Psychological Assessment
2000, Vol. 12, No. 1, 102–111

Copyright 2000 by the American Psychological Association, Inc.
1040-3590/00/$5.00 DOI: 10.1037//1040-3590.12.1.102

On the Sins of Short-Form Development

Gregory T. Smith, Denis M. McCarthy, and Kristen G. Anderson
University of Kentucky

The empirical short-form literature has been characterized by overly optimistic views of the transfer of validity from parent form to short form and by the weak application of psychometric principles in validating short forms. Reviewers have thus opposed constructing short forms altogether, implying researchers are succumbing to an inappropriate temptation by trying to abbreviate measures. The authors disagree. The authors do not oppose the development of short forms, but they do assert that the validity standards for short forms should be quite high. The authors identify 2 general and 9 specific methodological sins characterizing short-form construction and offer methodological suggestions for the sound development of short forms. They recommend a set of 6 a priori steps researchers should consider and 9 methodological procedures researchers can use to develop valid abbreviated forms of clinical-assessment procedures.

The literature on the development of short forms of clinical assessment procedures reflects an interesting dichotomy. On the one hand, researchers have been pursuing short forms for most of the 20th century, apparently beginning when Doll (1917) first asked whether it was necessary to use all of the Binet–Simon items to assess intelligence. Since that time, investigators have developed short forms for virtually every conceivable topic of clinical assessment: As far back as 1968, Levy (1968) catalogued a very wide-ranging set of topics assessed by short forms. On the other hand, numerous reviews, both of particular short forms (for the Minnesota Multiphasic Personality Inventory [MMPI], see Butcher & Hostetler, 1990; for the Wechsler scales, see Silverstein, 1990; Wechlser, 1967) and of general short-form methodology have been severely critical of the practice (Levy, 1968; Smith & McCarthy, 1995). Levy (1968) summarized his review, thus, "by and large the task of producing short forms . . . has become such a self-perpetuating academic activity that many of the essential components of the real problem have been forgotten" (p. 415). Wechsler (1967) asserted that reduction in the number of Wechsler subscales "as a time-saving device is unjustifiable and not to be encouraged" (p. 37). For those who felt there was not sufficient time for a full assessment, Wechsler's (1967) advice was to "find the time" (p. 37).

Of interest, it appears that none of these critical stances has slowed the development of short forms. Researchers have articulated a number of reasons for engaging in this enterprise: Some have tried to develop short forms that offer all the validity of the full-length form; some have developed short forms for "screening purposes"; some have abbreviated measures to fit them into large, multivariate studies; some have abbreviated existing instruments for use with children; and some have reduced behavioral-observation time to save costs. In the current managed health care environment, many clinicians may feel additional pressure to abbreviate assessments because of a failure of insurers to reimburse for more comprehensive evaluations. Thus, the temptation to find ways to measure constructs more quickly than the original test developers thought was necessary is as strong today as it ever was.

There are two classic objections to the development of short forms. The first objection is that it is virtually always a mistake to try to develop a short form. Rigorous, valid, comprehensive assessment is crucial for the evaluation and treatment of many psychological problems, and thus good assessment is worth the time. The inevitable loss of validity for the savings in time is simply not justifiable. The second objection is not directed at short forms per se, but rather at the ways in which they are most often constructed. Short forms have frequently been developed without careful, thorough examination of the new form's validity. From this viewpoint, if we can improve the methodology of short-form development, we may find useful abbreviated instruments.

We take the second view. If developing short forms can be thought of as a kind of succumbing to temptation, then our goals are to describe the methodological sins that are often committed when researchers do succumb to temptation and to propose a set of methodological criteria that would lead to the rigorous development and validation of short forms. A new conscientiousness about careful short-form test development could well presage scientifically supportable and hence valuable abbreviated instruments. We describe two general and nine specific sins of short-form development and offer methodological guidelines to improve the validity of the short forms we pursue so tirelessly. We believe the criteria for judging a short form should be strict: By definition, a short-form developer is attempting to measure a construct or answer a question that the original test developers concluded required a more lengthy assessment.

We surveyed the following journals for articles describing new short forms between 1991 and 1997: *Psychological Assessment*, *Assessment*, *Journal of Personality and Individual Differences*, *Journal of Consulting and Clinical Psychology*, *Addictive Behav-*

Editor's Note. This review was invited by the Editor.

Gregory T. Smith, Denis M. McCarthy, and Kristen G. Anderson, Department of Psychology, University of Kentucky.

Correspondence concerning this article should be addressed to Gregory T. Smith, Department of Psychology, University of Kentucky, Lexington, Kentucky 40506-0044. Electronic mail may be sent to gsmith@pop.uky.edu.

iors, *Journal of Personality Disorders*, *Journal of Personality Assessment*, *Journal of Studies on Alcohol*, *Journal of Personality and Social Psychology*, and *Behavioral Assessment*. Of the many articles that described or used short forms, 12 presented new short forms and the validity evidence for them. We reviewed each article using the criteria we describe below. Results of this evaluation are presented in the following review, and they give an indication of the current state of short-form development (see Table 1 for list of reviewed articles).

The nature of assessment varies as a function of the target construct being assessed. Self-report questionnaires, reports by significant others, behavioral-observation approaches, and other techniques may each be used, depending on the assessment goal. Similarly, appropriate indicators of reliability (e.g., internal consistency, test–retest, interrater) and validity (e.g., correlations with criteria, convergent and discriminant correlations, classification rates, etc.) vary as a function of the context. Some of the issues we raise and the suggestions we offer will only be appropriate for some purposes, but most of the principles we advance are generalizable across assessment and validation method. In our formal review, we largely follow true score theory, and then we briefly discuss item-response theory as an alternative means of short-form development.

