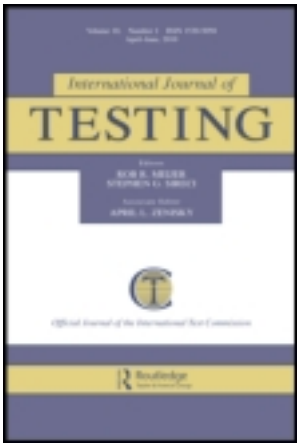

International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hijt20

On the Shortcomings of Shortened Tests: A Literature Review

Peter M. Kruyen [a], Wilco H. M. Emons [b] & Klaas Sijtsma [b]

[a] Institute for Management Research, Radboud University Nijmegen, The Netherlands

[b] Department of Methodology and Statistics, Tilburg University, The Netherlands

Published online: 31 May 2013.

To cite this article: Peter M. Kruyen, Wilco H. M. Emons & Klaas Sijtsma (2013) On the Shortcomings of Shortened Tests: A Literature Review, International Journal of Testing, 13:3, 223-248, DOI: 10.1080/15305058.2012.703734

To link to this article: http://dx.doi.org/10.1080/15305058.2012.703734

International Journal of Testing, 13: 223–248, 2013

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2012.703734

Routledge
Taylor & Francis Group

On the Shortcomings of Shortened Tests: A Literature Review

Peter M. Kruyen
Institute for Management Research, Radboud University Nijmegen, The Netherlands

Wilco H. M. Emons and Klaas Sijtsma
Department of Methodology and Statistics, Tilburg University, The Netherlands

To efficiently assess multiple psychological constructs and to minimize the burden on respondents, psychologists increasingly use shortened versions of existing tests. However, compared to the longer test, a shorter test version may have a substantial impact on the reliability and the validity of the test scores in psychological research and individual decision making. In this study, we reviewed the psychological literature for recent trends in the use of short tests and examined in depth how and to what extent test constructors and test users addressed the impact on reliability and validity, other potential consequences of using short tests. The sample consisted of shortened tests found in six peer-reviewed psychological journals in the period 2005–2010. Based on our review, we provided recommendations for psychologists considering test shortening.

Keywords: literature review, psychological tests, reliability of short tests, test shortening, validity of short tests

Psychological tests and questionnaires are widely used in psychological research and individual decision making in areas such as clinical, health, and medical psychology and personnel selection. To meet practical limitations on available time and resources, psychologists increasingly resort to shortened tests to increase efficiency of testing (e.g., Burisch, 1997; Shrout & Yager, 1989; Stanton, Sinar,

This research was supported by a grant from the Netherlands Organization of Scientific Research (NWO 400-05-179) (first author).

The authors would like to thank Stéfanie André for her research support.

This article was accepted under the previous co-editor team, Rob R. Meijer and Stephen G. Sireci.

Correspondence should be sent to Peter M. Kruyen, Radboud University Nijmegen, Institute for Management Research, P. O. Box 9102, 6500 HC Nijmegen, The Netherlands. E-mail: p.m.kruyen@fm.ru.nl

Balzer, & Smith, 2002). Examples include a 64-item and a 32-item version of the Inventory of Interpersonal Problems (Alden, Wiggins, & Pincus, 1990; Soldz, Budman, Demby, & Merry, 1995), the Mini-Markers, which is a 40-item version of the Unipolar Big-Five Markers (Saucier, 1994), and the 53-item and 18-item Brief Symptom Inventories, both derived from the Symptom Checklist-90-Revised (Derogatis, 2001; Derogatis & Melisaratos, 1983). Short forms exist even for tests that already included only a limited number of items, like the 13-item Beck Depression Inventory items (originally 21 items, Beck & Beck, 1972), the 10-item Marlowe-Crowne Social Desirability Scale (originally 33 items, Strahan & Gerbasi, 1972), and the 21-item Depression Anxiety Stress Scales (originally 42 items, Lovibond & Lovibond, 1995).

Shortened test may be more efficient from a practical viewpoint, but their use goes against the old psychometric wisdom that many items are needed for reliable and valid measurement (e.g., Anastasi, 1982, p. 192; Nunnally, 1978, p. 243). This is the reason why the original psychological tests consisted of large numbers of items, often 40 items or more. Moreover, these long tests were carefully constructed such that every item contributed to test-score reliability, and the items provided sufficient construct coverage. Omitting items from a psychometrically sound test almost inevitably results in lower test-score reliability and poorer construct coverage.

This loss in measurement quality may create bias in the estimated group means and correlations (Nicewander & Price, 1983; Sijtsma & Emons, 2011) and impair individual decision making (Emons, Sijtsma, & Meijer, 2007; Kruyen, Emons, & Sijtsma, 2012). The question is whether the current practice of leaving out items of good psychometric quality with the purpose to meet efficiency requirements lowers measurement quality to an unacceptable extent (e.g., Allen & Yen, 1979, p. 87; Lord & Novick, 1968, pp. 112, 114; Smith, McCarthy, & Anderson, 2000).

The purpose of this literature study was to explore current practices in test shortening and assess to what extent test shortening affects measurement quality. We reviewed the literature on test shortening published between 2005 and 2010 in six leading psychological journals that publish articles about psychological tests. This article is organized as follows. First, we present the research questions. Second, we discuss test-shortening strategies and define technical terms, including reliability, measurement precision, and validity. Third, we discuss the results from the literature study. Fourth, we provide recommendations with respect to the use of shortened tests.

RESEARCH QUESTIONS

The next three research questions were investigated on the basis of a literature search:

1. What is the current practice of using shortened tests? This includes: How often do researchers use shortened tests? What motivates researchers to use

shortened tests? How much shorter are shortened tests compared to the original, longer versions? How are shortened tests constructed?
2. What are the differences between reliability, construct-related validity, and prediction-related validity of the shortened test and the longer test?
3. To which extent do researchers pay attention to the potential advantages and potentially negative implications of using shorter tests for their own research?

TECHNICAL TERMS

Test-Shortening Strategies

To shorten tests, test constructors use statistics-driven strategies, a judgmental strategy, an ad hoc strategy, or a combination of these strategies (Coste, Guillemin, Pouchot, & Fermanian, 1997; Stanton et al., 2002, p. 44). Following a statistics-driven strategy, items are removed based on statistical criteria. A widely used statistical strategy is to produce a shorter test that has a test-score reliability that is close to the reliability of the longer test. Classical test theory (CTT) approaches maintain reliability at approximately the same level as in the longer test. This is achieved by keeping the items in the shortened version that have the highest correlations with the test score on the longer version (including the item under consideration) or the corrected test score (excluding the item). Factor analytic approaches select items with the highest factor loadings. Item-response-theory methods select the items that have larger item-information functions than other items. The first items contribute most to the test-information function, thus producing an estimated latent variable that enables the greatest measurement precision along the scale. Using these statistical strategies, the best items are selected (or, equivalently, the worst items are removed) even if their quality is not very good but there are no other items from which one can choose.

Test constructors also use statistics-driven strategies to maintain test validity at the same level as in the longer test version. For example, tests constructors select items having the highest correlation with another test measuring the same construct or with an interesting external criterion. Factor analysis may be used to remove items that have cross-loadings or that are involved in correlated errors as they may measure collateral constructs not deemed relevant for the test. Another possibility is to compare the specificity and the sensitivity of the short test with those of the longer test. Test shortening is considered successful when correlations with the other test are substantial or the number of classification errors is not much larger for the short test than the longer test.

