PSYCHOMETRIKA—VOL. 77, NO. 1, 4–20
JANUARY 2012
DOI: 10.1007/S11336-011-9242-4

FUTURE OF PSYCHOMETRICS: ASK WHAT PSYCHOMETRICS CAN DO FOR PSYCHOLOGY

KLAAS SIJTSMA

TILBURG UNIVERSITY

I address two issues that were inspired by my work on the Dutch Committee on Tests and Testing (COTAN). The first issue is the understanding of problems test constructors and researchers using tests have of psychometric knowledge. I argue that this understanding is important for a field, like psychometrics, for which the dissemination of psychometric knowledge among test constructors and researchers in general is highly important. The second issue concerns the identification of psychometric research topics that are relevant for test constructors and test users but in my view do not receive enough attention in psychometrics. I discuss the influence of test length on decision quality in personnel selection and quality of difference scores in therapy assessment, and theory development in test construction and validity research. I also briefly mention the issue of whether particular attributes are continuous or discrete.

Key words: change assessment, decision quality based on short tests, didactics of psychometrics, personnel selection, test-quality assessment, test validity, theory construction.

1. Introduction

From 2005 till 2010, I was chairman of the Dutch Committee on Tests and Testing (COTAN). COTAN works under the auspices of the Dutch Association of Psychologists (Nederlands Instituut van Psychologen; NIP). COTANs mission is to advance the quality of tests and the quality of test use in The Netherlands, and to inform test users about developments concerning availability of novel tests, test use, legislation, and so on. COTAN advances quality by evaluating as many tests and questionnaires as possible that are used in the Dutch practice of psychological and educational testing for individual assessment. Most of these applications are in the areas of job selection, career management, clinical diagnosis, and educational assessment. Test constructors and test publishers ask COTAN to evaluate their tests but in principle COTAN takes the initiative for evaluation with or without permission given by third parties. Tests and questionnaires nowadays are also constructed and used in medical and health practice, for example, for measuring quality of life, pain experience, and motor development. COTAN also assesses these tests but concentrates on psychology and education.

The 15 members of COTAN are all psychologists who are experienced test users or test constructors, and have at least a basic knowledge of test construction and psychometrics. They are affiliated with Dutch universities or work as applied psychologist in health care, trade or industry. Their main responsibilities are the assessment of tests and keeping the evaluation system used for test assessment up to date. Recently, Evers, Sijtsma, Lucassen and Meijer (2010) published an English-language description of the Dutch-language revision of the evaluation system. COTAN members assess tests and additional reviewers are recruited from a reviewer base that consists mainly of test constructors and psychometricians who have a keen interest in test construction and the practical use of tests. They must use the evaluation system for test assessment. Each test

This article is based on the author's Presidential Address, presented at the International Meeting of the Psychometric Society 2011, July 18–22, 2011, Hong Kong, China.

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

is assessed on seven criteria as insufficient, sufficient, or good. The seven criteria are: theoretical basis of the test, quality of the test materials, comprehensiveness of the manual, norms, reliability, construct validity, and criterion validity. Assessment results based on two independent reviews and edited by a COTAN member who serves as Editor are commercially published online in the Dutch language.

Evers et al. (2010) published test-quality assessments obtained in four waves: 1982 (# tests = 236), 1992 (# tests = 299), 2000 (# tests = 372), and 2009 (# tests = 540). Rather than reiterate detailed results, I focus on some main conclusions that are of interest to psychometricians. First, norms were problematic in 1982 and improved only gradually to a level that still is worrisome: In 2009, 14 percent of the tests were rated good, 28 percent sufficient, and 56 percent insufficient. The two main problems seem to be that the sample sizes are often too small and that samples are not representative, given the measurement goals of the test. For criterion validity, the trend does not show improvement at all: In 2009, the percentages were 7 (good), 21 (sufficient), and 59 (insufficient). Modest improvements are available for reliability and construct validity but percentages for assessment category "good" of 27 (reliability) and 18 (construct validity) seem low from an absolute point of view. To summarize, Dutch psychological test quality has improved over the past 30 years but much terrain remains to be conquered.

COTANs work has strategic edges that serve to boost test quality. For example, test constructors predominantly use classical test theory (CTT) and principal component analysis to analyze their data and draw conclusions with respect to their tests. My guess is that psychometricians would be inclined to provide quality assessments that are less favorable than those produced by typical COTAN reviewers, simply because test data are analyzed using older methods, but rarely using modern psychometric methods published in *Psychometrika* and other measurement journals. Even though a case can easily be made for the use of superior modern methods, the truth is that many researchers either are not familiar with those methods or are hard to convince of their merits over classical methods. Then, assessing tests by the standard of modern psychometrics when almost no one (yet) uses these methods would lead to an avalanche of negative assessments, the effect of which would soon be that the test community would turn their back to COTAN, rendering its work ineffective. COTANs policy is to gently push the Dutch test community toward using better methods without asking them to use methods the advantages of which are not entirely clear to them and that might only overstretch their technical skills. This means, for example, that item response theory (IRT) is not the norm for test construction even though most psychometricians would prefer its use to the use of CTT. Instead, a test constructed using principles of CTT receives positive assessments provided the work was done well according to COTANs evaluation system.

The Dutch psychological testing community holds COTAN in high regards. Governmental institutions often require a published COTAN assessment for tests to be used under their responsibility and insurance companies require a positive assessment before providing compensation for treatment to their clients. Thus, COTAN offers an excellent opportunity for psychometricians to feel proud of psychometrics. However, COTAN also made me painfully aware of the huge gap that exists between the practice of test construction and the theory of psychometrics. For example, until I was COTAN chair it never bothered me that much that test constructors attach so much value to coefficient alpha (Cronbach, 1951) and, moreover, that they do not know that alpha is one of the worst estimates of test-score reliability, that many alternatives superior to alpha exist, and that alpha cannot be used to assess whether the test is unidimensional or 'internally consistent' (Sijtsma, 2009a). This contribution discusses what I learned from my COTAN years for teaching psychometrics both to students and researchers, and about the research topics that may be of special interest to the practice of constructing psychological tests.

2. Teaching Psychometrics: Some Common Misunderstandings

I devote some space to the teaching of psychometrics. This is rather unusual in *Psychometrika*, which does not have something like a "teacher's corner", but I believe it is very useful for a field of research that can only flourish by the successful dissemination of its results among researchers, primarily in psychology but also in other areas. I mention some widespread misunderstandings about psychometrics that exist among test constructors and psychological researchers that construct or use tests in their research. I also briefly discuss possible causes of the misunderstandings and what psychometricians could do in their teaching to prevent these misunderstandings from occurring as much as possible. My treatment is based on personal experience and may be biased and incomplete. All examples use coefficient alpha, as it appears to be the most used method in practical psychometrics. I skip results from my previously published treatment of coefficient alpha (Sijtsma, 2009a).