Two General Sins of Short-Form Development

There are two fundamental misconceptions that many developers of short forms hold that may underlie the occurrence of any of nine specific sins described below.

The first general sin is that many investigators assume that all of the reliability and validity evidence of the original, full-length measure applies automatically to the abbreviated version. This assumption is false. One can think of a short form as an alternate form of a measure, with (by definition) reduced coverage of the target domain. Of course, a short form's items or observation scenarios are all included on the full-length form, and in that sense, it is not an alternate form. However, its reduced length and content coverage make it a different, alternative assessment, and these features are a built-in disadvantage for the short form. As with any alternate form, and moreso because of the reduced coverage, it is essential to establish independently the reliability and validity of the new, alternative measure. Thus, the burden is on the investigator to show that a new, shortened version of a test or observation procedure is both reliable and valid. A similar burden is on the clinician: Practitioners should refuse to use, as a matter of ethical practice, unvalidated abbreviated measures—whether such measures are presented in journal articles or are promoted by health insurance providers or managed care companies.

The second general sin is to assume that because the new measure is shorter, less validity evidence is required. Although one may be hard pressed to find a researcher who would voice this assumption, the behavior of short-form developers often seems to imply just this reasoning. The reasoning is, of course, patently false. In fact, psychometric theory suggests the opposite: It is harder to have reliability and full content coverage, and hence harder to have validity, with fewer items. A short-form developer must meet the same standards of validity as are required for any test.

Nine Specific Sins of Short-Form Development

Sin 1: Develop a Short Form of an Insufficiently Validated Measure

We judged 5 of the 12 short-form articles we reviewed to have committed this sin. For example, Francis (1996) introduced a short form of the Junior Eysenck Personality Questionnaire—Revised (EPQ–R) from Corulla's (1990) short form of the same instrument, that is, a short form of a short form. The Junior EPQ–R attempts to measure the constructs neuroticism, extroversion, and psychoticism. Francis (1996) cited two studies pertaining to the parent short form: Both studies addressed only the relationship between religiosity and personality as assessed by the parent short form. Those two studies do not show sufficient construct validity for Corulla's (1990) short form. Thus, Francis (1996) introduced a short form of an instrument that has not been shown to measure what it purports to measure. There is no justification for abbreviating an instrument that has yet to be adequately validated. Although it may be possible that the new short form would prove to have validity that has not been demonstrated on its parent short form, we found no new short forms whose validity exceeded that of the instruments from which they were derived. Of course, this concern is all the clearer when viewed from the standpoint of the consumer of the research. If a parent measure is insufficiently

Table 1
Articles in Which Short Forms Were Developed

Reference	Short form of scale
Adan & Almirall (1991)	Horne & Osterberg Morningness–Eveningness Questionnaire
Donders (1997)	Wechsler Intelligence Scale for Children—Third Edition
Francis (1996)	Revised Junior Eysenck Personality Questionnaire
Francis, Brown, & Philipchalk (1992)	Revised Eysenck Personality Questionnaire
Klepsch, Zaworka, Hand, Lunenschloss, & Jauernig (1991)	Hamburg Obsession/Compulsion Inventory
LaFreniere & Dumas (1996)	Social Competence and Behavior Evaluation
Mueser, Bellack, & Wade (1992)	Camberwell Family Interview
Recklitis, Yap, & Noam (1995)	Adolescent Version of the Defense Mechanisms Inventory
Santor & Coyne (1997)	Center for Epidemiologic Studies—Depression Scale
Sher, Wood, Crews, & Vandiver (1995)	Tridimensional Personality Questionnaire
Soldz, Budman, Demby, & Merry (1995)	Inventory of Interpersonal Problems Circumplex Scales
Whitley (1991)	Expanded Attributional Style Questionnaire

validated, then a psychologist will not (or should not) use it. Why, then, would a psychologist take an interest in an abbreviated version of such an instrument?

Sin 2: Fail to Show That Your Short Form Preserves the Content Coverage of Each Factor in the Measure

In any multifactor measure, it is essential to show the validity of each factor individually. When constructing a short form of such a measure, it must be shown that each factor is reliable and valid. Think of constructing multiple, related short forms: Every principle of short-form development ought to be applied at the individual factor level.

An obvious issue when evaluating short forms has to do with the content coverage of the target behavior or construct. Psychometric theory holds that each item of a test, or each behavior sampled, is itself an example of the target content domain. Thus, the assessment as a whole represents a sample of the true target behavior or construct. The degree to which the sample represents the target well is, in part, a function of the size of the sample. When the sample is reduced by making a short form, how well the assessment represents the target construct is also reduced.

Although reducing samples of items or observations can be necessary, the researcher has a burden to show that the target content domain is being adequately represented given such a reduction. However, only 3 of the 12 studies we reviewed even undertook content-domain checks. (One study used a classic criterion-keying method, which by principle eschews content considerations: For that study, content-domain checks would have been inconsistent with their approach.)

In lieu of doing content evaluations, many authors of questionnaires chose those items with the highest item–total correlations for a given factor (see Francis, 1996; Recklitis, Yap, & Noam, 1995; Whitley, 1991). This is a very popular approach. By choosing those items, the best chance of preserving a high internal-consistency estimate of reliability exists. Also, those items with the most error variance attached may be eliminated and, it is hoped, a purer measure of the target construct may result. These are good arguments. However, there is one potentially negative effect on the validity of an author's questionnaire by relying exclusively on item–total correlations for short-form item selection.