The judgmental strategy amounts to selecting items on the basis of expert judgment of the contents of the items. Experts decide which items best cover the construct of interest. The judgment may also include a decision on the relevance of particular items for construct measurement in particular subgroups or an

assessment of the appropriateness of the contents of items, for example, with respect to language use, or both. Examples include the work of Lawing, Frick, and Cruise (2010) who shortened the Impulsive/Antisocial Behavior scale of the Juvenile Sex Offender Assessment Protocol-II by excluding two items they considered irrelevant for the target population; and Scheier and Carver (1992) who shortened the Ironson–Woods Spirituality/Religiousness Index by selecting items that they judged the most important based on respondent interviews.

The ad hoc strategy has many appearances that have in common that neither statistics nor content play a role in selecting or removing items. Examples are the selection of the uneven numbered items or retaining the first ten items for the shortened test. Another possibility is to select items based on their format without considering their content, such as when the researchers aim at maintaining a balanced number of positively and negatively worded items in the short test.

Several authors recommend to combine statistics-driven strategies and a judgmental approach to shortening the test (e.g., Coste et al., 1997; Smith et al., 2000). First, the definition of the construct needs to be such that the most important items for assessing the construct can be identified. Using this definition, multiple experts should assess the validity of each item and statistical methods should assess experts' degree of agreement (see also American Educational Research Association, 1999, p. 19). Second, test constructors should decide which items to include in the shortened test based on the judgment of these experts and additional statistical evidence with respect to the contribution of every item to the reliability and the validity of the test. An example of such a combined strategy is the work of Paap and colleagues (2011) who used item response theory to derive shortened scales from the SCL-90-R, where each scale was built on a kernel of two items that content experts identified as best reflecting the attribute of interest.

Reliability and Measurement Precision

As CTT still is the dominant approach to test construction, we explain reliability and measurement precision from this perspective. Reliability is defined as follows. Let X_+ be the test score on a test containing J items. CTT assumes that test score X_+ is the sum of a true score T and a random measurement error E, such that $X_+ = T + E$. Test-score reliability is defined as the proportion of true-score variance in the group relative to the test-score variance, such that $\rho_{XX'} = \sigma_T^2/\sigma_{X_+}^2$. Estimates of test-score reliability include test-retest reliability, split-half reliability, coefficient alpha (Nunnally & Bernstein, 1994, pp. 251–255), and the greatest lower bound to the reliability (Bentler & Woodward, 1980; Sijtsma, 2009b; Ten Berge & Sočan, 2004).

Researchers assume that a test score that has a reliability in excess of a particular minimum value is suited for research purposes or individual decision making

(Charter, 2003). For research purposes in which group means and correlations are of interest, a minimum reliability of .80 is considered adequate. In contrast, a minimum reliability of .90 is often considered necessary for drawing inferences about individuals based on their test scores (Kline, 2000, p. 13; Nunnally & Bernstein, 1994, p. 265). Interestingly, Clark and Watson (1995) found that some test constructors considered reliabilities in the range .60 to .70 to be sufficient.

However, reliability is a group characteristic, but reliability is not informative for making decisions about individuals (Mellenbergh, 1996). Knowing that a test score has reliability equal to, say, .85, helps little to make a decision as to whether John should be assigned to a therapy group on the basis of his fallible test score. Instead, practitioners need the standard error of measurement (SEM), which is deduced from the definition of the reliability by first noting that $\sigma^2_{X_+} = \sigma^2_T + \sigma^2_E$, using this result to rewrite reliability as $\rho_{XX'} = 1 - \sigma^2_E/\sigma^2_{X_+}$, and then rewriting such that the SEM results,

$$SEM = \sigma_E = \sigma_{X_+}\sqrt{1 - \rho_{XX'}}.$$

The SEM is used to compute confidence intervals (CIs) for true scores. The smaller the CI, the higher the measurement precision. For example, we assume that the test score X_+ is an estimate of the true score T, such that $\hat{T} = X_+$, and that random measurement error is normally distributed with mean equal to T and variance σ^2_E, such that $E \sim N(T, \sigma^2_E)$. Given the standard deviation of the test-score distribution (S_{X_+}) and the estimated reliability coefficient ($r_{XX'}$), the sample value for SEM, denoted S_E, is computed by

$$S_E = S_{X_+}\sqrt{1 - r_{XX'}}.$$

Then, a 95% CI for T is obtained as $[X_+ - 1.96S_E; X_+ + 1.96S_E]$. This CI can be used, for example, to test at a 5% significance level whether John's test score is significantly different from a particular cut-score, X_c. If cut-score X_c is inside the estimated CI, then the test score does not differ significantly from the cut-score.

To investigate how measurement precision is affected as the test grows shorter, the scale length needs to be considered as well (Sijtsma, 2009a). The scale length is the maximum possible test score minus the minimum possible test score. As items are removed, CIs are narrower but the scale length also is shorter and shrinks at a larger pace (Sijtsma, 2009a). Consequently, CIs for the shortened test may encompass a larger proportion of the scale length than those for longer tests, thus leaving less room for test scores to differ significantly. To study the relationship between measurement precision and test length, we defined measurement precision by the ratio of the 95% CI and the scale length, which we called the relative

95% CI. The larger the relative 95% CI, the less precise measurements are relative to the scale length.

Finally, most textbooks use the Spearman-Brown formula to relate reliability to test length (e.g., Allen & Yen, 1979, p. 85; also, see Lord & Novick, 1968, p. 112). The formula predicts the reliability of a shortened test under the ideal circumstance that the test consists of parallel items. In practice, items are not parallel and the reliability of the shortened test is different from what the Spearman-Brown formula predicts. Therefore, the question to what extent the reliability and measurement precision of a real test reduces when the test is shortened is an empirical issue.

Validity

Validity is not unambiguously defined (Borsboom, Mellenbergh, & Van Heerden, 2004; Lissitz, 2009; Sijtsma, 2009a). Basically, there are two types of validity: construct-related validity and prediction-related validity (Evers, Sijtsma, Lucassen, & Meijer, 2010). Next, we discuss these two types of validity and possible effects of test shortening on the two types of validity.

Construct-related validity entails the degree to which a test measures the construct of interest, and is sometimes assumed to also include content validity. Construct validity is often ascertained by means of factor analysis so as to assess the internal structure of a test, and the correlations of the test scores with scores on tests that are assumed to measure the same construct, which establishes convergent validity, or a different construct, which establishes divergent validity. An unfortunate test-shortening that results in a poor or an incomplete construct representation impairs construct-related validity. The statistics-driven strategy is particularly vulnerable to this problem as this strategy is "blind" to item content, and the way some statistical methods select items may bias the shortened test's validity. In particular, strategies that select items on the basis of their correlation with other items have a tendency to select items that are similar in content, so that the content domain of the shortened test may be unintentionally narrowed (Clark & Watson, 1995; Stanton et al., 2002). Item selection based on high factor loadings in explorative factor analysis tends to produce the same bias, with shortened tests that no longer fully cover the content of the original scales (Coste et al., 1997; Reise, Waller, & Comrey, 2000). Such a shift in meaning may produce instrumentation bias (Cook & Campbell, 1979, p. 52).