2.1. Coefficient Alpha is a Lower Bound for Some Tests but not for Others

A well-known theorem in psychometrics is that coefficient alpha is smaller than or equal to the test-score reliability (e.g., Novick & Lewis, 1967). After I explained this result to an audience of test constructors, someone responded that alpha might be a lower bound for my tests but that he did not believe the result held for his tests. My first explanation of the problem was that many researchers probably do not know what a theorem is, and that they interpret the result that alpha is a lower bound to the reliability as a matter of circumstance rather than fact—for one test it may be true but for another test it may not.

Singh (1997, pp. 21–27) offers a more interesting explanation. He posits that researchers used to working with empirical data are familiar with the concept of relative proof: They have learned that a sample of data cannot prove a hypothesis beyond reasonable doubt, only make it believable to a certain degree. Also, they know that research findings may be difficult to replicate, so that findings in one sample may not re-appear in another sample. On the other hand, mathematics uses the concept of absolute proof. Once it is proven that alpha is a lower bound to test-score reliability, there can be no doubt about the truth of this result. It is true for each test, and there are no exceptions. This way of thinking is at odds with empirical proof in which results support expectations to a certain degree, thus leaving room for doubt and exceptions. It seems to me that researchers confuse relative proof with absolute proof, thus taking the lower bound example as a statement that is often true but not always.

As an aside, I notice that the lower-bound theorem is true under the CTT model that assumes that measurement errors correlate zero. This is the approach most test constructors and researchers use to assess the psychometric quality of their tests. Variations on the classical model that allow correlated errors lead to a different reliability definition than the CTT proportion-of-true-score-variance definition, and allow alpha to exceed the alternative reliability coefficient (Green & Yang, 2009; Raykov, 2001). Of course, there is nothing wrong adopting a different measurement model that leads to different results for alpha. However, as I noticed elsewhere (Sijtsma, 2009b), if one wants to know the degree to which test scores are the same upon repetition of the measurement procedure, the CTT approach assuming zero-correlated error is correct.

2.2. Values of Alpha and Other Methods Obtained from a Sample Are Parameter Values

Discussing psychometric results obtained for a particular test in a sample of less than 100 observations, I remarked that the sample alpha of 0.82 should be tested for significance against the rule of thumb that required alpha must be at least 0.80 to be regarded sufficiently high (I refrain here from discussing what sufficiently high means exactly; the reader may notice that the practice of statistical data analysis is replete with similar rules of thumb). My remark was not considered

particularly helpful as 0.82 clearly was in excess of 0.80. Indeed, statistical tests of hypotheses concerning expectations about alpha or confidence intervals for alpha are rarely reported. An explanation for their absence might be the following. Coefficients like alpha and other reliability methods express the degree to which test-score variance is unaffected by random measurement error but values of alpha and other methods are also affected by sample size and hence liable to sampling error. This double stochasticity is a difficult phenomenon to understand. I assume this difficulty causes the disbelief that values of alpha cannot be taken at face value and hypotheses about alpha values should be tested or that confidence intervals might be highly useful, just as they are for any other statistical parameter.

2.3. Using a Higher Lower Bound than Alpha Means that One Tries to Cheat

After having explained to a colleague that there are better methods for estimating test-score reliability than alpha and that one should pick the one producing the highest value—for lower bounds this would be the greatest lower bound (GLB; Bentler & Woodward, 1980; Ten Berge, Snijders, & Zegers, 1981)—he replied without hesitation that this would amount to cheating. What he meant was that researchers should stick with the method they chose and not shop around. Using the same method creates consistency in methodology and comparability of different results and using different methods ruins this without apparent necessity. For reliability, the prevailing opinion seems to be that different methods establish different kinds of "reliability" that are impossible to compare across different studies.

Here, psychometrics is not without blame. Textbooks on test theory and other publications often suggest that parallel-test reliability, retest reliability, and coefficient alpha estimate different kinds of reliability, and that different lower bounds are based on different conceptions of error again producing different reliability assessments. I believe this is wrong. Instead, what each of these methods does is provide an approximation to the definition of reliability as the product-moment correlation between the test scores obtained in two independent administrations of the same test to the same group of examinees. Let the two test scores be denoted X and X', and reliability $\rho_{XX'}$. Parallel-test reliability coincides with this definition if the two tests are parallel to one another (Lord & Novick, 1968, p. 48); retest-reliability if the repetition of the same test at a later occasion may be conceived as a test parallel to the first administration; and lower bounds like alpha, lambda2 (Guttman, 1945), and GLB if the test parts on which the computations are based are essential tau-equivalent (Lord & Novick, 1968, p. 50). In real data, parallelism and essential tau-equivalence are too restrictive. As a result, each of the reliability methods estimates a different parameter, and all are different from reliability $\rho_{XX'}$.

My point is that each of the methods attempts to estimate the same test-score reliability $\rho_{XX'}$ but that the failure of the assumptions of the methods to be satisfied in the test data produces different reliability values. The literature on reliability regularly misinterprets the different methods as representing different kinds of reliability, and this clearly sets researchers on the wrong track.

For an imaginary test in an imaginary population (i.e., no data involved), Figure 1 shows true test-score reliability value $\rho_{XX'} = 0.85$, and true values for coefficients alpha and lambda2, and GLB, and for parallel-test reliability (denoted ρ_{parallel}) and retest-reliability (denoted ρ_{retest}). Had the test parts been essential tau-equivalent, different test versions been parallel, and the retest been parallel with respect to its first administration, all five values would have coincided with test-score reliability $\rho_{XX'} = 0.85$. Hence, the figure shows an instance in which the assumptions were not satisfied, thus producing different values for all five methods that are all unequal to 0.85. Then it is also clear that $alpha \leq lambda2 \leq GLB$, which is a necessary ordering both in the population and in the sample (Sijtsma, 2009a). Parallel-test reliability ρ_{parallel} and retest-reliability ρ_{retest} do not have a formal relation to alpha, lambda2, and GLB, and their values relative to the three lower bounds depend on the specific test and population.