At issue is the breadth of the construct trying to be measured. A measure with an average interitem correlation of .50 is thought to be less broad than a measure that has an average interitem correlation of .30, and more broad than a measure with an average interitem correlation of .70. Items reflecting a broad construct will, on average, correlate less highly with each other than will items reflecting a narrow, more tightly defined construct, because each item can only represent a smaller portion of the broad construct. (We have argued elsewhere that once constructs are very broad, the presence of subfactors should be investigated; Smith & McCarthy, 1995; but for these purposes we assume there is only one construct being measured.) If a researcher has a measure with an average interitem correlation of .50 and chooses for a short form only those items with the highest item–total correlations, then the researcher will have chosen those items with the highest average interitem correlations by definition. The result may be that the content domain being measured has been inadvertently narrowed. By removing from consideration those items with lower item–total correlations (or lower interitem correlations), some part of the construct's domain may have been systematically omitted. The implication of this is that, although the reliability estimate remains high, the validity of the short form has been lessened: A reliable measure of a more narrow construct has been created.

If items are chosen solely on the basis of item–total correlations, then it is not known whether error variance associated with the least valid items has simply been removed or whether the construct being measured has been narrowed. In a recent meta-analysis, McCarthy and Smith (1996) found that length of alcohol expectancy measure accounted for 18% of the variance in effect sizes, independent of reliability. In other words, the short measures correlated substantially lower with criteria even though they were just as reliable. This finding can occur when the shorter measure does not sample the full content domain of the target construct.

To avoid the uncertainty about the impact of removing items, we recommend that investigators engage in a thorough content analysis along with, for questionnaires, examination of item–total correlations. Investigators should carefully describe the content domain of the original test or behavioral assessment procedure, use multiple judges to assess content preservation, quantify those judgments using a clear scaling procedure, and ensure that items or tasks in the original measure are represented proportionally in the short form (Haynes, Richard, & Kubany, 1995). The results of this analysis should then be reported as part of the validation process for the short form (Haynes et al., 1995). A careful content examination combined, where appropriate, with review of item–total correlations should give investigators the opportunity to remove the weakest items without unduly sacrificing content coverage. Alternatively, this approach would at least make it possible for researchers to clarify precisely which content they are omitting, which is, of course, also helpful. In this latter situation, a slightly different, more narrow construct has been really defined: This change should be made clear in the definition and labeling of the short form.

Sin 3: Fail to Show That Your Short Form Measures Each Factor Scale Reliably

There are different types of reliability that are differentially important depending on the purpose of the assessment. For many psychological tests, some form of internal consistency reliability is crucial: It is essential to show that each item is an indicator of the same construct. Interrater agreement is important for most behavioral-observation assessments. There are also many times in which stability over time is the primary consideration. Developers of short forms should make clear the type of reliability that is most germane, as should any test developer. We discuss reliability in a general way here, because most of our comments apply to the broad concept.

Our review of recent short-form publications yielded mixed results on this dimension. On the one hand, all 12 of the studies calculated and reported reliability coefficients. On the other hand, 5 of the 12 studies reported at least one reliability coefficient of .65 or lower: We considered any coefficient below .70 to be inadequate. Obviously, reduced reliability is a likely consequence of reducing test length or number of observations. Although poor reliabilities are certainly more likely with short forms, it is still true that when reliability equals .65, then 35% of the variance on that measure is random or error (Nunnally & Bernstein, 1994). In other words, as understandable as it may be, it is still a significant measurement problem. We believe that researchers and clinicians should expect strong reliability in short forms, just as they expect in full-length forms. Clinical decisions will be made from the short form, and therefore the short form has to be reliable. If it is impossible to construct a short form of a given construct with strong reliability, then the short form should not be constructed.

Sin 4: Fail to Show That Your Short Form Has Adequate Overlapping Variance With the Full Form, Using Independent Administrations

Showing adequate overlapping variance between a short and a full form is an important part of demonstrating the validity of the short form. It can be thought of, roughly, as an example of alternate forms reliability: The short form is akin to an alternate form of the full-length measure, and a strong correlation between the two forms is essential to support an argument that the short form is valid. In this section, we first review standard psychometric theory regarding correlations between shorter and longer measures of the same construct, and we then discuss the most common errors investigators make when they correlate short and long forms.

Consider a hypothetical example of a psychological test of 30 items, for which the average interitem correlation is .20. If we assume that all 30 items are drawn from the same content domain, then the internal consistency reliability of this test is given by Equation 1, where *n* refers to the number of items, *r*(*ij*) refers to the average interitem correlation, and *r*(*kk*) refers to the reliability:

$$r(kk) = \frac{n \times r(ij)}{1 + (n-1)r(ij)} \tag{1}$$

or

$$r(kk) = \frac{30 \times .20}{1 + (29 \times .20)} = .88$$

Now, if we shorten this measure to 10 items, again with an average interitem correlation of .20, the estimate of reliability becomes

$$r(kk) = \frac{10 \times .20}{1 + (9 \times .20)} = .71.$$

Clearly, this is a substantial loss of reliability. Assume that although short and long forms may yield different observed scores, each form does estimate the same true score for a given person (the assumption of essential tau equivalence; Lord & Novick, 1968, p. 50). We can then go on to estimate the correlation between the short form and the full-length form, with *r*(*fs*) reflecting the correlation, *r*(*ss*) reflecting the reliability of the short form, and *r*(*ff*) reflecting the reliability of the full form with Equation 2:

$$r(fs) = r(ss)r(ff) \tag{2}$$

or

$$r(fs) = (.71 \times .88) = .79.$$

These formulas are taken from Nunnally and Bernstein (1994).