Prediction-related validity refers to the degree to which test scores accurately predict a criterion. Prediction-related validity is often assessed by correlating test scores with scores on a criterion measure or by studying the test score's classification accuracy. Assuming that the items in the longer test all measure the same construct and that removal of items leads to lower test-score reliability, prediction-related validity decreases as the greater influence of random measurement error lowers the correlation of the resulting test score and the criterion score (Allen &

Yen, 1979, p. 98). However, if test shortening leads to construct misrepresentation then the influence of the misrepresentation on the test-criterion correlation is difficult to predict. If the items that represent particular aspects of the construct correlating with the criterion measure are left out, then the resulting test score correlates lower with the criterion. If items are removed that do not correlate with important aspects of the criterion measure, then the test-criterion correlation might even increase, provided greater unreliability does not drive the correlation downward.

Evidence for the validity of the original test does not automatically transfer to the shortened test version (Smith et al., 2000). Validity of the shortened test therefore needs to be demonstrated in additional validation research, also if the longer version had adequate validity. Previous reviews (Coste et al., 1997; Levy, 1968; Smith et al., 2000) suggested that test constructors tend to take the validity of the shortened tests for granted and refrain from additional validation research at all.

The validity of the shortened test may be judged by experts or investigated by means of statistics-driven strategies. In practical test-shortening studies, statistical validation studies are performed using the following types of data samples (Smith et al., 2000). First, statistical analyses can be done by reanalyzing the data which was also used to construct the shortened test or by reanalyzing data collected by means of the longer test in a different sample but without the deleted items. However, results from the reduced data set are vulnerable to chance capitalization due to item selection, especially if the sample size was modest. Also, answers to items included in the shortened test may have been influenced by answers to the deleted items. This influence would have been absent if only the shortened test version would have been administered. Second, both test versions may be administered to the same respondents but here validation results may be confounded by memory and learning effects because the same items were administered twice to the same persons. The best strategy is to validate the shortened test in a new, independent sample drawn from the target population, but this strategy is rarely used in practice (Coste et al., 1997; Smith et al., 2000; Stanton et al., 2002). In the present study, we investigated to what extent validity issues related to test shortening are studied and if so, which validation strategies are used.

METHOD

Sample

The sample consisted of shortened tests reported in articles published in six peer-reviewed psychological journals. The unit of analysis was test pairs (longer tests and shortened tests) and the unit of observation was research articles. The journals included in our review were selected as follows. We screened a large number of leading peer-reviewed psychology journals for the presence of research using

tests or questionnaires. Of all journals regularly reporting on test and questionnaire construction, we chose six journals such that four major areas in psychology were covered. For four journals, *Journal of Personality Assessment* (personality psychology: 6 issues per year), *Personnel Psychology* (personnel psychology: 4 issues per year), *International Journal of Selection and Assessment* (personnel psychology: 4 issues per year), and *Psychological Assessment* (clinical psychology: 4 issues per year), we listed all shortened tests found in articles starting in 2005 and up to and including 2010. The fifth and sixth journals, *Personality and Individual Differences* (personality psychology: 16 issues per year) and *Journal of Psychosomatic Research* (medical psychology: 12 issues per year), published a considerably larger number of issues compared to the other reviewed journals. For both journals, we included only issues that appeared in 2009 and 2010 and added four randomly drawn issues from the previous period (2005–2008) to prevent these two journals from dominating the results.

To identify shortened tests, we looked for all references to tests that mentioned the adjectives "abbreviated," "abridged," or "shortened," or were referred to as "short form." This search strategy may have excluded shortened tests that were not explicitly labeled as such but a pilot study suggested such cases to be rare. Some reviewed articles contained references to short tests as possible alternatives to the measures used in the study but these tests were not actually used; hence, we did not include them in our review.

If different shortened versions of the same test were found, each version was included as a separate shortened test in the sample. Shortened tests that were used in multiple articles were only included once in our analyses. When information about the reliability and validity of a test was provided in multiple articles, this information was combined to obtain a better picture of the psychometric properties of the test. Some shortened tests were constructed without a reference to the longer test, but they were presented instead as a shortened version of a previously shortened test. In these instances, we considered the initially shortened form as the original test and compared the properties of both shortened versions with each other.

The selection criteria resulted in the exclusion of some commonly used shortened tests. For example, the NEO-FFI includes a number of items that were not incorporated in the original NEO-PI-R (Costa & McCrae, 1992); the Short Form Health Survey consists of 36 items taken from various longer tests (Ware, Kosinski, & Keller, 1996); both the 100- and 50-item versions of the International Personality Item Pool were derived from the same 2000+ item pool (Goldberg, 1992); and the shortened 370-item version of the Minnesota Multiphasic Personality Inventory (MMPI)-2 contains all the original scales from the MMPI (Butcher & Hostetler, 1990).

The first author (PK) screened all articles for the relevant information. For this purpose, he used a detailed coding scheme, which was developed in a pilot study. After two months, all articles were coded a second time to secure consisting

coding. If the required information was not reported in the reviewed article, PK consulted the article or the book that presented the longer test or the shortened form to obtain the necessary data.

Variables

Table 1 lists all studied variables, which were selected on the basis of a review of the psychometric literature and a pilot study that we performed, and Table 1 also lists the variables' operationalizations. A few remarks are in order. First, some tests are a composite of separately used subtests, each measuring a different construct. For example, the shortened Depression Anxiety Stress Scales yields two scale scores that should be interpreted separately (Lovibond & Lovibond, 1995). We considered a test to be a composite if the authors indicated that test scores should be reported and interpreted at the subtest level. For composite tests, we studied the magnitude of test shortening and the effects of test shortening on reliability and measurement precision at the subtest level.

Second, different methods for estimating test-score reliability may produce different reliability values, and therefore cannot be used interchangeably (American Educational Research Association, 1999, p. 32). Our pilot study showed that most studies report coefficient alpha and rarely report other estimates such as test-retest reliability. Therefore, to study the effects of test shortening on reliability we rely on the reported coefficient alphas.

Third, for the same test, test-score reliability has different values in different populations. For example, reliability is lower in homogeneous populations (i.e., small true-score variance) than in heterogeneous populations (i.e., large true-score variance; Allen & Yen, 1979, pp. 34–36). Hence, we only compared reliability and measurement precision of tests pairs for which we could assume that these properties were estimated in random samples from the same population. We made this assumption when both samples came from either a general population or a clinical population *and* both tests were administered in the same language, which was often English. We did not further distinguish clinical subpopulations.

Fourth, our pilot study revealed instances of test shortening in which the items from all scales were pooled and a selection of these items was used to form one or more new subtests. An example is the 24-item Locus of Control Scale, which comprises three subtests each covering one of three dimensions, whereas the shortened 9-item version is based on a clustering of items from the three subtests into one general test (Levenson, 1973; Noone, Stephens, & Alpass, 2010). As a result, the items in the shortened subtests came from different longer subtests. Such a redistribution of items into smaller scales confounds the effect of test-shortening on test-score reliability and measurement precision. Therefore, we only compared test length, reliability, and measurement precision of subtest pairs for which all items in the short subtests came from the corresponding longer subtests.