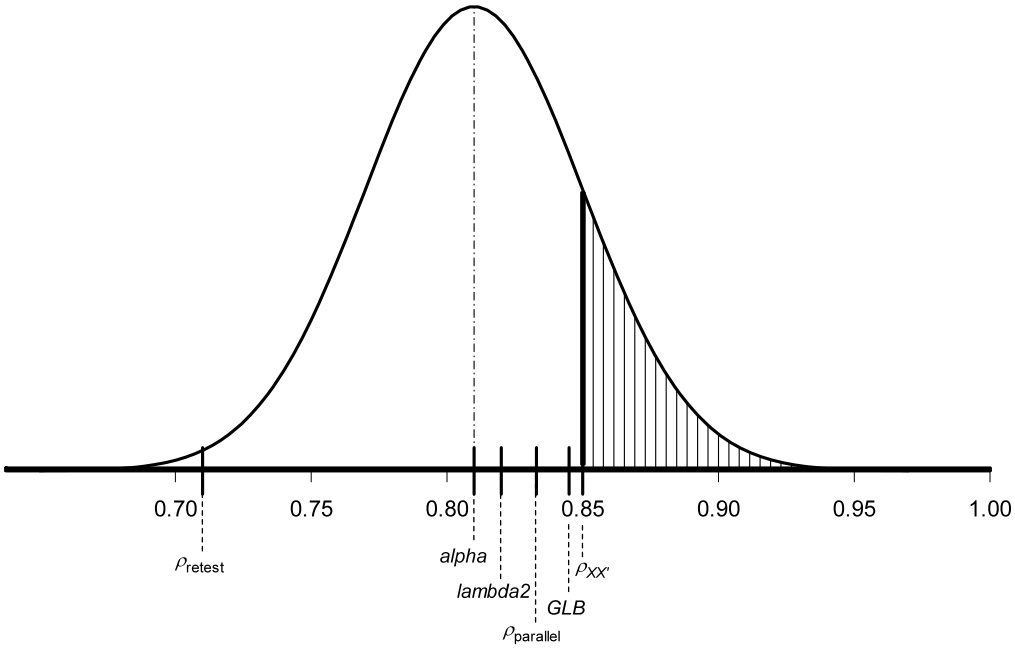


FIGURE 1.
Values of reliability methods when test forms are not parallel and test parts are not essential tau-equivalent, and sampling distribution of coefficient alpha.

Retest-reliability may even exceed test-score reliability when, for example, upon retesting, say, three months later, the examinees remembered many items from the test and the answers they gave, and the memory effect is the same for most examinees. This is a rare result, however, and the typical result the COTAN editor (Evers, personal communication) communicated to me is that on average retest-reliability is approximately 0.10 units lower than Cronbach's alpha (hence, a value equal to 0.71 was used in Figure 1). For example, a trait, such as depression, is unstable over time and the instability lowers the correlation between test and retest.

Figure 1 also shows an imaginary sampling distribution for alpha. It shows that in some samples alpha may overestimate true reliability $\rho_{XX'}$. A colleague (Markus, personal communication) pointed out to me that this result is difficult to understand for many people. Especially if one realizes that researchers tend to interpret sample alpha values as parameter values, the result that a sample estimate of a theoretical lower bound can be higher than the true reliability is confusing. The failure to distinguish sample and population is responsible for the confusion—sample values are interpreted as if they were population values, and the lower bound result for alpha that is true theoretically ($alpha \leq \rho_{XX'}$) may not stick in the sample in which due to sampling fluctuation statistic alpha may exceed parameter $\rho_{XX'}$.

2.4. Conclusion: Dissemination of Psychometric Knowledge Is Important

Researchers do not easily seem to pick up knowledge about the issues discussed here from psychometrics and, as a result, they may eventually abandon psychometrics and rely on the common lore of do's and don'ts that exist in their own research communities. It is of great importance to disseminate the right conceptual knowledge to researchers. Many textbooks nowadays concentrate on providing computational (e.g., SPSS, SAS) and numerical examples and explain at great length how to feed the right information to a computer program and how to interpret the output. There is nothing wrong with that as long as learning conceptual knowledge precedes it: What is

absolute proof in relation to well-known psychometric results? What is the difference between measurement error and sampling error? What is the difference between population and sample? How should different reliability methods be interpreted and used? I think psychometricians may show more interest in teaching their area to the people who are motivated to use it, the test constructors and researchers in general. The importance of the dissemination of knowledge through effective communication in a language with as little technical jargon as possible must not be underestimated.

On an optimistic note, I mention the sudden use in Dutch test construction of lambda2 (Hermans, 2011; Korkman, Kirk, & Kemp, 2010; Schlichting & Lutje Spelberg, 2010) and GLB (Kapinga, 2010) since the publication of the revision of the COTAN evaluation system in 2009 (Evers et al., 2010; the previous system did not mention lambda2 and GLB). Lambda2 is in SPSS but GLB is not, and GLBs inclusion might further stimulate its use in test construction and test revision.

3. Research Topics in Test Construction

The seven criteria COTAN uses for test assessment may each be scrutinized for topics that are candidates for improvement but not all criteria are equally interesting for psychometricians, in particular, *quality of the test materials* and *comprehensiveness of the manual*. *Theoretical basis of the test* refers to the question of whether the test construction was based on theory about the attribute of interest and *construct validity* refers to the research that was done to ascertain whether the test measures the attribute as intended. I will argue later that these two criteria represent different phases in the development of a valid instrument. The reason that COTAN distinguishes the two topics as separate assessment criteria is that one may be rated insufficient whereas the other may be rated good or sufficient. For example, in the absence of a theoretical basis (rating "insufficient") the test constructor may have done empirical research aimed at investigating ad hoc whether the test measures the intended attribute (rating "good" or "sufficient").

Prior to discussing *theoretical basis of the test* and *construct validity* under the unifying heading of validity, I discuss an issue in *reliability*, which is the tendency to use short to very short tests in individual assessment problems such as personnel selection and clinical change measurement. The assumption underlying the use of short tests is that a small number of high-quality items justifies making decisions about individuals, and can compensate for a much longer test that also contains items that are less well tailored for the particular decision problem. I argue that, in general, short tests may readily produce many more decision errors than longer test versions and that the use of short tests should be considered with care and caution.

The two topics I skip here are *norms* and *criterion validity*. Compared to educational test agencies, psychological researchers often have much smaller samples at their disposal, in particular for different age groups, educational-level groups, socio-economic status groups, and other subgroups that may be of interest to their measurement ambition. A recurring problem is that to have norm distributions of sufficient accuracy, one needs a sample size per subgroup that is unrealistic given financial or other constraints on test development research. A question then is whether data from different groups can be used to estimate norms in particular groups. This is known as continuous norming (Zachary & Gorsuch, 1985), of which different variations exist. For example, Zhu and Chen (2011) discuss inferential norming, which estimates different best fitting non-linear regression curves for the test-score means, standard deviations and possibly the skewness and the kurtosis on, say, age group in years. The regressions are used to interpolate means and other distribution characteristics between observed years, and the interpolations are used to estimate the test-score distributions, for example, for unobserved groups. Van Breukelen and Vlaeyen (2005) used multiple regression analysis for the test score on several interesting

background variables, checked for the assumptions of the regression model, and then interpreted the resulting standardized residuals as normed scores. The topic obviously is of great interest to test construction but little psychometric research has been done so far to study the method. Such research is needed but is beyond this article.

Criterion validity research results are often insufficient due to the vast effort it takes to collect the relevant kind of data in a prediction context that is often longitudinal. This is more a methodological obstacle than a psychometric problem, and for that reason I refrain from further discussing it, however important it is in practical test use. The next two subsections discuss the influence of test length on the decision quality in personnel selection and the assessment of individual change, and the validity issue of whether measurement is possible without a theory about the attribute of interest as a point of departure for test construction.