Investigators can apply these formulas to obtain an a priori estimate of both the reliability of an intended short form and the overlap between the short form and the full-length version, and they can modify their short-form plans as a function of what outcome is predicted and what outcome is desirable. This estimation procedure, however, is no substitute for an empirical test. For example, the equations rest on the assumptions that each item is drawn from the same content domain reflecting the same set of true scores and that the short-form items will have the same level of interrelatedness as do the items in the long form. The first assumption is often inadequately tested, in part because of a failure to conduct thorough content analyses of chosen items. The second assumption is routinely violated, albeit with the best of intentions. As noted above, researchers frequently choose those items with the highest item–total correlations, and hence the highest average interitem correlations, for their short forms. When that is done without also considering items based on a content validity assessment, Equations 1 and 2 no longer apply, because the assumption that the two pools of items reflect the same content domain is no longer tenable: There is good reason to think that the short form reflects a more narrow domain. Most typically, such short forms will have higher reliability than what Equation 1 would predict, but quite likely a lower correlation with the full form than what Equation 2 would predict.

We advocate that investigators always conduct empirical tests of the overlap between short and long forms. However, there is a methodological sin that is so routinely committed that it may be becoming almost standard practice. The sin is to calculate the correlation between the short and the long forms on the basis of one test administration. Investigators frequently give the long form to one sample, score the short form from that same administration, and then correlate the short and long forms and report that correlation as an index of overlap between the two forms.

This approach is a methodological sin because it will, by definition, lead to an overestimate of the correlation between the two forms. All of the responses to the items in the short form are being counted twice: They are on both sides of the correlation. Thus, any error or random variance in the responses to any of the short-form items is, of course, completely reproduced in the long form. When taking this approach, one is in fact partially correlating error to itself. What is more, if there are any systematic error effects on item responses, perhaps due to the influence of neighboring items, those systematic effects will be present in both the short and the long forms. Clearly, those effects will not be present, or will be different, when the short form is used by itself. Correlating short and long forms from one administration may be a useful pilot step, but it is not part of rigorous validation of a short form.

An alternative approach that is sometimes considered is the following: Assume we have our 30-item measure and its 10-item short form. To avoid correlating error variance, we correlate the 10-item short form with the other 20 items of the long form. This approach is not a solution. Here, clearly taking steps away from answering the original question of overlap between short and long form is occurring. On the one hand, we are no longer relating our short form to its parent, and thus we may underestimate the overlap. On the other hand, we do not know how systematic error effects, from neighboring items and the like, may be influencing the correlation. Of course, if we have taken the approach of choosing for the short form the items with the highest interitem correlations, then we would then be correlating those items with the items with the lowest interitem correlations; this would also likely attenuate the correlation. The odds that we would end up correlating two slightly different constructs are high.

The solution to this problem is straightforward. Both the short form and the full-length form should be administered, separately, to the same participants. With proper instructions, and in some cases filler questionnaires between the two forms, this often can be done during one testing session. If this is judged inappropriate, then the two forms should be administered on two different occasions. Of course, in the latter case, test–retest factors are combined with the general overlapping variance concern. In such a case, the correlation between the short and long forms should be compared with the test–retest correlation of the long form. This approach is the only way to avoid correlating error with itself, and the only way to correlate the short form as it will be administered with the full-length form. We believe this strategy provides the best estimate of the overlap between the two forms. In our review of the 12 short-form articles, none performed the appropriate procedure. The field is regularly getting biased estimates of correlations between short and long forms. This bias is relatively easy to limit: Give both versions to participants.

Sin 5: Fail to Show Empirically That Your Short Form Reproduces the Factor Structure of a Multifactoral Instrument

These comments assume an investigator's goal is to preserve the original factor structure. We discuss the issue in correlational and factor analysis terms, but the same principles apply to any multidimensional assessment. There are two common approaches to this issue that are flawed. The first approach is to choose those items with the highest correlations with the grand, overall score. If items are chosen on the basis of their correlation with the overall score, and not based on their correlation with the appropriate individual factor score, then the retained set of items is unlikely to represent each factor sufficiently or evenly. If the retained items are then used to calculate a score for each original factor, then there is a good chance that some of the factors will not be represented well. This outcome is especially likely if not all factors correlate to the same degree with the overall score, which is more the rule than the exception. The result is a short form that lacks validity for some of its factor domains. In this typical strategy, a new factor analysis is not conducted on the short form, relying instead on the factor analysis of the full-length form. This decision causes the added problem that the factor structure of the short form may actually be different from the one presumed from the long form.

Another option is to follow this item-selection strategy but conduct a new factor analysis on the short-form items. A likely outcome here is to obtain a slightly different factor structure than the one obtained with the long form. When that happens, a slightly different set of constructs is being measured. When the difference between the short and long forms is substantial, it is incorrect to label the new measure as a short form of the original measure. Doing so obscures important differences in the likely validity domain of the new measure.

The second erroneous approach is to choose items, perhaps on the basis of item–total correlations, separately by factor, but then fail to verify the factor structure of the retained items. In this approach, it is assumed that because items representing each factor are retained, the original factor structure is still valid. One possible problem with making such an assumption relates to our earlier discussion of content validity. If each factor is represented with fewer items, then the content domain represented by some factors may have been reduced to a level that alters the meaning of the factor. If that occurs, then both the factor structure and the validity of the factors may be different for the short form.

Here again, there is a straightforward methodological solution. Investigators should select items separately by factor, using both item–factor scale correlations and content validity considerations; give the short form to a new sample; and conduct a factor analysis of the short form. Then the degree to which the content of the original factors has been preserved, as well as the degree to which the factor structure has been preserved, can be described. In most circumstances, because a defined factor structure is being replicated, confirmatory factor analysis is the appropriate choice. Confirmatory methods provide indices of how well the factor structure fits the hypothesized structure, leaving readers in a strong position to evaluate the extent of the short form's fidelity to the original test. Only 2 of 12 short-form articles described factor analyses on the short forms, and none reported factor-level content analyses. There are measures for which certain factors have not proven valid, or are not valid for the intended purpose of the short form. In such cases, eliminating the factors is obviously indicated.

