2004) holding 7 items, and the Test Anxiety Inventory (Taylor & Deane, 2002), consisting of 5 items. Authors of short tests generally claim that the psychometric properties are good enough to justify the use of the test scores for individual decision making.

The majority of tests and questionnaires consist of tens of items and are located somewhere between very short and very long. I will concentrate on the very short tests, and argue that they are insufficiently reliable for individual decision making. This conclusion does not rule out that test scores based on few items may be adequate for use in scientific research in which group mean scores and correlations are of primary interest. This kind of test use is perfectly justifiable. I also acknowledge that the use of short tests has its practical advantages and that this use may be necessitated by financial restraints on testing time and physical and mental restraints on examinees. For example, little children are not able to concentrate for a long time and patients may simply be too confused or too sick to be able to answer tens of demanding questions. As real and compelling as these restraints may be, they cannot make unreliable measurement reliable.

Short Tests and Standard Measurement Errors

Neither classical test theory nor item response theory assume that the error variance is the same for different individuals. However, applications of tests constructed by means of classical test theory use one standard measurement error for all individuals. Applications of item response theory use a standard deviation of the estimated latent variable conditional on the true value, which varies in magnitude across the scale. The individual's propensity distribution can be used to illustrate that short tests involve great risks in individual decision making. For simplicity, I follow the classical practice of using one standard measurement error for all measurement values.

For person v, let the true score be defined as $T_v = E(X_{+v})$ and estimated simply by $\hat{T}_v = X_{+v}$ (to keep things simple, I refrain from using Kelley's formula; Kelley, 1947). The standard error for $\hat{T}_v$ is known as the standard measurement error. In practical applications of test scores it is assumed to be the same for each individual from the population of interest. Let the standard deviation of the test score be denoted by S_{X_+}, then the standard measurement error equals $S_E = S_{X_+}\sqrt{1 - r_{XX'}}$ (in this section, I use sample notation). The standard measurement error is used for constructing confidence intervals for the true score T_v, for example, a 95% interval, defined by the bounds $X_{+v} \pm 1.96 S_E$. When this interval contains a cut-score, say, X_c, it is concluded that $\hat{T}_v$ does not differ significantly from the cut-score; and when the cut-score is outside the interval, it is concluded that $\hat{T}_v$ differs significantly from the cut-score. It may be noted that estimating confidence intervals like this assumes that the test score has a normal distribution. In particular for short tests, this assumption is wrong but I will ignore this in this section.

For person v and person w, the question whether two test scores X_{+v} and X_{+w} are significantly different from one another is answered as follows. The null hypothesis is H$_0$: $T_v = T_w$ and the standard measurement error of the difference $D_{vw} = X_{+v} - X_{+w}$ equals $S_{E(D)} = \sqrt{2} S_E$. A 95% confidence interval is obtained from $D_{vw} \pm 1.96 S_{E(D)}$. When this interval contains the value 0 the null hypothesis is accepted, and when 0 falls outside the interval the null hypothesis is rejected. Thus, one may check whether $|D_{vw}| < 1.96 S_{E(D)}$.

I present results from a small computational study that clarifies the relation between test length, reliability, standard measurement error, and scale length. This study provides insight into the chance mechanism responsible for the failure of short tests as decision instruments, even when they consist of high-quality items. Data sets were simulated using the Rasch (1960) model. Let θ denote the latent variable, δ_j the difficulty of item j, and a the item discrimination constant, which is the same for all J items. The Rasch model is often scaled such that $a = 1$, but given a fixed standard normal distribution of θ as I will use here, this is impossible and a becomes visible in the model equation; that is,

$$P(X_j = 1|\theta) = \frac{\exp[a(\theta - \delta_j)]}{1 + \exp[a(\theta - \delta_j)]}.$$