TABLE 1
Variables and Operationalizations

Category	Explanation
Function of the Shortened Test in the Reviewed Article	
Constructed only	The authors constructed the shortened test but did not use the test for empirical research
Constructed and used	The authors constructed the shortened test and used the test for empirical research
Used only	The authors used an available shortened test for empirical research
Psychological Domain	
Mental abilities	e.g., intelligence
Personality traits	e.g., extraversion
Interests, values, and attitudes	e.g., disgust
Psychopathology	e.g., depression
Motivation to Develop or Use a Shortened Test Instead of the Longer Form in the Reviewed Article	
Savings in time and increased efficiency	
Investigating the psychometric properties	The reliability and validity of the shorter test were investigated
Improved psychometric properties	The shortened test was considered better than the longer test
Inappropriate content of some items in the longer test	
Shortened test was effectively used in previous research	
No motivation given	
Test Length of the Longer Form and the Shortened Form	
Number of items and scale length	Per subtest for composite tests; we considered a test to be a composite if authors indicated that scores should be interpreted at the subtest level
Scale range	Difference between the maximum possible test score and the minimum possible test score
Test-Shortening Strategy	
Statistics-driven strategy	
Judgmental strategy	
Ad-hoc strategy	
Unknown	The information was neither reported in the reviewed article nor the article presenting the shortened form

Statistics-Driven Strategy (If Applicable)	
Classical test theory	e.g., coefficient alpha, correlations, proportions, shape of the test-score distribution
Item response theory	e.g., test information
Factor analysis	e.g., loadings, model fit
Regression analysis	e.g., regression weights
Classification accuracy	e.g., Cohen's kappa, sensitivity, specificity
Purpose of the Statistics-Driven Strategy (If Applicable)	
Maintaining test-score reliability close to the test-score reliability of the longer test	e.g., removing items with the lowest loadings in factor analysis
Maintaining validity close to the validity of the longer test	e.g., removing items with cross-loadings in factor analysis
Reliability and Measurement Precision of the Longer Form and the Shortened Form	
Sample used to estimate test-score reliability	General population versus clinical population
Language in which the test was administered	
Method used to estimate test-score reliability	Coefficient alpha
Standard deviation of the test-score distribution (S_{X_+})	
Estimate of the standard measurement error (S_E)	
Included information	We applied the following rules when multiple reliability estimates were given for the same population (i.e., general versus clinical): • When coefficient alpha and S_{X_+} or S_E were reported for various samples of either a general or clinical population (e.g., males and females), we took the mean values. Differences between reported values for these samples were smaller than .05. • When coefficient alpha was reported for various samples, but either S_{X_+} or S_E was reported for only a subset of these samples, we considered only those samples for which coefficient alpha and either S_{X_+} or S_E were reported. • When coefficient alpha was reported in a single article for multiple time points or replications, we took the first value for which either S_{X_+} or S_E was given. If either S_{X_+} or S_E were unreported, we included the first reported coefficient alpha.

(Continued on next page)

TABLE 1
Variables and Operationalizations *(Continued)*

Category	Explanation
Types of Evidence Given for the Validity of the Shortened Test	
Construct-related validity	Evidence given both in the reviewed article and the article presenting the shortened form
Prediction-related validity	Evidence given both in the reviewed article and the article presenting the shortened form
Only references in the reviewed article	No evidence given, but the reviewed article states that the shortened test "has excellent psychometric properties," "is a valid measure," and so on, and cites validation studies
No validity evidence	No evidence given and the reviewed article does not refer to validation studies
Statistics-Driven Strategy Used to Investigate the Validity of the Shortened Test (If Applicable)	
Classical test theory	
Item response theory	
Factor analysis	
Regression analysis	
Classification accuracy	
Sample Used to Investigate the Validity of the Shortened Test (If Applicable)	
Using the data that were also used to construct the shortened version	
Reanalyzing data of the longer version obtained in a different sample	
New data on the shortened test obtained in a sample of respondents who had previously completed the longer version	
New sample, which filled out the shortened test but not the longer test	
Discussion Pay-Off Advantages Versus Losses in Reliability and Validity in the Reviewed Article	
Discussion advantages realized	
Discussion of pay-off	

RESULTS

Research Question 1: What Is the Current Practice of Using Shortened Tests?

Short-Test Use. Among 2273 reviewed articles, we found 164 shortened tests reported in 170 articles. The longer tests comprised 380 subtests (range was 1–22 subtests per test) and the shortened tests 346 subtests (range was 1–20 subtests per test). We identified 281 pairs of subtests, coming from 133 tests, for which all items of the shortened version came from the corresponding longer subtest. For the other 31 shortened tests, the number of subtests varied between the short and the long versions or insufficient information was available to determine whether all items in the short subtests came from the corresponding longer subtests.

In *Psychological Assessment* and *Personnel Psychology*, we found the highest percentages of articles in which shortened tests were reported but percentages varied little among the six journals (Table 2). Each of 146 tests were reported in one reviewed article and 18 tests were reported in at least two articles. Frequently used shortened tests included the shortened 30-item Trait Emotional Intelligence Questionnaire (Cooper & Petrides, 2010, 153 items included in the longer form, used in 7 studies); the Mini-markers (40 items), which is a reduced version of Goldenberg's 100-item Unipolar big-five markers (Saucier, 1994, used in 6 studies); and the 12-item Health Survey (Ware et al., 1996, 36 items included in the longer form, used in 5 studies).

Among the 164 shortened tests, 7 tests measured mental abilities, 56 tests measured personality traits, 53 tests measured interests, values, or attitudes, and

TABLE 2
Prevalence of Shortened Tests

	Articles		
Journal	# Reviewed	# Shortened Tests Included	Percentage
Journal of Personality Assessment	406	20	4.93
Personnel Psychology	160	14	8.75
International Journal of Selection and Assessment	230	13	5.65
Psychological Assessment	366	42	11.48
Personality and Individual Differences	764	63	8.25
Journal of Psychosomatic Research	347	18	5.19
Total reviewed	2273	170	7.48

48 tests measured psychopathology. Among all shortened tests, 27 tests were proposed as an alternative to the longer version but not applied in empirical research, 57 tests were constructed and used for empirical research in the same article, and 80 tests were used for empirical research but developed in another study (including all tests which were found in multiple articles). For 114 shortened tests, we had access to the article or the book in which the longer form was presented. For 55 shortened tests that were developed in another study, we were able to retrieve the publication in which the shortened test was presented.

Motivation to Construct and Use Shortened Tests. Motivations for test shortening could be traced for 70 tests. Motivations for constructing a shortened test in the reviewed articles were the following. Twenty-one short tests were constructed to achieve savings in time and produce increased efficiency; 22 tests were shortened because test constructors believed that the shortened tests had superior psychometric properties (i.e., improved reliability and validity) compared to the longer version; and 10 tests were shortened because some items had inappropriate content for the target population. For example, McCarthy, Pedersen, and D'Amico (2009) found it inappropriate to ask children about sexual behavior and therefore removed two items from the 15-item short form of the Comprehensive Effects of Alcohol questionnaire (Ham, Stewart, Norton, & Hope, 2005). Motivations for using a shortened test that was developed in another study included: savings in time (mentioned 5 times), superior psychometric properties compared to the longer version (mentioned 6 times), investigating the psychometric properties of the shortened test (mentioned 14 times), and proven useful in previous research (mentioned 9 times).