Sin 6: If Your Short Form Omits Subfactors and Preserves Only Overall Factors, Then Fail to Show That The Short Form Preserves the Content Domains Represented by the Subfactors

Some measures represent constructs at different levels of hierarchical organization. For example, the NEO Personality Inventory—Revised (NEO-PI–R) has five broad factors: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Each factor can be subdivided into six facets. For example, the broad factor Extraversion or one facet of Extraversion, such as Gregariousness, can be studied.

When constructing a short form of such a measure, researchers sometimes seek only to preserve the broad factors, reasoning that specific facet-level analysis will be impossible with an abbreviated measure. The NEO-Five-Factor Inventory (NEO-FFI) is a short form of the NEO-PI–R that does exactly that: It preserves the five broad factors but not the subfacets (Costa & McCrae, 1992). This short form is not part of our systematic review, because it is described in a test manual, and we reviewed only peer-reviewed journal publications. Nevertheless, it is a useful example of developing a short form of an instrument that measures constructs at different hierarchical levels. None of the reviewed articles sought to preserve only overall factors.

The strategy Costa and McCrae (1992) chose was to select, for each broad factor, the 12 items with the highest loadings on that factor. After preliminary analyses, Costa and McCrae made 10 item substitutions: These were made on both rational and empirical grounds. Among the rational grounds was the goal of diversifying item content (Costa & McCrae, 1992). The issue we discussed above concerning preservation of content domains is important for evaluating this strategy. In the long form, each factor is represented by six subfactors, or facets. However, in the short form, no systematic effort was made to ensure that the same six facets remain proportionally represented. Although one may accept this decision on the grounds that facet-level interpretation will be impossible anyway, there remains the concern that the short-form broad factors may actually represent different, more narrow constructs than do the long-form broad factors. When items are selected on the basis of their factor loadings, or, similarly, item–total correlations, the facets that correlate most highly with the broad factor are likely to be overrepresented, and the facets that represent only a small portion of the broad factor are likely to be underrepresented. The result may well be a meaningfully more narrow factor. Costa and McCrae (1992) appeared to have recognized this issue to some extent, because they substituted items to diversify item content, but they described no effort to ensure that they have preserved the full diversity of factor content by representing each subdomain (facet) in the short form. Systematic content validity analyses would make it possible for readers to make some evaluation of the fidelity of the short form to its parent. To Costa and McCrae's (1992) credit, they do report validity analyses on the short form from samples other than the one from which it was derived. These findings are, of course, reassuring.

Sin 7: Fail to Show That Each Factor in the Short Form Has Validity on an Independent Sample

In this section, we discuss measurement of a single construct: The discussion applies to each factor represented in a short form. It is frequently assumed, and sometimes explicitly stated (Costa & McCrae, 1992; LaFreniere & Dumas, 1996), that once the evidence has been shown that the short form correlates highly with the long form, the validity evidence for the long form applies automatically to the short form. This assumption is overly optimistic. Frequently, short forms do not represent the full content domain of the long form, and hence they measure different sets of true scores. When this is true, a high correlation between forms does not guarantee the forms will have similar correlations with other measures.

However, if a short form is developed so carefully that the assumption can be made that the short form represents the same set of true scores as the long form, having proportionally sampled each facet of the target construct, psychometric theory does permit an estimate of the impact of item reduction on validity correlations (Nunnally & Bernstein, 1994). Consider this example, in which $r(fc)$ represents the correlation between the full form and a criterion, $r(ff)$ represents the reliability of the full form, $r(sc)$ represents the estimated correlation between the short form and the criterion, and $r(ss)$ represents the reliability of the short form. The estimated short-form correlation with the criterion is given by Equation 3 (Nunnally & Bernstein, 1994):

$$r(sc) = r(fc) \times \frac{r(ss)}{r(ff)} \tag{3}$$

Suppose $r(fc) = .50$; $r(ff) = .90$; and $r(ss) = .70$. Then,

$$r(sc) = \frac{(.50 \times .70)}{(.90)} = .39.$$

This estimation procedure can be used by investigators to anticipate the likely impact on validity of their item reduction, but only when the assumptions described above are met. Even when the assumptions are met, the estimation procedure is no substitute for an empirical evaluation of a short form's validity. When there is some doubt about whether the assumptions are met, there are many factors that limit the confidence with which we can infer that the parent form's validity likely translates to the short form:

1. As noted above, short forms are frequently developed from measures for which there is only limited evidence of validity. In such cases, potential problems with the validity of the long form have not been fully explored. When a short form is then derived, those potential problems certainly remain. Simply choosing the strongest validity coefficient and applying the above formulas will probably overestimate a short form's validity in this case, or at least obscure potential problems in the validity picture for the short form. We also believe there is a tendency for readers to forget, or not attend to, the limitations of the original measure once there is a short form: as if the existence of a short form implies that the parent measure must have been of the highest quality.

2. The most frequently used strategy for assessing the overlap between short and long forms (correlating the two from one, single administration) is biased, leading to an overestimate of the overlap. If the overlap is exaggerated, then the relevance of the long-form validity is also exaggerated.

3. As noted above, most short-form derivation approaches pay insufficient attention to preserving content coverage. To the extent that one's short form actually represents a slightly different, more narrow construct, the validity evidence for the long form has reduced value.

4. When one's short form is substantially less reliable than the parent form, one has, by definition, limited the validity coefficients of the short form relative to the long form (Nunnally & Bernstein, 1994). This effect is indicated in the above example, but recall that many short forms have scales with reliability coefficients of .65 or lower, indicating that the problem in many cases may be more severe than that described in our example.