I simulated 20 data sets for 500 respondents, taking all combinations of $J = 6$, 8, 10, 12, 20 and $a = 1, 2, 3, 4$. The item difficulties were equidistant between –1 and 1. Increasing a values stands for increasing item quality: for a standard normal θ, $a = 1$ is modest, $a = 2$ is good, $a = 3$ is very high, and $a = 4$ is extremely high,

TABLE 1
Guttman's Lambda2, Standard Measurement Error, Half Confidence Interval for True Score, and Half Confidence Interval for True Score Difference

Realistic	J	a	*Lambda*2	S_E	$1.96S_E$	$1.96\sqrt{2}S_E$
Yes	6	1	.5515	1.0711	2.0994	2.9689
		2	.7378	.9212	1.8055	2.5534
No		3	.8085	.8360	1.6386	2.3173
		4	.8486	.7699	1.5090	2.1341
Yes	8	1	.6006	1.2381	2.4267	3.4318
		2	.8103	1.0699	2.0970	2.9656
No		3	.8656	.9533	1.8685	2.6424
		4	.8785	.9054	1.7824	2.5096
Yes	10	1	.6565	1.3964	2.7369	3.8706
		2	.8331	1.2072	2.3661	3.3462
No		3	.8848	1.0753	2.1076	2.9806
		4	.9119	1.0034	1.9667	2.7813
Yes	12	1	.7091	1.5213	2.9817	4.2168
		2	.8686	1.2968	2.5417	3.5945
No		3	.9101	1.1667	2.2867	3.2339
		4	.9301	1.0838	2.1242	3.0041
Yes	20	1	.7990	1.9816	3.8839	5.4927
		2	.9168	1.6747	3.2824	4.6420
No		3	.9431	1.5109	2.9614	4.1880
		4	.9566	1.3966	2.7373	3.8712

almost impossible to reach in practice. I will consider tests with $a = 3$ and $a = 4$ as upper benchmarks.

Table 1 shows that for fixed J, as a increases, Guttman's (1945) lambda2 increases and the standard measurement error decreases. For $J = 6$ and $a = 1$ (moderate discrimination), the table shows that when a test score is less than 3 units in range of a cut-score, it is not significantly different from that cut-score. The same conclusion is true for score differences; $|D_{vw}| < 3$ is not significant. It must be noticed that these differences span half the scale. Imagine a ruler of 6 cm in length, which could only reliably distinguish objects that differ at least 3 cm in length. Of course, a scale for psychological attributes is different from a ruler, but the analogy may help to grasp the degree of inaccuracy that we face when using a very short test containing items of moderate discrimination. For $J = 6$ and $a = 2$ (good discrimination), the situation improves for testing against a cut-score but not for testing score differences, and for higher, albeit unrealistic upper benchmark a values, the situation remains the same.

For increasing J, score differences that have to be exceeded to be significant increase slowly. What happens here is the following. As the test becomes longer, the true scores T_v and T_w run further apart and so do the observed scores X_{+v}

and X_{+w}, on average. Also, as Table 1 illustrates, the standard measurement error grows and, as a result, the confidence interval grows. The fascinating thing is that as one keeps adding items to the test the distance between true scores and between corresponding observed scores grows faster than the length of the confidence interval. Consequently, score differences D_{vw} that initially are inside the confidence interval, move out of it as J grows, resulting in the rejection of the null hypothesis of equal true scores. This is readily formalized for strictly parallel tests (Lord & Novick, 1968, pp. 47–50). Assume a J-item test is extended with $K - 1$ other J-item tests, and that all K tests are parallel. Then, it can be shown that the initial true score difference, say, T_{vw}, becomes KT_{vw} and the initial standard measurement error $\sqrt{K}S_{E(D)}$. This shows that the true-score difference grows by a factor $\sqrt{K}$ faster than the standard measurement error of the difference and the corresponding confidence interval, thus clarifying my point.