3.1. Reliability: Test Length, Decision Quality and Change Measurement

A trend in the construction and use of questionnaires for traits and attitudes but also in cognitive measurement is to use short tests, containing, say, 5 to 10 items. These items are chosen to capture the essence of the attribute and are located on the scale near a cut-score that divides the scale into, for example, a rejection region and a selection region (for a job, a therapy, a course). Their relief of the burden on the examinee, their increased efficiency and their lowered cost has motivated the use of short tests. The use of short tests in research has always been well accepted because research concentrates on group characteristics such as group means but does not require a particular level of accuracy for individual test scores. Short tests or questionnaires are also used for research purposes in marketing and sociology but in this case individual assessment is not an issue and the question of whether only a few items are enough to have sufficiently accurate individual measurement is immaterial. However, in psychological testing the focus usually is on individual measurement and then the question is justified whether a short test can provide enough accuracy, even if items are well chosen with respect to the application at hand.

Articles reporting construction of short tests often discuss test-score reliability but Mellenbergh (1996) correctly noted that the group characteristic of reliability is not very informative about the precision of individual measurement. Hence, reliability values of 0.8 or even 0.9 do not guarantee accurate individual measurement. A common reaction I experienced to the observation that compared to longer tests, shorter tests produce less accurate individual test scores and more decision errors based on those test scores, is that this is a result that has been known in psychometrics since Spearman (1910) and Brown (1910) introduced their prophecy formula for reliability as a function of test length. Hence, the observation is allegedly trivial. Be that as it may, then I ask why test constructors and test users prefer to use very short tests indeed, thus greatly increasing the risk of making decision errors in spite of the common knowledge about these risks. Thus, I assume it is useful to look further into the influence of deleting items from a test on decision quality in personnel selection and individual-change measurement.

3.1.1. Test Length and Decision Quality in Personnel Selection I assume that a test consists of K parallel test parts (Lord & Novick, 1968, pp. 47–50) indexed k. Test parts are individual items or tuples of n items; for example, a 20-item test may be thought of as consisting of five 4-tuples ($K = 5$) that are parallel without the individual items being parallel. Test scores are defined as sums of item scores and denoted X, test scores on parallel test parts are denoted X_k, true scores as T_k and random measurement errors as E_k. Then, given parallelism true-score variances are equal ($\sigma^2_{T_1} = \cdots = \sigma^2_{T_K}$) and so are the error variances ($\sigma^2_{E_1} = \cdots = \sigma^2_{E_K}$). For the whole test, true-score variance and error variance can be expressed in terms of part-test contributions: $\sigma^2_T = K^2\sigma^2_{T_k}$ and $\sigma^2_E = K\sigma^2_{E_k}$. For standard deviations, the true-score standard deviation equals $\sigma_T = K\sigma_{T_k}$ and σ_E, also known as the standard error of measurement (SEM), equals $\sigma_E = K^{1/2}\sigma_{E_k}$. Shortening a test consisting of K parallel parts until one test part remains, yields

TABLE 1.
Confidence intervals (CIs) for the true score as function of test length ($J = \#$ items).

J	X_c	$\hat{T}$	*Alpha*	*SME*	90% CI
5	12.0	10	0.70	1.58	[7.40; 12.60]
6	14.4	12	0.74	1.72	[9.16; 14.84]
7	16.8	14	0.76	1.88	[10.90; 17.10]
8	19.2	16	0.78	2.02	[12.68; 19.32]
9	21.6	18	0.80	2.14	[14.48; 21.52]
10	24.0	20	0.81	2.25	[16.29; 23.71]
11	26.4	22	0.82	2.36	[18.12; 25.88]
12	28.8	24	0.84	2.46	[19.95; 28.05]
13	31.2	26	0.85	2.56	[21.79; 30.12]
14	33.6	28	0.86	2.65	[23.64; 32.36]
15	36.0	30	0.87	2.74	[25.48; 34.52]
16	38.4	32	0.87	2.83	[27.34; 36.66]
17	40.8	34	0.88	2.91	[29.21; 38.79]
18	43.2	36	0.89	3.00	[31.07; 40.93]
19	45.6	38	0.89	3.08	[32.93; 43.07]
20	48.0	40	0.90	3.17	[34.78; 45.22]

shrinkage of true-score standard deviation by a factor K^{-1} and SEM by $K^{-1/2}$. Thus, the share of error variance in the test-score variance σ_X^2 increases because SEM shrinks slower and becomes more dominant. The use of short tests presupposes that this effect can be acceptable in terms of the numbers of correct and incorrect decisions based on the test scores. My question is to what degree this is true.

For non-parallel items, the impact of test shortening on measurement accuracy is still impressive (Sijtsma & Emons, 2011). For example, going upwards in Figure 2, the first scale is based on 20 five-point rating scale items consistent with the graded response model (Samejima, 1969) and scored 0–4, so that $0 \leq X \leq 80$, and $alpha = 0.90$; the second scale is based on 15 items ($alpha = 0.87$); the third scale on 10 items ($alpha = 0.81$); and the fourth scale on 5 items ($alpha = 0.70$). Table 1 shows results for all tests lengths between 20 and 5 items. Each test had a cut score X_c and items were maximally discriminating around the cut score. Each shorter test was a subset from the previous, longer test. For each scale, the cut score X_c was at 60% of the scale (Table 1), and a 90% confidence interval (CI) was based on an estimated true score located at 50% of the scale (Table 1).

Figure 2 shows that for the 20- and 15-item tests, cut score X_c is outside the 90% CIs, for the 10-item test it is just outside the CI, and for the 5-item test it is well inside the CI. More precisely, Table 1 shows that also for the 9-item test cut score X_c is just outside the CI but that for the 8-item test it is inside the CI. For 5 to 8 items, alpha values range from 0.70 to 0.78, and many researchers consider such values quite satisfactory for these short tests, implying that score differences are "reliable", but they may not realize that SEM and the corresponding CI tell a different story. Next, the question is how a shorter test affects the degree to which decisions based on test scores are affected.

Kruyen, Emons, and Sijtsma (in press) studied the effects of shortening tests in five different personnel selection scenarios. The authors studied a design using simulated test data, in which test length was an independent variable. I consider a scenario in which test score Y is the sum of the test scores on five subtests of equal length, with correlations between subtest true-scores equal to 0.2. The compensatory test score Y is used to select a fixed percentage of simulated applicants that have the highest test scores. Thus, the cut score for this problem is determined by the distribution of the test scores and may vary across different samples of equal size. Polytomous