Magnitude of Test Shortening. For 201 subtests coming from 106 test pairs, we were able to retrieve the number of items for both the longer and the shortened versions. Figure 1 shows test lengths for the shortened and the longer tests, and Figure 2 shows the frequency distribution of test-shortening factor K (i.e., the ratio of the novel test length and the original test length). The mean number of items reduced from 19.28 to 9.36. Most subtests were shortened by removing 40% to 60% of the items. However, Figure 2 also shows that for 21 subtests test shortening involved a reduction of items equal to 80% or more. For these tests, the mean number of items decreased from 29.86 to 6.33. We did not find a relationship between the magnitude of the decrease of test length and the psychological domain.

Test-Shortening Strategies. We were able to retrieve information about the test-shortening strategy used for 111 tests. Statistics-driven strategies were used for 92 tests, the judgmental approach was used for 45 tests, and the ad hoc strategy was used for 17 tests. Single strategies shortened 76 tests, combining two strategies

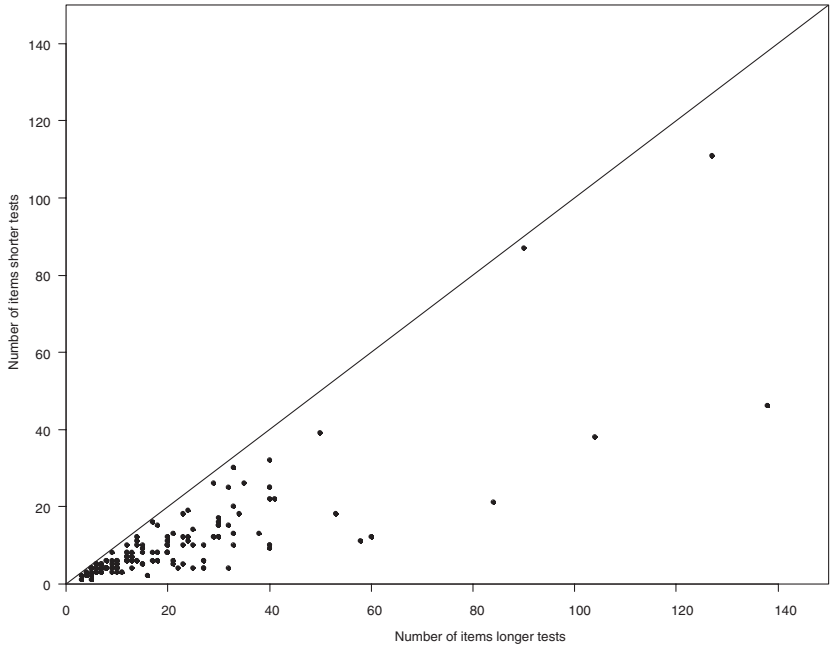

FIGURE 1
Comparison of test length of the longer tests and the shortened tests.

shortened 27 tests and combining all three strategies shortened eight tests. For 53 shortened tests, information about the test-shortening strategy was absent.

Table 3 (columns 1–4) shows that among the statistics-driven, classical-test-theory strategies and factor analysis were dominant. For only ten tests item response theory was used to remove items. This is a remarkably low number given

TABLE 3
Frequencies of Statistics-Driven Strategies Used to Shorten Tests

	Item-Selection Purpose			
Strategy	Maintaining Reliability	Maintaining Validity	Both	Validation
Classical test theory	44	2	10	115
Item response theory	10	0	0	8
Factor analysis	32	4	26	54
Regression analysis	1	1	0	23
Classification accuracy	0	1	0	11

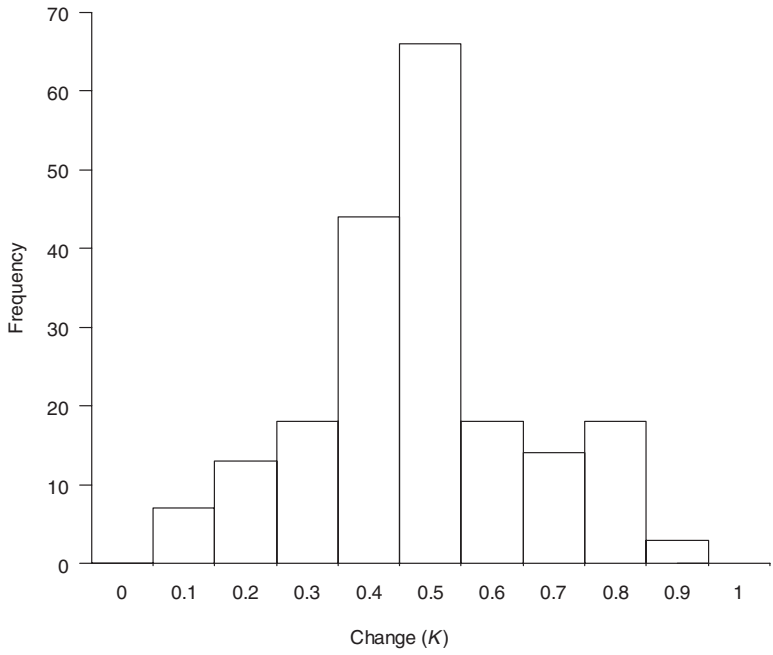

FIGURE 2
Test shortening factor K for the shortened tests.

the growing popularity of item response theory in psychological assessment nowadays. Most statistics-driven strategies removed items that contributed the least to the reliability and thus minimized reliability reduction. When statistical procedures were used, either to optimize reliability or validity, the final decision about which items to remove was often somewhat arbitrary. For example, it remains unclear why Tangirala and Ramanujam (2008) selected the five items with the highest factor loadings from Mael and Ashforth's (1992) identification scale and not, for example, four items.

A combination of a statistics-driven strategy and a judgmental strategy was used for 28 tests. However, for most of the shortened tests the procedure was not performed as thoroughly as advised in the literature (Coste et al., 1997; Smith et al., 2000). Exceptions included the SKILLSCOPE® (Kaiser, Lindberg, & Craig, 2007) and the Material Values Scale (Richins, 2004). For example, items for the shortened SKILLSCOPE® were selected as follows. First, three experts had to judge the relevance of all items in an iterative procedure based on background information about the measured constructs. Next, Kaiser and colleagues (2007) determined which items to include in the shortened version using the opinion of these experts and the

results of several statistical analyses. In general, when test constructors used a judgmental approach, they typically reported that they decided which items to retain on the basis of the items' perceived relevance (e.g., Mason, Linney, & Claridge, 2005).

Research Question 2: What Are the Differences Between Reliability, Construct-Related Validity, and Prediction-Related Validity of the Shortened Test and the Longer Test?

Reliability. Coefficient alpha was reported for 291 subtests included in 116 shortened tests. The reliabilities of these subtests ranged from .09 to .96 with a mean equal to .78; 129 subtests had a reliability below .80 and 47 subtests had a reliability below .70. For 137 pairs of longer subtests and shortened subtests coming from 62 tests, coefficient alphas were available for the same type of population (i.e., either a general population or a clinical population). For these pairs, Figure 3 shows the relationship between the reliability of the original subtests and the reliability of the shortened subtests. Ignoring the tests that were shortened to improve the psychometric properties of the test (triangles in Figure 3), reliability decreased as a result of removing items for 102 subtests, was similar for 5 subtests, and improved for 17 subtests (mean improvement was .03).