The issue of measurement inaccuracy could be tackled from different angles, such as item response theory, and the use of the standard measurement error surely is not optimal. However, use of the standard measurement error is simple, makes the point well and, not unimportant, the standard measurement error continues to be more popular in test construction and test use than any other measure of inaccuracy. I conclude that very short tests are highly inaccurate measurement instruments, no matter how good the quality of the items used, assuming one stays within realistic limits. The amount of statistical information simply is too small. This problem may be alleviated readily by using more items without ending up with an excessively long test (see Table 1).

Short Tests and Classification Consistency

The gravity of two scores having to be at least, say, 3 points apart to be significant depends on the gravity of the decisions made by means of the test score. This can be studied mathematically in simple and lucid examples, thus providing straightforward lessons for the length of real tests. Assume a simple binary decision with respect to treatment using a cut-score that separates the scale in two exhaustive and disjoint segments. In general, the probability of making the right decision for an individual is larger the further an individual's true score lies away from the cut-score. Two proportions are important here (Emons, Sijtsma, & Meijer, 2007; for different approaches, see Bechger, Maris, Verstralen, & Béguin, 2003; Ercikan & Julian, 2002; Hambleton & Slater, 1997).

Consider an individual v's propensity distribution of which the true score T_v is the expected value, as before. Let T_v be located at the right of a cut-score X_c ($T_v > X_c$) that is used to classify people. Thus, based on his true score this person belongs in the category to the right of the cut-score, which corresponds, for example, with treatment. When I would repeatedly draw scores X_{+v} at random from individual v's propensity distribution, a sensible question is which proportion (say, P_v) of these scores would classify individual v at the right side of the cut-score. Now, assume that for John this proportion equals $P_{John} = .6$ and that for Mary it equals $P_{Mary} = .8$. Assuming normal propensity distributions (which are unrealistic for short tests), this means that John's true score is closer to the cut-score than Mary's. It may be noted that knowing these proportions does not provide information about the test's quality as a decision-making instrument. A way to formalize this information is the following.

Let us assume that a decision-maker will only feel confident to make a decision about an individual when the proportion of correct decisions based on draws from the individual's propensity distribution exceeds an *a priori* determined lower bound. This lower-bound proportion is called the *certainty level* (Emons et al., 2007). The certainty level results from policy considerations, which depend on the type of problem for which one seeks to make decisions and also on financial recourses and so on. Important decisions made under strict financial and other constraints will require a high certainty level. For the sake of argument, I choose the certainty level (denoted π) to be equal to $\pi = .9$, meaning that I only feel confident about a decision when at least 90% of the draws from an individual's propensity distribution classify him correctly.

Given a particular test and a particular decision-making policy reflected by certainty level π, the crucial question is which proportion of individuals whose true score is in one particular category are classified in this category (i.e., the right decision) on the basis of a proportion of draws from their propensity distributions that is at least equal to π. This proportion of individuals is the test's classification consistency (CC) in a particular population. Given that $\pi = .9$, for an artificial population consisting only of John and Mary (and assuming that their true scores place them in the same category, for example, treatment), we have that $P_{John} = .6 < \pi$ and also $P_{Mary} = .8 < \pi$; thus, the test's *CC* for treatment equals 0. This result reflects that the test does not provide enough evidence for making individual decisions with sufficient certainty that they are the right decisions for providing treatment. For the other category of no treatment the typical CC may also be determined.

An example of the situation sketched thus far is that a particular medical treatment must be provided to patients who really need it but not to others because of potentially damaging side effects (for simplicity, I will only consider the treatment group). A cut-score delineates the two groups and a high certainty level, such as $\pi = .9$, reflects a particular level of caution. It is important that patients whose true score is in the treatment category receive treatment and thus as few of these patients as possible should be classified incorrectly as not needing treatment. A perfectly reliable test would do the job (assuming perfect construct validity), but real tests have fallible reliability. For fallible tests the question thus is which *CC* value they produce. The closer to 1 the *CC* value is, the better the test functions as a classification instrument.