For these reasons, the burden is on the short-form developer to show that the short form is a valid measure of the intended construct. The key empirical evidence should not be based on a sample in which the full, long form was administered. One should show that the short form, as it will be used, performs as hypothesized. In our review of the literature, we found that 6 of 12 studies reported some form of convergent validity evidence, 2 of 12 reported some form of discriminant validity evidence, and 4 of 12 reported criterion-related validity evidence of some kind. In all, only 4 of 12 tested the short form's validity on an independent sample (not a sample in which the full form was administered). We urge that future short-form developers include independent tests of the validity of the form they are deriving.

Sin 8: Fail to Show That Classification Rates Remain High With the Short Form

One product of many assessment procedures is some form of classification. Behavioral observations or test administrations may lead to the assignment of diagnoses, they may lead to inclusion in some target group, they may be a screening to identify individuals for further evaluation, or they may lead to classification of either the antecedents or the consequences of some target behavior.

One's goal with respect to classification depends in part on the intended use of the short form. If a short form is developed to conduct screening to identify those individuals who should be tested more comprehensively, the appropriate goal may well be to reduce false negatives (failure to identify a syndrome that is present) to near zero, even if that leads to a high number of false positives (identifying a syndrome that is not present). This way, virtually all at-risk individuals are assessed, and the accuracy of the full assessment will be depended on to identify the individuals who were false positives on the basis of screening. In such a case, the classification goals of the short form may well be different from those of its parent. Investigators should be clear in advance about such a goal, because it may influence choice of short-form screening items. This is another situation in which choice of short-form items is likely to be influenced by content considerations and not just item correlations with the full measure. Investigators should also make this choice and their rationale for it clear when introducing the new short form.

For many other short-form uses, the goal is to preserve the same accuracy of classification that is present in the parent form. When classification is an appropriate goal, researchers must investigate the accuracy of the new, briefer procedure just as they would investigate any aspect of a short form's validity: on an independent sample. Certainly it bodes well for a short form if the items, or the reduced behavioral sample, classify as well as the full-length version when the short form is extracted from the full-length version during one test administration. Nevertheless, there is no substitute for showing, with the short version administered as it would be administered in the future, successful classification on an independent sample. Two of four short-form articles investigated classification rates on the short form, and neither investigated them on an independent sample. Of course, just as with any index of validity, the likelihood is that the briefer measure will not classify participants quite as well as the parent measure. This fact requires two procedures from short-form developers: The first is, quite simply, to only abbreviate measures that have truly exceptional classification hit rates.

The second procedure is to estimate, in advance, the likely drop in classification accuracy one may encounter, so an a priori decision can be made about the merits of developing the short form. In the case in which one seeks to classify individuals between two outcomes, one can use the formulas given above to estimate the drop in correlation between the test and the outcome group membership. Assuming one is conducting a point–biserial correlation, using an interval scale variable to differentiate between two groups, one can go on to calculate approximate hit rates for a given correlation. The procedure and its assumptions, illustrated with a hypothetical example, are as follows.

Suppose we have developed and standardized a 30-item Police Officer Readiness Test (PORT), which we believe will differentiate between individuals chosen to become police officers and those turned down by the police force after a more extended, lengthy evaluation period. Suppose the PORT has a mean of 100 and a standard deviation of 15. To test the classification accuracy of my measure, we study a sample of 100 recent applicants, of whom 50 were accepted as officers and 50 were turned down. The officers have a mean score of 109, and the rejected applicants have a mean score of 91. The point–biserial correlation in this sample can be calculated (Cohen & Cohen, 1983, p. 38) where *r*(pbis) is the point biserial correlation, Y(p) is the mean score for police officers, Y(a) is the mean score for rejected applicants, *p* is the probability of being a police officer, and *q* is the probability of being a rejected applicant, and *SD*(y) is the standard deviation of the PORT:

$$r(\text{pbis}) = \frac{[\text{Y}(\text{p}) - \text{Y}(\text{a})]pq}{SD(\text{y})} = \frac{(109 - 91 \times .5 \times .5)}{15} = .60.$$

Because we have an equal number of participants in each group, and because we assume each distribution is normal, an obvious cutting score will be the point where the two curves cross, which is halfway between the two means (Cliff, 1987, p. 406). In this case, that point is a score of 100. This cutting score will have a *z* score of .60 with respect to the rejected applicant mean of 91, and a *z* score of −.60 with respect to the officer mean of 109 (see Cohen & Cohen, 1983, p. 39). The use of *z* tables leads to the proportional classification table given in the top half of Table 2.

Suppose we reduce our 30-item test to 10 items, retaining those items with the highest item–total correlations. We anticipate that we might have reduced content coverage with this short form, therefore we anticipate a point–biserial correlation of .40 between the short form and group membership. Using the process just described, we give the estimate of proportional classification accuracy in the bottom half of Table 2.

An investigator may use such an estimation to decide that the short form is likely to perform adequately, or an investigator may decide that retention of more items is essential if some necessary classification accuracy is to be achieved. When one plans to differentiate between groups of vastly different sizes, the estimation procedure becomes more complicated and is beyond the scope of this presentation (see Cliff, 1987). Again, this a priori estimation procedure is not a substitute for an independent, empirical test of a short form's accuracy.