Test shortening caused mean reliability to reduce from .84 to .77 (Table 4, columns 3–4). For 33 original tests, reliability was below .80 and for 6 original tests below .70. Test shortening resulted in 31 additional tests with reliability below .80 and 18 additional tests with reliability below .70. Among the 110 subtests that had lower reliability, 36 subtests showed a reduction of reliability that was less than .05, 35 subtests showed a reduction between .05 and .10, and 39 subtests showed a reduction in excess of .10. The largest decrease was found for the subtest Unassuming-Ingenuous Scale of the Battery of Interpersonal Capabilities (Hofsess & Tracey, 2005). For this subtest, which was shortened from 16 to two items, reliability decreased from .80 to .09. The effect of separate test-shortening strategies on reliability could not be assessed. For most pairs of longer and shortened subtests for which coefficient alphas were available, either a statistics-driven strategy or a combination of at least two of the strategies was used to shorten the tests and, moreover, the mean K varied between .43 and .68 across strategies.

Measurement Precision. For 21 pairs of longer subtests and shortened subtests coming from 15 tests, we computed the relative 95% CI for both subtests. For all others tests, we lacked complete information about the scale range, coefficient alpha, and the standard deviation of the test scores. Table 5 shows descriptive statistics for the relative 95% CIs (mean K was .49). Measurement precision decreased little, and on average the CIs covered about one-third of the test-score

TABLE 4
Overview of the Test-Shortening Strategies and Reliability of the Longer Tests and the Shortened Tests

			Reliability											
			Mean		Standard Deviation		Minimum		Maximum		$r_{XX'} < .80$		$r_{XX'} < .70$	
Strategy	Frequency[a]	$\bar{K}$	L	S	L	S	L	S	L	S	L	S	L	S
S	44	.52	.84	.80	.06	.08	.69	.57	.95	.92	12	18	1	4
J	3	.68	.86	.75	.05	.10	.80	.63	.89	.82	0	2	0	1
A	1	.60	.90	.89			.90	.89	.90	.89	0	0	0	0
S & J	37	.52	.80	.75	.10	.12	.67	.50	.97	.92	20	21	5	11
S & A	13	.43	.86	.74	.04	.06	.80	.68	.93	.88	0	10	0	2
J & A	0													
S & J & A	13	.47	.89	.85	.03	.02	.85	.80	.94	.88	0	0	0	0
Unknown	26	.36	.89	.73	.05	.20	.77	.09	.97	.93	1	13	0	6
Total	137	.49	.84	.77	.07	.12	.67	.09	.97	.93	33	64	6	24

Notes. $\bar{K}$ = mean K; L = longer test; S = shorter test. Strategies: S = statistics-driven strategy; J = judgmental strategy; A = ad-hoc strategy.
[a] Frequency = pairs of subtests.

TABLE 5
Measurement Precision of the Longer Tests and the Shortened Tests
(for 17 Pairs of Subtests)

Statistic	Relative 95% CI	
	Long	Short
Mean	.26	.32
Standard deviation	.07	.11
Minimum	.15	.16
Maximum	.41	.54

Note. Relative 95% CI equals the ratio of the width of the 95% CI and scale length.

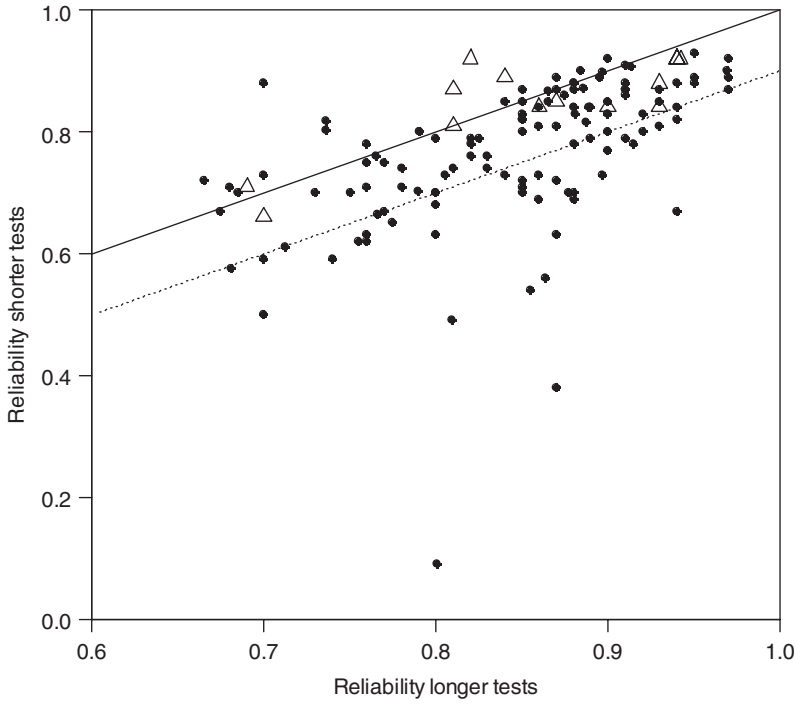

FIGURE 3
Difference between reliability of the longer tests and reliability of the shortened tests.
Notes. Triangles represent subtests shortened to improve the psychometric properties of the original version; dots represent all other subtests. The solid line is the identity line; subtest pairs at the solid line have equal reliability. The dashed line represents a difference in reliability of .10; the difference between the reliability of the longer and the shortened test was more than .10 for subtest pairs below the dashed line.

TABLE 6
Frequencies of Validity Evidence for the Shortened Tests

Source of the Validity Evidence	Status of the Shortened Test: Constructed Only	Constructed and Used	Used Only	Total
Construct-related validity	16	8	45	69
Prediction-related validity	1	0	1	2
Both construct-related validity and prediction-related validity	8	0	13	21
Only references in the reviewed article	—	—	8	8
No validity evidence given	2	49	13	64
Total	27	57	80	164

scale. The 95% CI of the shortened version of the social desirability scale of the Revised Junior Eysenck Personality Questionnaire (George, Connor, Gullo, & Young, 2010) extended even across more than half of the scale range (i.e., a relative 95% CI of .54). However, given the small number of subtests for which information was available, caution should be exercised in generalizing these results to other shortened tests.

Validity. For 100 tests it was reported that the validity of the shortened version had been studied. Detailed information about the validation strategies followed was available for 92 tests (Table 6). For the other eight tests, test users provided references to validation research without providing information about validation strategies they followed. Construct-related validity evidence was reported for 90 tests, while prediction-related validity evidence was only available for 23 tests. This is interesting because tests are often used to predict particular outcomes (e.g., therapy, behavior, employee success). For 49 of the 57 shortened tests that were constructed *and* used to collect data for research, evidence supporting their validity was absent. As these tests were nevertheless used in empirical research, this result suggests that the researchers took the validity for granted.

For the 92 tests for which validation strategies were reported, validity of the shortened test was always assessed by applying statistics-driven strategies. However, validation studies showed much variation with respect to the number of different statistics-driven strategies used. Twenty-two tests were validated on the basis of only one particular statistical strategy, and three tests used all five statistical strategies listed in Table 3. Most validation studies were based on a comparison of psychometric properties of the shortened test and the longer version using classical test theory statistics and factor analysis (Table 3, fourth column). According to the test constructors, the higher the value of the statistic (e.g., the correlation

between test scores of the short and the longer versions) or the more similar the statistic for both test versions (e.g., the factor solution), the stronger the evidence that test shortening did not affect validity.