For given certainty level π, Emons et al. (2007) mathematically investigated the influence of test length on *CC*. The latent variable was normally distributed. Also, persons' true scores were assumed known. This was necessary to compute *CC*s. Decisions about treatment or no treatment were based on one test score and one cut-score. Although these are strong simplifications of real decision-making situations, they allow for clear-cut answers to difficult questions, thus getting better grips on the problem. For a fixed value of certainty level π and all but one design factor kept constant, Emons et al. (2007) found that *CC* for treatment ($T \geq X_c$) was larger as the number of items J was larger, item discrimination was higher, item difficulty was closer to the cut-score on the latent-variable scale, and items were polytomous instead of dichotomous. *CC* was smaller as the cut-score was more extreme (i.e., fewer people had a true score that necessitated treatment).

Some interesting detailed results for certainty level $\pi = .9$ were that for a 6-item test consisting of dichotomous items for which $a = 1.5$ (i.e., between modest and good) and all items located exactly at the cut-score (which is the ideal situation), and half of the group needing treatment (based on their true scores), the *CC* was only .46. All other things kept equal, for $J = 8, 10, 12, 20$, the *CC*s increased to .53, .58, .61, .70, respectively. For $a = 2.5$, which is a very high item discrimination power, for $J = 6, 8, 10, 12, 20$, the *CC*s were equal to .66, .70, .74, .76, .82, respectively. As the group who needed treatment became smaller, the *CC*s also became smaller; for example, when this group consisted of the 5% highest T scores, for $J = 6$ and $a = 1.5$, the $CC = .17$. The interested reader can find the complete results in Emons et al. (2007).

It is illuminating to let the meaning of these *CC*s sink into ones mind: If the decision is important ($\pi = .9$), a short 6-item test consisting of items with quite good discrimination power will classify only 46% of the patients who really need the treatment with enough certainty. For a 20-item test, the $CC = .70$. This may look good but actually means that for 30% of the people who are in need of treatment, the test does not provide enough certainty to select them as candidates.

The conclusion from this research was that tests must be long if one wants to make good decisions. This is true even if the test consists of good-quality items. A sample of only a few item scores cannot consistently classify people correctly. Although this conclusion may be at odds with practical constraints typical of particular testees such as young children and patients, reliable information comes at a price, which involves a relatively long test and a long administration time. Of course, this time may be spent administering several short tests, possibly at different occasions, and combined afterwards to reach a consistent decision.

GENERAL CONCLUSION

The three issues in test theory discussed in this article are by no means new, and the insights presented not revolutionary. However, the practice of test construction does not easily absorb the developments offered by psychometrics and sometimes makes choices in the face of conflicting psychometric evidence (also, see Borsboom, 2006a), thus running into bad practices and in the longer run even bad habits. Novel psychometric results may not be accepted for very different reasons. One reason may be that test constructors may not recognize a new method as particularly relevant for their problems. Perhaps the method only solved a problem that looked small and unimportant from the test constructor's perspective, or perhaps the method was potentially important but the psychometrician failed to provide convincing examples of its application in test construction. Another reason may be that new methods have a tendency to be statistically more complex than old methods, and thus more difficult to access, and then they have to provide big and visible improvements over the older methods to become accepted in practice. But improvements may be small or fail to be obvious.

However, I think that the issues discussed in this article are plagued by different problems. Validity assessment simply is difficult, reliability estimation seems to be dominated by wrong habits, and diagnosing individuals is under pressure by practical demands. In particular, construct validity is difficult because the scientific status of the attributes is liable to dispute and attribute theories as a rule are simplistic, incomplete, vague, abstract, or sometimes almost absent. This causes lack of direction, fragmentary exploratory validation research, but also solutions that are so strict that they may fail to be applicable to the majority of attributes.

Reliability estimation is not so much difficult but plagued by strong habit, which has created a persistence in using old but inferior lower bounds, coefficient alpha in particular. The problem also is in the statistical complexity of alternatives, such as the glb, and estimation based on generalizability theory and structural equation modeling, which are not readily available to test constructors through a simple mouse click (Sijtsma, 2009b). Also, the use of alpha as index for the test's internal consistency is known to be untenable, but this has not withheld test constructors from using this interpretation.