Sin 9: Fail to Show That Your Short Form Offers Meaningful Time or Resource Savings for the Loss in Validity

The obvious rationale for developing a short form is to save valuable time or important resources, and there is clearly a trade-off between assessment time and validity. It is essential that short-form authors address this trade-off directly. This can be done

Table 2
Hypothetical Classification Rates as a Function of Item Reduction

Test	Rejected applicant	Officer
Full 30-item		
Predicted officer	.27	.73
Predicted rejected	.73	.27
Reduced 10-item		
Predicted officer	.35	.65
Predicted rejected	.65	.35

Note. Top half of Table 2 gives hypothetical classification rates, using a test where *r*(pbis) = .60 between the test and acceptance or rejection as a police officer. Bottom half of Table 2 gives hypothetical classification rates when the test is reduced from 30 items to 10 items, and *r*(pbis) is reduced to .40 between the test and acceptance or rejection as a police officer.

in an a priori way using psychometric theory. Suppose we have a 60-item test with internal consistency reliability of .90. On average, it takes 30 min to complete, or 30 s per item. We consider reducing our test to 40 items, thus saving close to 10 min. We can apply the Spearman–Brown formula (Nunnally & Bernstein, 1994) to estimate the reliability of the short form, thus,

$$r(\mathrm{ss}) = \frac{\mathrm{k}\; r(\mathrm{ff})}{1 + (\mathrm{k} - 1)r(\mathrm{ff})}, \qquad (4)$$

where r(ss) is the reliability of the short form, r(ff) is the reliability of the long form, and k is the length of the short form divided by the length of the long form. We would estimate the reliability of the 40-item short form to be

$$\frac{2/3(.90)}{1 + (2/3 - 1 \times .90)} = .86.$$

Using Equation 3 and assuming a long-form validity coefficient of .50, we estimate the validity coefficient of our short form to be .49. In almost all cases, this would represent a trivial loss of validity for the savings of 10-min assessment time. We could then consider a reduction to 20 items. Application of Equations 3 and 4 yield the following estimates: 20 item short-form reliability = .75; short-form validity coefficient = .46. In this case, a time savings of nearly 20 min gives an estimated loss of shared variance with a criterion of 4%. Depending on the purpose of the short form, such a loss might be either acceptable or unacceptable. As short-form developers, we can then decide how much time it is reasonable to try to save. Of course, once we develop the short form, we must then conduct empirical tests of the reliability and the validity of the new measure.

It is crucial to note that this example depends on development of the short forms using the procedures described in this article. The above formulas do not apply unless each facet of the original content domain is proportionately sampled, and the average correlation among items is the same as for the long form. Once again, these estimations are likely to be optimistic with any procedure that fails to preserve the content domain of the long form.

In our review, none of the articles offered an a priori evaluation of the savings versus validity trade-off. Four of the 12 articles did address the issue, leaving readers in a position to evaluate the persuasiveness of the authors' contentions. For the other 8 articles, the reader is not in a position to evaluate the issue. Investigators should show empirically and rationally that their proposed short form offers meaningful savings as compared with the reduction in validity. What is meaningful depends on the assessment context and the purpose of the short form.

Item Response Theory (IRT) and Tailored Testing

IRT is an approach to test development that is not based on classic true score theory, but which is quite useful when the goal is brief, precise assessment. There are a number of excellent resources describing IRT (cf. Hambleton, Swaminathan, & Rogers, 1991; Mellenbergh, 1996): We describe the procedure very briefly. With a development or standardization sample, for each item a curve that describes the ability or intensity level at which the item maximally discriminates can be defined. These item characteristic curves (ICCs) are often S-shaped and the item maximally discriminates where the slope is the steepest (with ability or intensity level on the *x*-axis). Items differ with respect to the attribute intensity level at which they discriminate. For example, a Neuroticism item "I tend to worry about my job when the company announces that 20% of us will be layed off next month," will probably only differentiate among people with very little neuroticism. The item, "Although management has never engaged in layoffs, I cannot help worrying that I could lose my job that way," probably differentiates among individuals much higher on neuroticism. Thus, for every item in a pool, curves can be developed that describe the ability or intensity range at which the item will be maximally useful. ICCs can then be used in a tailored testing approach. One begins with some estimate of intensity of an attribute (perhaps the standardization sample average intensity level) and administers an item that maximally discriminates at that level. If the examinee endorses or passes the item, then an item that discriminates at a higher intensity level is administered. If the examinee does not endorse or fails the first item, then an item that discriminates at a lower intensity level is administered instead. With the examinee's response to each new item, more information is available to estimate the examinee's intensity level of the attribute. Statistically, once the examinee has endorsed at least one item and failed to endorse at least one item, the examinee's attribute intensity can be formally estimated. Administering items is then continued, leading to modifications in the estimate of attribute intensity, until the changes in the estimates from one item to the next become trivially small. At that point, one has a reliable (reproducable) estimate of attribute intensity for that examinee. It appears that reliable estimates can be obtained frequently with less than half of the number of items used to measure the same attribute using true-score theory (Hambleton et al., 1991). Computers make tailored testing fairly simple. Assuming a sound item-development procedure, this approach can lead to shorter assessments without all of the methodological problems that are often evident with true-score theory-based short-form construction.

Summary and Recommendations

Thirty years ago, Levy (1968) reviewed short forms and concluded that many essential elements of the methodological task of constructing short forms had been forgotten. Our memories have not improved in the ensuing 3 decades. Short forms are continually constructed with such methodological weaknesses that it is tempting to argue for a halt to the process. We have taken a different view: Short forms constructed with appropriate rigor, absent the methodological sins currently plaguing the short-form literature, may well prove to have a valuable place in clinical assessment. We have recommended the following steps to sin-free short-form development. These steps are most relevant for either classic, true-score theory approaches or behavioral-assessment approaches. Tailored testing using IRT is a promising alternative approach for many assessment purposes.

A Priori Steps in Rigorous Short-Form Development

1. Ensure that the parent measure has, itself, been sufficiently validated for the intended purpose.

2. Clarify the intended use of the short form (e.g., screening measure vs. diagnosis measure) and choose items or observation classes that meet the goal.

3. Compute an a priori estimate of the short form's reliability.

4. Compute an a priori estimate of the likely overlap between the short form and its parent.

5. Compute an a priori estimate of validity correlations of your short form with key criteria.

6. Where appropriate, compute an a priori estimate of the classification accuracy (whether hit rates or elimination of false negatives) of your short form.