Information about the samples used for validation was given for 89 tests. Forty-three tests were validated using the data that were also used to construct the shortened version. For 46 shortened tests, validity was investigated by reanalyzing data of the longer version obtained in a different sample. For two shortened tests, new data on the shortened test were obtained in a sample of respondents who had previously completed the longer version. Forty-three tests were validated in a new sample that completed the shortened test but not the longer test.

Research Question 3: To Which Extent Do Researchers Pay Attention to the Potential Advantages and Potentially Negative Implications of Using Shorter Tests for Their Own Research?

For all 22 tests that were shortened to improve reliability and validity, the advantages of test shortening for the reliability and the validity were mentioned. For 11 of the 26 tests that were shortened to have more efficient instruments to be used in empirical research, the pay-off between efficiency gain and potential loss of reliability and validity was discussed. Conditions were discussed under which psychologists could use the shortened test instead of the longer version. Actual time savings were only reported for 6 of the tests that were shortened with the aim to save administration time. For example, for the Short Children's Behavior Questionnaire, which includes of 94 of the 177 items of the original version, testing time was reduced by 30 minutes (Putnam & Rothbart, 2006). For the shortened version of the Profile of Moods States, testing time was reduced by about one half (Shacham, 1983). For the other 125 shortened tests, authors of reviewed articles did not discuss realized advantages and potential disadvantages of using shortened versions or possible pay-offs if the longer version would have been used.

DISCUSSION

We reviewed 2273 articles that appeared in six leading journals (2005–2010) and identified 164 shortened tests in 170 articles. Most shortened tests were found in the noncognitive domain (e.g., personality traits and psychopathology). Given the huge number of psychological tests that are available nowadays, 164 shortened tests may not give the impression that test shortening is a significant trend, but one should keep in mind that the reviewed journals do not have as their goal to publish shortened tests. A quick scan of other psychological journals confirmed that the use of shortened tests is quite common. Moreover, we found numerous translations of shortened tests but ignored these tests due to the specific problems associated with

translated tests. Hence, we believe that the finding of 164 shortened tests suggests a significant trend rather than an incident, and this trend needs further scrutiny.

With the exception of a few tests, for most tests the shortened version had a reliability that was only little lower than that of the longer version, and still satisfied minimum reliability as required in the psychometric literature. However, it was particularly striking that shortened tests were derived from longer tests that often had only modest reliability to begin with. The literature often did not provide reasons for test shortening when initial reliability was already modest and one can only guess why researchers shortened their tests. We speculate that researchers tend to report reliability more as a ritual and by default consider a modest reliability, say, between .70 and .80, high enough for the application envisaged, and therefore readily accept losing half of the items for reasons of practical efficiency.

We further speculate that many researchers are unaware of the important distinction between the group-level characteristic of reliability and the individual-level characteristic of measurement precision. Because only 15 of 164 tests allowed us to compare measurement precision between the longer version and the shortened version, we believe researchers often do not realize that shortening their modest-reliability test to only half or less than half its length while losing only little reliability does not mean that measurement precision remains the same. On the contrary, measurement precision tends to be impaired considerably when the test grows shorter and a reliability that drops only little does not show this (Emons et al., 2007; Sijtsma & Emons, 2011). Loss of measurement precision means that decisions about individuals are more uncertain to a degree that uncertainty may become unacceptable for many decisions. The result may be a large number of decision errors (Emons et al., 2007; Kruyen et al., 2012).

Another concern with respect to low measurement precision is that test users increasingly are required to report scores per subdomain in addition to total scores so as to provide diagnostic information to the client. Sinharay, Puhan, and Haberman (2010) showed that the diagnostic value of subdomain scores often is limited because subdomain scores are based on a few items and, as a result, they are unreliable. Shortening a test may result in even fewer items per subdomain and thus further diminishes the diagnostic value of domain scores, and this happens even though reliability reduces just a little.

The literature review revealed that statistics-driven strategies were often used, sometimes in combination with other strategies, to maximize coefficient alpha of the shortened test. As highly correlating items produce high alphas, the selected items probably correlated highly whereas items correlating lowly were not selected. Items correlating highly often have similar content and, consequently, selecting such items tends to narrow construct coverage (e.g., Reise & Waller, 2009). Another problem is that the shortened test may consist of similarly worded items. Consequently, respondents noticing these similarities may purposively but

incorrectly match the responses from different items to ensure response consistency. This response strategy reduces the number of items that effectively contribute to measurement, and reliability of the shortened test based on coefficient alpha is artificially high. For tests that were shortened using a judgmental approach, the selection of items was typically informal and based on perceived relevance. When an ad hoc selection strategy was used, construct coverage was not considered at all.

Given the importance of using valid tests, it is remarkable that validity evidence was only available for half of the shortened tests. For the other tests, researchers ignored validity issues or assumed that validity evidence for the longer test transferred to the shortened test. These results are consistent with previous reviews (Coste et al., 1997; Levy, 1968; Smith et al., 2000). Studies that did investigate short-test validity suffered from methodological shortcomings, mostly related to using the same sample that also had been used for item selection rather than relying on a new sample. Because validity of the longer test does not automatically transfer to the shortened test, we conclude that test constructors and test users should give more attention to validity issues, both from a substantive and a methodological perspective.

Ignoring reliability and validity issues, an important and complex question is what the practical implications are of using shortened tests to address substantive research questions. This question does not have a single answer, but an important concern is whether test shortening produces bias in the statistical outcomes of psychological research. For example, we know from CTT that estimates of associations by means of the correlation coefficient become more biased when less reliable measurements are used (Lord & Novick, 1968, pp. 114–118). Sijtsma and Emons (2011) provided several numerical examples showing, for example, that if the population correlation is .60, the sample correlation coefficient reduces from .54 to .36 if the reliability of the variables reduces from .90 to .60. Reduced and therefore biased construct-related validity may further distort product-moment correlations. However, the effects of reduced reliability on the power of statistical tests seem to less problematic. For example, Sijtsma and Emons (2011) found little effect of lowered reliability on the power of the Student's t-test even when reliability decreased from .90 to .70. Thus, the effect of lowered reliability on statistical results should be considered separately for each statistical problem.

We provide four recommendations. First, it seems that researchers who shorten their test often aim at arriving at the minimum reliability prescribed by rules of thumb, thus assuming that this is a safe value for test applications. We recommend researchers to aim higher so as to provide the best possible quality for the persons who are tested. Second, without exception it seems that researchers are unaware that aiming for particular reliability values is a ritual, and that for individual decision making one rather needs to know the test's measurement precision. Information about measurement precision was scarce but should be standard in test

reports (American Educational Research Association, 1999, pp. 29, 31). Third, we recommend researchers not to automatically transfer validity results for the longer test version to the shorter test version, but to investigate the validity of shorter test in a separate sample. Fourth, we recommend researchers who consider the use of a shortened test to gauge the implications for the statistical analyses in the application envisaged and appraise the pay-off between gained practical efficiency and losses in statistical accuracy.

REFERENCES

Alden, L. E., Wiggins, J. S., & Pincus, A. L. (1990). Construction of circumplex scales for the inventory of interpersonal problems. *Journal of Personality Assessment*, *55*, 521–536. doi: 10.1080/00223891.1990.9674088

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

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

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York, NY: Macmillan.