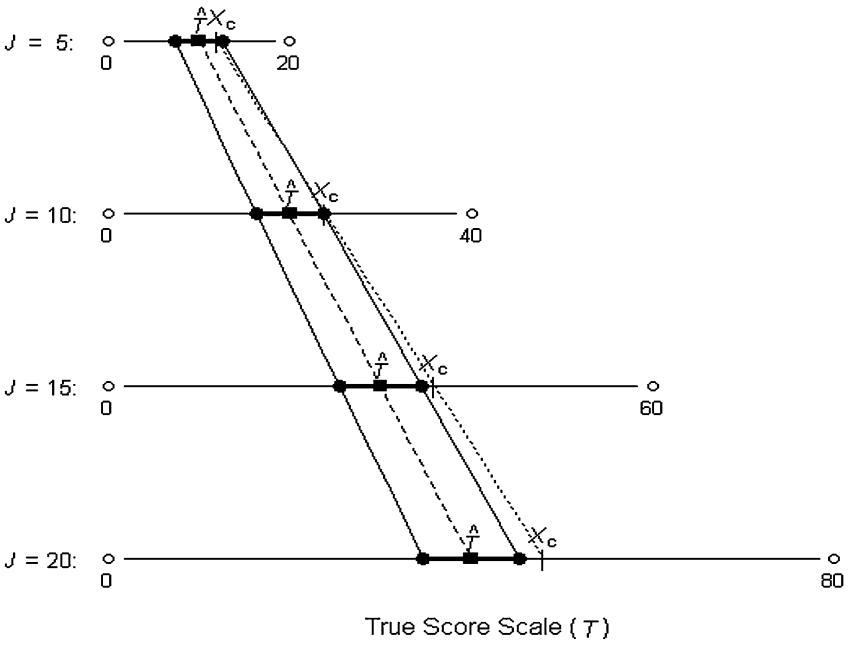


FIGURE 2.
90% confidence intervals for tests consisting of 5, 10, 15, and 20 items (J = # items); item scores 0, 1, 2, 3, 4; cut score at 60% of the scale and test score at 50% of the scale.

TABLE 2.
Quality of decisions for compensatory top-down selection based on five test scores; two selection ratios and five test lengths (J = # items).

J	*Alpha*	Sel. ratio 25%				Sel. ratio 10%			
		Spec.	Sens.	$CC-$	$CC+$	Spec.	Sens.	$CC-$	$CC+$
5	0.70	0.94	0.81	0.77	0.41	0.97	0.72	0.88	0.26
10	0.85	0.95	0.86	0.84	0.57	0.98	0.81	0.92	0.43
15	0.91	0.96	0.89	0.87	0.65	0.98	0.84	0.94	0.51
20	0.92	0.97	0.90	0.89	0.69	0.99	0.87	0.95	0.58
40	0.96	0.98	0.93	0.93	0.78	0.99	0.91	0.96	0.70

Note: Spec. = Specificity; Sens. = Sensitivity; $CC-$ = Classification consistency for true rejections; $CC+$ = Classification consistency for true selections.

items scored 0–4 were simulated using the graded response model (and dichotomous items using the 2-parameter logistic model), item discrimination was chosen such that particular coefficient-alpha values were realized for different test lengths, and item locations were close to the cut score; Kruyen et al. (in press) provide more technical details. Two useful situations are the following.

In one situation, Kruyen et al. (in press) considered the proportion of the group that is correctly rejected—both true score T_Y and test score Y are located to the left of the cut score—and the proportion that is correctly selected—both true score and test score are located to the right of the cut score; see Figure 3, left panel. Averaged across a large number of samples, these proportions yield the specificity and the sensitivity of the procedure. Table 2 (Kruyen et al., in press, did not discuss these results) shows that, as the selection ratio (i.e., the proportion of applicants eligible for selection) was lower, specificity was higher and sensitivity was lower. As tests became

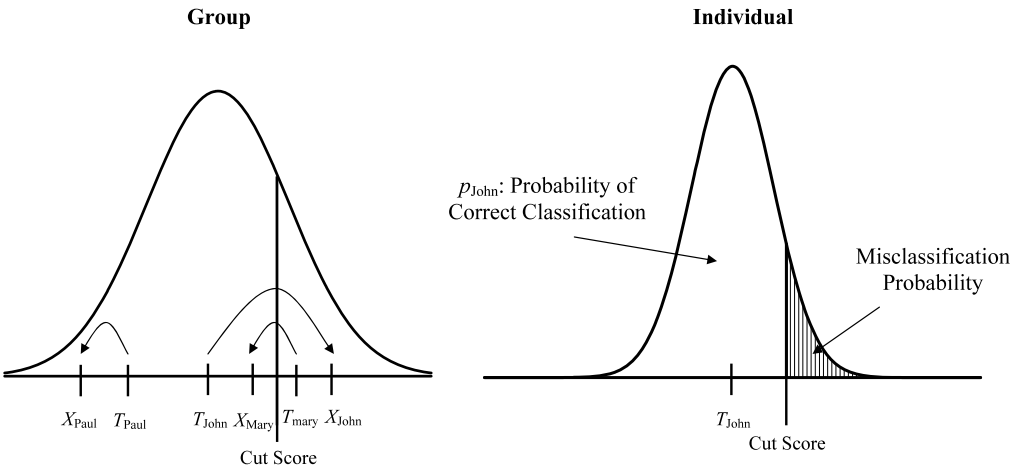


FIGURE 3.
Left-hand panel: true-score distribution in group of examinees; *right-hand panel*: propensity distribution of observable test scores for one individual.

shorter, specificity decreased more slowly than sensitivity, and specificity was only little affected whereas sensitivity was more affected, 0.72 being the lowest sensitivity value.

In another situation, Kruyen et al. (in press) addressed the question that is pervasive in statistics: What would happen if we did it all over again? That is, if we repeat the test procedure several times, to what degree do we classify an individual for whom $T_Y < X_c$ to the left of the cut score on the basis of his test score Y? A similar question was addressed for individuals located to the right of the cut score. For rejection, Figure 3 (right panel) shows the corresponding probability of correct classification, denoted p, under an individual's propensity distribution (Lord & Novick, 1968, pp. 29–30). Emons, Sijtsma, and Meijer (2007) proposed to set a minimum to p, which they called the certainty level π; this is the minimum proportion of correct decisions about an individual upon repetition of the test procedure that is acceptable to the organization responsible for the selection procedure. For example, choosing $\pi = 0.9$ expresses the organization's policy to maintain high standards for decisions about individuals. If it were possible to observe p values, we would require that $p \geq \pi$ were decisions to be made, and for individuals for whom $p < \pi$, we would conclude that not enough information was available for making a decision.

Although propensity distributions are unobservable, I concur that what one really would like to know is to what degree repetition of the measurement procedure would classify, say, John, correctly with enough certainty. This is of interest, for example, when it must be decided whether John should receive treatment and one does not want to withhold the treatment when he really needs it nor give the treatment when he does not need it. The individual is central here. Table 2 (again, Kruyen et al., in press, did not discuss these results) shows for rejections and selections the proportions of individuals for whom $p \geq 0.9$ ($\pi = 0.9$); these proportions are called classification consistency, denoted $CC-$ (rejected group based on true scores) and $CC+$ (selected group). As selection ratio was lower, $CC-$ was higher and $CC+$ was lower. As tests were shorter, $CC-$ remained quite high, but $CC+$ was low; $CC+ < 0.5$ for $J = 5$ (selection ratio 0.25) and for $J = 10$ and $J = 5$ (selection ratio 0.10). The proportions in Table 2 can be upgraded when a lower certainly level is chosen, but then one accepts lower quality standards. The conclusion is that test shortening has a profound negative effect on decision quality.