Short tests are becoming more popular, for example, in clinical and medical psychology, health-related quality of life measurement, but also in organizational psychology because they relieve the burden on testees and because it is believed that the high reliability that can sometimes be realized for short tests is sufficient for accurate decision-making. However, it can be shown that a high reliability in combination with a small score range makes it difficult to accurately distinguish a test score from a cut score, and thus is likely to result in many classification errors.

My point is not that these positions are new but that they are not well known nor well accepted in large areas of test construction and test use. Based on the discussion presented here, my recommendations are the following.

Construct validity does not follow from first assembling a measurement instrument and then trying to derive its meaning from a painstaking exploration of parts of the nomological network of the attribute of interest, so as to obtain evidence afterwards. Instead, whether the instrument measures the attribute has to follow from the testing of sound substantive theory (Borsboom et al., 2004), which supports the construction of the test for a particular attribute. Part of construct validation is the testing of this theory through the data collected by means of the test. Further support for the test's validity comes from investigating the test's nomothetic span (e.g., Embretson & Gorin, 2001), which limits the nomological network to a manageable size. However, it seems to me that this size may be increased given the degree to which the attribute under consideration has been developed. Experience accumulated thus far suggests that construct validation remains a long-term enterprise, mainly due to the unfinished nature of psychological theories. McGrath (2005; for reactions, see Barrett, 2005, Kagan, 2005, and Maraun & Peters, 2005) provides a discussion on problems in construct validity of personality traits, and Zumbo (2007) provides a general overview of validity issues.

Reliability is most often estimated by means of coefficient alpha, which somewhere after 1951 began a dual life as a reliability coefficient and an index for internal consistency. Alpha is not a sensible indicator of a test's internal consistency (e.g., Cortina, 1993; Schmitt, 1996; Sijtsma, 2009a). For that purpose, one should use factor analysis or other methods specifically designed to unravel the dimensional structure of the data. Alpha is a lower bound to test-score reliability, but it has been shown a long time ago that it is a small lower bound compared with other, higher, and sometimes readily available lower bounds. There is no reason not to use these greater lower bounds or other methods for reliability estimation, such as structural equation modeling.

Another issue is that to know that a test's reliability equals .8 or its standard measurement error equals 1.4 does not mean much when this is not related to the kind of decision making for which the test is intended. This becomes a more acute problem as test length is smaller, which is a trend in present-day test use in clinical, medical, and organizational contexts. Tests must contain, say, at least 20 good-quality items for that purpose. More important is that test constructors' attention shifts from reliability, standard measurement errors, and other indicators of accuracy to classification consistency and related concepts. Attention in test construction and test use must shift again (and back) to decision making using test scores (e.g., Cronbach & Gleser, 1957, 1965). Clearly much exiting work remains to be done here.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington D.C.: American Educational Research Association.

Barrett, P. (2005). What if there were no psychometrics?: Constructs, complexity, and measurement. *Journal of Personality Assessment, 85*, 134–140.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*, 319–334.

Bentler, P. A. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137–143.

Bentler, P. A., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika, 45*, 249–267.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*, 175–184.

Borsboom, D. (2005). *Measuring the mind. Conceptual issues in contemporary psychometrics.* Cambridge UK: Cambridge University Press.

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Borsboom, D. (2006b). Can we bring about a velvet revolution in psychological measurement? A rejoinder to commentaries. *Psychometrika, 71*, 463–467.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Bouwmeester, S., & Sijtsma, K. (2004). Measuring the ability of transitive reasoning, using product and strategy information. *Psychometrika, 69,* 123–146.

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review, 27,* 41–74.

Brainerd, C. J., & Kingma, J. (1984). Do children have to remember to reason? A fuzzy-trace theory of transitivity development. *Developmental Review, 4,* 311–377.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24*, 396–439.