Beck, A. T., & Beck, R. W. (1972). Screening depressed patients in family practice. A rapid technic. *Postgraduate Medicine*, *52*, 81–85.

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. doi: 10.1007/BF02294079

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, *11*, 303–315. doi: 10.1002/(SICI)1099-0984(199711)11:4<303::AID-PER292>3.0.CO;2-#

Butcher, J. N., & Hostetler, K. (1990). Abbreviating MMPI item administration: What can be learned from the MMPI for the MMPI-2? *Psychological Assessment*, *2*, 12–21. doi: 10.1037/1040-3590.2.1.12

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology*, *130*, 290–304. doi: 10.1080/00221300309601160

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319. doi: 10.1037//1040-3590.7.3.309

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Cooper, A., & Petrides, K. V. (2010). A psychometric analysis of the Trait Emotional Intelligence Questionnaire-Short Form (TEIQue-SF) using item response theory. *Journal of Personality Assessment*, *92*, 449–457. doi: 10.1080/00223891.2010.497426

Costa, P. T., Jr., & McCrae, R. R. (1992). *The NEO PI/FFI manual supplement*. Odessa, FL: Psychological Assessment Resources.

Coste, J., Guillemin, F., Pouchot, J., & Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology*, *50*, 247–252. doi: 10.1016/S0895-4356(96)00363-0

Derogatis, L. R. (2001). *Brief Symptom Inventory (BSI)-18 administration, scoring and procedures manual*. Minneapolis, MN: NCS Pearson.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13*, 595–605. doi: 10.1017/S0033291700048017

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105–120. doi: 10.1037/1082-989x.12.1.105

Evers, A. V. A. M., 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. doi: 10.1080/15305058.2010.518325

George, S. M., Connor, J. P., Gullo, M. J., & Young, R. M. (2010). A prospective study of personality features predictive of early adolescent alcohol misuse. *Personality and Individual Differences*, *49*, 204–209. doi: 10.1016/j.paid.2010.03.036

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*, 26–42. doi: 10.1037//1040-3590.4.1.26

Ham, L. S., Stewart, S. H., Norton, P. J., & Hope, D. A. (2005). Psychometric assessment of the Comprehensive Effects of Alcohol Questionnaire: Comparing a brief version to the original full scale. *Journal of Psychopathology and Behavioral Assessment*, *27*, 141–158. doi: 10.1007/s10862-005-0631-9

Hofsess, C. D., & Tracey, T. J. G. (2005). The interpersonal circumplex as a model of interpersonal capabilities. *Journal of Personality Assessment*, *84*, 137–147. doi: 10.1207/s15327752jpa8402_03

Kaiser, R. B., Lindberg, J. T., & Craig, S. B. (2007). Assessing the flexibility of managers: A comparison of methods. *International Journal of Selection and Assessment*, *15*, 40–55. doi: 10.1111/j.1468–2389.2007.00366.x

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing, 12,* 321–344. doi: 10.1080/15305058.2011.643517

Lawing, K., Frick, P. J., & Cruise, K. R. (2010). Differences in offending patterns between adolescent sex offenders high or low in callous–unemotional traits. *Psychological Assessment*, *22*, 298–305. doi: 10.1037/a0018707

Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, *41*, 397–404. doi: 10.1037/h0035357

Levy, P. (1968). Short-form tests. A methodological review. *Psychological Bulletin*, *69*, 410–416. doi: 10.1037/h0025736

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

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

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd ed.). Sydney, Australia: Psychology Foundation.

Mael, F. A., & Ashforth, B. E. (1992). Alumni and their alma mater: A partial test of the reformulated model of organizational identification. *Journal of Organizational Behavior*, *13*, 103–123. doi: 10.1002/job.4030130202

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, *78*, 293–296. doi: 10.1016/j.schres.2005.06.020

McCarthy, D. M., Pedersen, S. L., & D'Amico, E. (2009). Analysis of item response and differential item functioning of alcohol expectancies in middle school youths. *Psychological Assessment*, *21*, 444–449. doi: 10.1037/a0016319

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299. doi: 10.1037//1082-989X.1.3.293

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new result. *Psychological Bulletin*, *94*, 52–533. doi: 10.1037/0033-2909.94.3.524

Noone, J. H., Stephens, C., & Alpass, F. (2010). The Process of Retirement Planning Scale (PRePS): Development and validation. *Psychological Assessment*, *22*, 520–531. doi: 10.1037/a0019512

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

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

Paap, M. C. S., Meijer, R. R., Van Bebber, J., Pedersen, G., Karterud, S., Hellem, F. M., . . . Haraldsen, I. R. (2011). A study of the dimensionality and measurement precision of the SCL-90-R using item response theory. *International Journal of Methods in Psychiatric Research*, *20*, e39–e55. doi: 10.1002/mpr.347

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, *87*, 103–113. doi: 10.1207/s15327752jpa8701_09

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. doi: 10.1146/annurev.clinpsy.032408.153553

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297. doi: 10.1037//1040-3590.12.3.287

Richins, M. L. (2004). The material values scale: Measurement properties and development of a short form. *Journal of Consumer Research*, *31*, 209–219. doi: 10.1086/383436

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's Unipolar Big-Five Markers. *Journal of Personality Assessment*, *63*, 506–516. doi: 10.1207/s15327752jpa6303_8

Scheier, M. F., & Carver, C. S. (1992). Effects of optimism on psychological and physical well-being: Theoretical overview and empirical update. *Cognitive Therapy and Research*, *16*, 201–228. doi: 10.1007/BF01173489

Shacham, S. (1983). A shortened version of the Profile of Mood States. *Journal of Personality Assessment*, *47*, 305–306. doi: 10.1207/s15327752jpa4703_14

Shrout, P. E., & Yager, T. J. (1989). Reliability and validity of screening scales: Effects of reducing scale length. *Journal of Clinical Epidemiology*, *42*, 69–78. doi: 10.1016/0895-4356(89)90027-9

Sijtsma, K. (2009a). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*, 167–194. doi: 10.1080/15305050903106883

Sijtsma, K. (2009b). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. doi: 10.1007/S11336–008-9101–0

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*, 565–572. doi: 10.1016/j.jpsychores.2010.11.002

Sinharay, S., Puhan, G., & Haberman, S. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 1–21. doi: 10.1080/00273171.2010.483382

Smith, G. T., McCarthy, D. M., & Anderson, K. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 102–111. doi: 10.1037/1040-3590.12.1.102

Soldz, S., Budman, S., Demby, A., & Merry, J. (1995). A short form of the Inventory of Interpersonal Problems Circumples Scales. *Assessment*, *2*, 53–63. doi: 10.1177/1073191195002001006

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167–194. doi: 10.1111/j.1744-6570.2002.tb00108.x

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlow-Crowne (sic) social desirability scale. *Journal of Clinical Psychology*, *28*, 191–193. doi: 10.1037/h0047358

Tangirala, S., & Ramanujam, R. (2008). Employee silence on critical work issues: The cross level effects of procedural justice climate. *Personnel Psychology*, *61*, 37–68. doi: 10.1111/j.1744-6570.2008.00105.x

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. doi: 10.1007/BF0228-9858

Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, *34*, 220–233. doi: 10.1097/00005650-199603000-00003