3.1.2. Test Length and Individual-Change Assessment The same individual's test scores obtained before and after therapy may be compared so as to decide whether the therapy effect

was statistically significant. Individual change is the difference between a patient's test score after treatment, X_2, and his/her test score before treatment, X_1. Difference scores like $X_2 - X_1$ have sometimes been dismissed (Cronbach & Furby, 1970) due to their alleged low reliability, suggesting that difference scores are mainly due to measurement error. However, reliability is a group characteristic and low reliability leaves open the possibility that individual change is accurate and hence statistically significant (Mellenbergh, 1999).

The RC index (Jacobson & Truax, 1991) is the most popular statistical method for testing change. It uses the standard error of measurement for difference scores, SEM_{diff}, which is the standard deviation of the measurement error of difference scores. This quantity is derived from the SEM for one test score, $SEM = \sigma_X\sqrt{1-\rho_{XX'}}$. Assuming normal propensity distributions (hence, normally distributed measurement error) and zero correlation between measurement errors at different measurement occasions, one can derive that $SEM_{\text{diff}} = \sqrt{2}SEM$. The RC index is defined as

$$RC = (X_2 - X_1)/SEM_{\text{diff}}.$$

RC is used to test the null hypothesis of no change against the alternatives whether observed change is in the expected direction of recovery ($X_2 - X_1 > 0$) (significant if $RC > 1.645$, 5% level) or whether observed change is in any direction, recovery or deterioration ($X_2 - X_1 \neq 0$) (significant if $|RC| > 1.96$).

Statistical significance is a necessary but not a sufficient condition for clinical significance (Jacobson & Truax, 1991). A patient whose change is clinically significant can be declared cured and no longer in need of further therapy. A patient thus may show change that is statistically significant but too small to also be clinically significant (Ogles, Lunnen, & Bonesteel, 2001). Clinically significant change requires the individual to cross a cut score X_c that can be determined in different ways depending on the overlap of the test-score distribution in the dysfunctional group that is in need of therapy and the test-score distribution in the functional group that has recovered. Jacobson and Truax (1991, p. 16) suggested statistically testing whether a post-treatment score X_2 is significantly greater than cut score X_c (assuming that higher test scores reflect the desired direction of change). Obviously, if one requires an individual having a significant RC value to also show clinical significance, assuming that X_c is located in the rejection region for the test of significant statistical change, this reduces the power of the whole procedure compared to only requiring RC to be statistically significant; also see Jacobson and Truax (1991) for a discussion of this issue.

The RC index has been criticized because it uses the wrong S_E estimator, ignores the regression-to-the-mean effect, and uses only two test scores rather than multiwave data. Bauer, Lambert, and Nielsen (2004) compared five statistics for individual-change assessment that address the latter two problems and argued in favor the Jacobson–Truax method because it is easy to compute, clinical cutoff scores are available for many widely known instruments, and the method moderates some of the extreme effects of other methods. Atkins, Bedics, McClinchey, and Beauchaine (2005) drew a similar conclusion.

Sijtsma and Emons (2011) studied the effect of shortening a test on the power of the RC test, using the same setup of graded-response model data that lay at the basis of Table 1 and Figure 2. They simulated effect sizes that were small (0.25 SD), medium (0.5 SD), large (1 SD) and very large (2 SD). For a one-tailed 5% significance level, thus testing for improvement, Sijtsma and Emons (2011, Table 4) found that power was only substantial, say, in excess of 0.70, for all test lengths (5 to 20 items) if effect size was large or very large. For medium effect size, short tests generated too little power; that is, for 10 items power was 0.58 and for 5 items power was 0.39. For small effect size, all test lengths failed to generate enough power, which ranged from 0.19 for 5 items to 0.37 for 20 items. For two-tailed tests, power was smaller and only in excess of 0.60 for all test lengths if effect size was large or very large. The authors concluded that only for

tests consisting of more than 10 items is the *RC* test powerful enough to detect at least medium change.

Reise and Haviland (2005) suggested using the standard error for the latent person variable from IRT models, denoted $SE(\theta)$, for individual-change assessment, as it is individual-oriented using the test information function rather than the group-based reliability coefficient; also, see Mellenbergh (1999). However, test length is also an important issue for IRT-based change assessment that is scale-dependent, and also in this context I expect that short tests generate too little power to detect less than very large change.

3.1.3. Conclusion: Use Long Tests for Individual Decisions Short tests may have quite high reliabilities but their SEMs (and presumably $SE(\theta)$s) may tell a different story. Although additional research remains to be done, I recommend using relatively long tests in individual decision-making. Here, I used SEM because it is consistent with the predominant use of CTT in test construction and test use but the scale-dependent standard error $SE(\theta)$ may be a strong alternative to SEM.

3.2. Validity: Measurement Without Theory?

3.2.1. What is Valid: The Test or the Test Score? The literature on validity has long revolved around the issue of whether the test or the test score is valid (Lissitz, 2009), and critics (e.g., Borsboom, Mellenbergh, & Van Heerden, 2004; Michell, 1999; Sijtsma, 2011) have posited that too much attention went to validation of the test score (e.g., see *Standards for Educational and Psychological Testing*; AERA, APA, & NCME, 1999, p. 9) and too little attention to the question what the test measures. Why did the question what the test measures move to the background in favor of what one can do with the test score?

Michell (1999) noticed that already in the first half of the twentieth century psychologists embraced operationism as their philosophy of measurement—the attribute is what the test measures, after Boring's famous quote that intelligence is what the test measures (Boring, 1923)—and that later Stevens' (1946) definition of measurement as the assignment of numbers to objects according to rules without the necessity to empirically justify this assignment made things only worse. Thus, psychologists felt little pressure to engage in the thorough development of theories of attributes to a degree that the operationalization of the attribute into a set of prescriptions for instrument construction is possible. As a result, psychological measurement often does not rest on well-founded theory but instead on results from ad hoc theorizing about the attribute or from rather abstract, vague and incomplete theories that are impossible to test empirically due to lack of detail. In defense of psychological measurement I notice that there are many reasons why the establishment of exact psychological theories of attributes is difficult. Examples are the impossibility to effectively manipulate or control cognitive processes in experiments, and the subjects' reflexivity to the situation in which they find themselves, such as an experiment, which allows them to (consciously or unconsciously) influence the outcomes of the research beyond the control and the intentions of the experimenter.

In the absence of well-established theories, items used for constructing a measurement instrument often coincide with the attribute one wants to measure, thus defining the attribute to a large degree. Validation then rests on factor analyzing the correlation matrix of the items and correlating the test score with other variables such as tests believed to measure the same or nearly the same attribute—an approximation of convergent validation—or attributes that are related to the attribute of interest but not the same—divergent validation. The evidence of what a test measures thus is circumstantial and circles around directly addressing cognitive, affective and other processes.