Brennan, R. L. (2001a). *Generalizability theory.* New York: Springer.

Brennan, R. L. (2001b). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317.

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

Churchill, Jr., G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64–73.

Clark, L. A. (2006). When a psychometric advance falls in the forest. *Psychometrika, 71*, 447–450.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418.

Cronbach, L. J., & Gleser, G. C. (1957, 1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

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.

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

Diamantopoulos, A. (2005). The C-OAR-SE procedure for scale development in marketing: A comment. *International Journal of Research in Marketing, 22,* 1–9.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (1991). A multidimensional latent variable model for measuring learning and change. *Psychometrika, 56*, 495–515.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343–368.

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

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12,* 105–120.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*, 269–294.

Finn, A., & Kayande, U. (2005). How fine is C-OAR-SE? A generalizability theory perspective on Rossiter's procedure. *International Journal of Research in Marketing, 22,* 11–21.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 2*, 189–198.

Green, S. A., & Yang, Y. (2009a). Commentary on coefficient alpha: a cautionary tale. *Psychometrika, 74*, 121–135.

Green, S. A., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155–167.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19–28.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Heiser, W. J. (2006). Measurement without copper instruments and experiment without complete control. *Psychometrika, 71*, 457–461.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Kagan, J. (2005). A time for specificity. *Journal of Personality Assessment, 85*, 125–127.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2006). In praise of pluralism: A comment on Borsboom. *Psychometrika, 71,* 441–445.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Koppes, L. L. J., Twisk, J. W. R., Snel, J., Mechelen, W. van, & Kemper, H. C. G. (2004). Comparison of short questionnaires on alcohol drinking behavior in a nonclinical population of 36-year-old men and women. *Substance Use and Misuse, 39*, 1041–1060.

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

Maraun, M. D., & Peters, J. (2005). What does it mean that an issue is conceptual in nature? *Journal of Personality Assessment, 85*, 128–133.

McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment, 85*, 112–124.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). Washington, D.C.: American Council on Education.

Michell, J. (1999). *Measurement in psychology. A critical history of a methodological concept.* Cambridge UK: Cambridge University Press.

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

Piaget, J. (1947). *La psychologie de l'intelligence*. Paris: Collin.

Raijmakers, M. E. J., Jansen, B. R. J., & Van der Maas, H. L. J. (2004). Rules and development in triad classification task performance. *Developmental Review, 24,* 289–321.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*, 195–212.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145–154.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19,* 305–335.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development,* 46(2), Serial No. 189, 1–74.

Sijtsma, K. (2006). Psychometrics for psychologists: Role model or partner in science? *Psychometrika, 71*, 451–455.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*, 169–173.

Taylor, J., & Deane, F. P. (2002). Development of a short form of the Test Anxiety Inventory (TAI). *The Journal of General Psychology, 129*, 127–136.

Ten Berge, J. M. F., & Kiers, H. A. L. (2003). *The minimum rank factor analysis program MRFA*. Internal report, Department of Psychology, University of Groningen, The Netherlands; retrieved from http://www.ppsw.rug.nl/∼kiers/.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613–625.

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika, 46,* 201–213.

Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43,* 575–579.

Trabasso, T. (1977). The role of memory as a system in making transitive inferences. In R. V. Kail, J. W. Hagen, & J. M. Belmont (Eds.), *Perspectives on the development of memory and cognition* (pp. 333–366). Hillsdale, NJ: Erlbaum.

Van Maanen, L., Been, P. H., & Sijtsma, K. (1989). Problem solving strategies and the Linear Logistic Test Model. In E. E. Ch. I. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–287). New York/Berlin: Springer.

Verweij, A. C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23*, 241–264.

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods, 3*, 300–321.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children.* New York: Psychological Corporation.

Zumbo, B. D. (2007). Validity: foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam: Elsevier, North Holland.