Borsboom et al. (2004) and Borsboom, Cramer, Kievit, Zand Scholten, and Franić (2009) suggested that validity research must focus on the question what the test measures, and argued

that the development of theory for the attribute is necessary. Studying what subjects do when they respond to the items best contributes to developing attribute theory. The question is which cognitive processes the item activates before the subject responds. To this end, the researcher can manipulate formal characteristics of the items thus obtaining more insight in the processes. These processes are inaccessible to direct observation but IRT models and other latent variable models may be used to study the processes indirectly (e.g., De Boeck & Wilson, 2004; Bouwmeester, Vermunt, & Sijtsma, 2007). The IRT models hypothesize a formal structure for a cognitive process and their fit to the data can be tested, leading to the support or the rejection of the hypothesis. In the latter case, model misfit diagnostics provide suggestions on how to improve the model. Latent class models are flexible tools that allow a subdivision of the investigated group into subgroups that use different cognitive processes and a formalization of the process per group.

3.2.2. Theory, Tests, Data, and Psychometrics I concentrate on the role of psychometrics in test construction and validation, first in the absence and then the presence of a well-established theory of the attribute. Let us assume a theory that would otherwise guide the operationalization of the attribute into a measurement instrument is absent. Then the researcher can do nothing other than assemble a set of items on the basis of available but incomplete theoretical notions, habit, tradition, and intuition. The set comprises the experimental test version. The data collected by means of this test are analyzed, for example, using IRT models, to establish whether a scale can be constructed, and structural equation modeling to test ideas about relationships of test scores with other variables. In a nutshell, this is very much how modern psychometrics can be used to develop scales and validate them by establishing relationships in a network of relevant variables.

It is important to realize that in the absence of a guiding theory about the attribute, IRT models take over the substantive theory's role and become the criterion for what the scale will be. What does it mean that an IRT model takes over from substantive theory? Does IRT provide formalized models of the structure of an attribute? The answer is negative. IRT models define measurement properties for scales but nothing else. In this sense, the IRT model is prescriptive of what measurement should be but it does not tell us what, for example, Type D (D = distressed) personality (Denollet, 2000) or proportional reasoning (Siegler, 1981) is. These and other attributes are empirical entities, and theory should describe the structure of Type D personality or the processes driving proportional reasoning. Empirical research should be done to seek support for the theories or to obtain suggestions of how to modify the theories to improve their explanatory power. I contend there is no compelling reason why the theoretical structure of an attribute coincides with the mathematical structure of an IRT model (or another psychometric model).

Despite the absence of a compelling congruence between attribute theory and IRT model structure, IRT has been quite successful in scale construction. The reason is that researchers assemble tests by preselecting items that cover particular facets of an attitude logically derived in a facet design, or items that are variations on a particular principle, for example, as in a set of tasks used for the mental rotation of objects to determine whether they are equal to one of four given choice options. The point is that, without a guiding attribute theory, researchers assemble items that are homogeneous in the broad sense of "belonging" together, and this creates data showing much structure. IRT models and factor models easily pick up the common denominator in the data, which then may be the basis of a scale. However, whether this leads to theory formation on an ad hoc basis is doubtful. Theories do not rise up from the data; they have to be designed before data collection and then tested.

I discuss the Type D Scale-14 (DS14; Denollet, 2005), a 14-item questionnaire for Type D personality that is based on a host of empirical research. Type D personality refers to the joint tendency toward negative affectivity (NA; 7 items) and social inhibition (SI; 7 items), and is related to poor cardiac prognosis. Subjects are scored on continuous scales for NA and SI, which

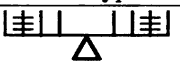
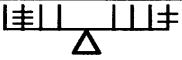
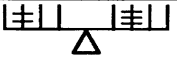
Item types	Strategies S1	S2	S3	S4	S5
 'Balance'	balance (100)	balance (100)	balance (100)	balance (100)	balance (100)
 'Weight'	left down (100)	left down (100)	left down (100)	left down (100)	left down (100)
 'Distance'	balance (0)	left down (100)	left down (100)	left down (100)	left down (100)
 'Conflict weight'	left down (100)	left down (100)	guess (33)	balance (0)	left down (100)
 'Conflict distance'	right down (0)	right down (0)	guess (33)	balance (0)	left down (100)
 'Conflict balance'	right down (0)	right down (0)	guess (33)	balance (100)	balance (100)

FIGURE 4.
Six balance-scale item types (*rows*) and five solution strategies (*columns*); in cells, predicted outcome and predicted percentage of correct answers (between parentheses).

refer to continuous traits, and classified as Type D using cut scores $X_{NA} \geq 10$ and $X_{SI} \geq 10$, where Type D refers to a discrete personality configuration (Denollet, 2000, p. 258) supported by cluster analysis results (ibid., p. 257).

As the discreteness of Type D is not based on well-founded theory but derived from empirical results and the need for practical diagnosis, Ferguson et al. (2009) asked whether Type D personality is continuous or categorical and used taxometric analysis (Ruscio, Haslam, & Ruscio, 2006) of DS14 data to arrive at the conclusion that Type D likely is continuous. Emons, Denollet, Sijtsma, and Pedersen (2011) used mixture IRT models and drew the same conclusion. In the absence of a theoretical foundation, however, both Ferguson et al. (2009) and Emons et al. (2011) had no other choice than to let the data speak, but Emons et al. (2011) also noted that data analysis without theoretical guidance leaves a wide margin for uncertainty.

What if a well-developed theory about an attribute stood at the basis of test construction? I discuss the cognitive ability of proportional reasoning; this the ability of identifying the relevant task dimensions and understanding the multiplicative relation between these dimensions (Jansen & Van der Maas, 1997, 2002; Siegler, 1981; Van Maanen, Been, & Sijtsma, 1989). Proportional reasoning is measured using balance-scale problems (Figure 4), and children follow one of several rules or strategies for solving tasks, resulting in the response that the scale tips to the left or the right or is in equilibrium. The different rules stand for different developmental phases. The tasks used in the test have properties that give way to the use of different rules, and children using the same rule are expected to produce the same pattern of incorrect-correct scores on the set of tasks.

Van Maanen et al. (1989) set up a 25-item test with five 5-tuples representing different classes of problems, together eliciting the different solution rules (Figure 4; the items in the first row, representing equilibrium, were too easy and were left out of the test). As could be anticipated, Van Maanen et al. (1989) found that the Rasch model did not fit the test in the complete group, then used cluster analysis on five items, one from each problem type, to identify the hypothesized four different rule groups, and finally fitted a linear logistic test model (Fischer, 1995) representing the cognitive operations involved in the use of a particular rule. The linear

logistic test models fit reasonably well, although the small sample sizes reduced the power of the chi-squared model test. In rule groups, standard deviations and coefficient alpha's of total scores were low, thus acknowledging that each group was homogeneous with respect to item-score pattern and, hence, the total score. Jansen and Van der Maas (1997, 2002) used latent class analysis for the same data and newly collected data sets and were able to partly confirm and also partly modify Siegler's predictions about rule use, and found children to be consistent users of one rule.

Type D personality, although well documented, lacks the theoretical foundation that predicts dimensionality or discreteness, and data analysis soon runs in circles, as no decisive evidence for either hypothesis is available from the data. The theory of proportional reasoning is detailed and well established enough to predict rule classes that indeed are supported by empirical research. Children can be classified as particular-rule users based on estimated probabilities of belonging to a rule-class, and class membership provides useful information about cognitive development. Proportional reasoning shows how a well-founded theory can be the basis of test construction, leading to a set of classes that can be ordered on the basis of developmental theory instead of a metric scale on which children are located using a total score or a latent variable.

3.2.3. Conclusion: Measurement Must Be Based on Theory As long as tests are not based on well-developed substantive attribute theory, validity will remain a problem and it will remain appealing to ask what the test can do—no doubt a sensible question—rather than what it measures—a prerequisite for asking sensible questions about utility. Good theories are a necessary condition to construct tests that measure a particular attribute and that serve particular practical purposes. This is where the COTAN assessment criteria of *theoretical basis of the test* and *construct validity* and even *criterion validity* meet one another. This is also a research area where psychometrics can provide a helping hand to psychology. Good examples of research mainly done by psychometricians involve studying solution rules in Raven's Progressive Matrices test (Verguts & De Boeck, 2002), testing theory for transitive reasoning (Bouwmeester et al., 2007), and studying the process structure of the emotion of guilt (Smits & De Boeck, 2003).

4. Discussion

The practice of psychological test construction and test use struggles with many didactical misunderstandings and relevant research problems that deserve attention from psychometricians. My membership in COTAN made me more aware of these needs and inspired this presidential address. My personal research program for the next few years concentrates on reliability issues including the consequences of using short tests for decision making, the accurate measurement of change, the influence of reliability on the power of statistical tests (not discussed here, but see Sijtsma & Emons, 2011; also Nicewander & Price, 1983), the promotion of theory development as the basis of test construction, and additional topics such as continuous norming (Zhu & Chen, 2011).

Acknowledgements

I am thankful to Wilco Emons, Arne Evers, Keith Markus, Roger Millsap, Jos ten Berge, and Carol Woods for their suggestions and their comments on a previous draft of this article. Of course, I am responsible for the views expressed in this article and its contents.

References

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

Atkins, D.C., Bedics, J.D., McGlinchey, J.B., & Beauchaine, T.P. (2005). Assessing clinical significance: does it matter which method we use? *Journal of Consulting and Clinical Psychology*, *73*, 982–989.

Bauer, S., Lambert, M.J., & Nielsen, S.L. (2004). Clinical significance methods: a comparison of statistical techniques. *Journal of Personality Assessment*, *82*, 60–70.

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.

Boring, E.G. (1923). Intelligence as the tests test it. *New Republic*, *35*, 35–37.

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Zand Scholten, A., & Franić, S. (2009). The end of construct validity. In R.W. Lissitz (Ed.), *The concept of validity. Revisions, new directions, and applications* (pp. 135–170). Charlotte: Information Age Publishing, Inc.

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

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.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

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

Cronbach, L.J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin*, *74*, 68–80.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models. A generalized linear and nonlinear approach*. New York: Springer.

Denollet, J. (2000). Type D personality: a potential risk facor refined. *Journal of Psychosomatic Research*, *49*, 255–266.

Denollet, J. (2005). DS14: standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine*, *67*, 89–97.

Emons, W.H.M., Denollet, J., Sijtsma, K., & Pedersen, S.S. (2011). *Dimensional and categorical approaches to the Type D personality construct* (in preparation).

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

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R.R. (2010). The Dutch review process for evaluating the quality of psychological tests: history, procedure and results. *International Journal of Testing*, *10*, 295–317.

Ferguson, E., et al. (2009). A taxometric analysis of Type D personality. *Psychosomatic Medicine*, *71*, 981–986.

Fischer, G.H. (1995). The linear logistic test model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications* (pp. 131–155). New York: Springer.

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

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

Hermans, H.J.M. (2011). *Prestatie Motivatie Test voor Kinderen 2 (PMT-K-2) (Performance motivation test for children 2)*. Amsterdam: Pearson Assessment.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

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.

Jansen, B.R.J., & Van der Maas, H.L.J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology*, *81*, 383–416.

Kapinga, T.J. (2010). *Drempelonderzoek. Didactische plaatsbepaling binnen het voortgezet onderwijs en praktijkonderwijs. 5^{e} versie 2010 (Threshold investigation. Didactical location within secondary education and practical education. 5th Version 2010)*. Ridderkerk: 678 Onderwijs Advisering.

Korkman, M., Kirk, U., & Kemp, S. (2010). *NEPSY-II-NL. Nederlandstalige bewerking (A developmental neuropsycological assessment, II, Dutch version)*. Amsterdam: Pearson Assessment.

Kruyen, P.M., Emons, W.H.M., & Sijtsma, K. (in press). Test length and decision quality in personnel selection: when is short too short? *International Journal of Testing*.

Lissitz, R.W. (2009). *The concept of validity. Revisions, new directions, and applications*. Charlotte: Information Age Publishing, Inc.

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

Mellenbergh, G.J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299.

Mellenbergh, G.J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, *23*, 87–89.

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

Nicewander, W.A., & Price, J.M. (1983). Reliability of measurement and the power of statistical tests: some new results. *Psychological Bulletin*, *94*, 524–533.

Novick, M.R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.

Ogles, B.M., Lunnen, K.M., & Bonesteel, K. (2001). Clinical significance: history, application, and current practice. *Clinical Psychology Review*, *21*, 421–446.
Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.
Reise, S.P., & Haviland, M.G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*, 228–238.
Ruscio, J., Haslam, N., & Ruscio, A.M. (2006). *Introduction to the taxometric method: a practical guide*. Mahwah: Erlbaum.
Samejima, F. (1969). *Psychometrika monograph: Vol. 17. Estimation of latent ability using a response pattern of graded scores*. Richmond: Psychometric Society.
Schlichting, L., & Lutje Spelberg, H. (2010). *Schlichting Test voor Taalproductie—II (Schlichting test for language production—II)*. Houten: Bohn Stafleu van Loghum.
Siegler, R.S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2, Serial No. 189).
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.
Sijtsma, K. (2011). Psychological measurement between physics and statistics. *Theory & Psychology*.
Sijtsma, K., & Emons, W.H.M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*, 565–572.
Singh, S. (1997). *Fermat's last theorem*. London: Harper Perennial.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.
Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.
Smits, D.J.M., & De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research*, *38*, 161–188.
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.
Van Breukelen, G.J.P., & Vlaeyen, J.W.S. (2005). Norming clinical questionnaires with multiple regression: the pain cognition list. *Psychological Assessment*, *17*, 336–344.
Van Maanen, L., Been, P.H., & Sijtsma, K. (1989). Problem solving strategies and the linear logistic test model. In E.E.C.I. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–287). New York: Springer.
Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's progressive matrices test. *European Journal of Cognitive Psychology*, *14*, 521–547.
Zachary, R.A., & Gorsuch, R.L. (1985). Continuous norming: implications for the WAIS-R. *Journal of Clinical Psychology*, *41*, 86–94.
Zhu, J., & Chen, H.-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment*. doi:10.1177/0734282910396323.

Published Online Date: 3 DEC 2011