7. Compute a priori estimates of the time saved and the validity lost.

Taking each of these steps in advance can help an investigator to choose the best balance between time–resource savings and loss of validity. One can certainly try out a number of different short-form lengths and choose the one most likely to meet the intended need.

Methodological Steps During Short-Form Validation

1. Show time or resource savings and their relation to loss of validity, empirically.

2. Conduct and describe content analyses of each factor in the measure to preserve as much content coverage as possible. Describe clearly any decisions to eliminate a content domain from the parent measure.

3. Administer the short form on an independent sample, to enact Steps 4 through 8.

4. Show that each factor meets reasonable reliability standards.

5. Calculate an estimate of overlapping variance between the two versions. As noted above, extracting the short form from the full form during one administration will overestimate the overlap, which could lead to unanticipated validity problems down the road. Correlating the short form with the unused, remaining items from the full form underestimates the overlap and is not recommended.

6. Demonstrate the factor structure or dimensionality of the short form. Make readers aware of any significant differences between the factor structures or dimensions of the short- and full-length versions.

7. Validate the short form. By definition, it is harder to validate a short measure, and the validity evidence for the full-length measure cannot be presumed. It is important to validate the short form in the form it will be used, rather than by extracting its items or observations from the full-length assessment.

8. Where appropriate, assess the short form's classification rates.

9. If subfacets that are included in the full-length form are omitted and only aggregate factors are kept, then content analyses must be conducted to show preservation of the meaning of the aggregate factors or to show that the aggregate factors represent a more narrow construct domain.

We believe that if these methodological guidelines are followed, then researchers will be in a strong position to argue that their short form is a reliable, valid alternative to a fuller, more comprehensive assessment.

References

References marked with an asterisk indicate the reviewed articles.

*Adan, A., & Almirall, H. (1991). Horne & Osterberg Morningness–Eveningness Questionnaire: A reduced scale. *Journal of Personality & Individual Differences, 12,* 241–253.

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.

Cliff, N. (1987). *Analyzing Multivariate Data.* New York: Harcourt Brace Jovanovich.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavior Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corulla, W. J. (1990). A revised version of the Psychoticism scale for children. *Personality & Individual Differences, 11,* 65–76.

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

Doll, E. A. (1917). A brief Binet–Simon scale. *Psychological Clinic, 11,* 197–211.

*Donders, J. (1997). A short form for the WISC–III for clinical use. *Psychological Assessment, 9,* 15–20.

*Francis, L. J. (1996). The development of an abbreviated form of the Revised Junior Eysenck Personality Questionnaire (JEPQR–A) among 13–15 year olds. *Journal of Personality & Individual Differences, 21,* 835–844.

*Francis, L. J., Brown, L. B., & Philipchalk, R. (1992). The development of an abbreviated form of the Revised Eysenck Personality Questionnaire (EPQR–A): Its use among students in England, Canada, the U.S.A. and Australia. *Journal of Personality & Individual Differences, 13,* 443–449.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (Vol. 2).* Newbury Park, CA: Sage.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

*Klepsch, R., Zaworka, W., Hand, I., Lunenschloss, K., & Jauernig, G. (1991). Derivation and validation of the Hamburg Obsession/Compulsion Inventory—Short Form (HOCI–S): First results. *Psychological Assessment, 3,* 196–201.

*LaFreniere, P. J., & Dumas, J. E. (1996). Social Competence and Behavior Evaluation in children ages 3 to 6 years: The short form (SCBE-30). *Psychological Assessment, 8,* 369–377.

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

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

McCarthy, D. M., & Smith, G. T. (1996, June). *A meta-analysis of alcohol expectancy research.* Paper presented at the annual meeting of the Research Society on Alcoholism, Washington, DC.

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

*Mueser, K. T., Bellack, A. S., & Wade, J. H. (1992). Validation of a short version of the Camberwell Family Interview. *Psychological Assessment, 4,* 524–529.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory.* New York: McGraw-Hill.

*Recklitis, C. J., Yap, L., & Noam, G. (1995). Development of a short form of the Adolescent version of the Defense Mechanisms Inventory. *Journal of Personality Assessment, 64,* 360–370.

*Santor, D. A., & Coyne, J. C. (1997). Shortening the CES–D to improve its ability to detect cases of depression. *Psychological Assessment, 9,* 233–243.

*Sher, K. J., Wood, M. D., Crews, T. M., & Vandiver, P. A. (1995). The Tridimensional Personality Questionnaire: Reliability and validity studies and derivation of a short form. *Psychological Assessment, 7,* 195–208.

Silverstein, A. B. (1990). Short forms of individual intelligence tests. *Psychological Assessment, 2,* 3–11.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

* Soldz, S., Budman, S., Demby, A., & Merry, J. (1995). A short form of the Inventory of Interpersonal Problems Circumplex Scales. *Assessment, 2,* 53–63.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence.* New York: Psychological Corporation.

* Whitley, B. E., Jr. (1991). A short form of the Expanded Attributional Style Questionnaire. *Journal of Personality Assessment, 56,* 365–369.

Received October 19, 1998
Revision received August 24, 1999
Accepted August 27, 1999 ■

AMERICAN PSYCHOLOGICAL ASSOCIATION
SUBSCRIPTION CLAIMS INFORMATION

Today's Date:__________________

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do **NOT** duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

____
PRINT FULL NAME OR KEY NAME OF INSTITUTION

____
MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL)

____
ADDRESS

____
DATE YOUR ORDER WAS MAILED (OR PHONED)

____

____PREPAID ____CHECK ____CHARGE
CHECK/CARD CLEARED DATE:________________

____
CITY STATE/COUNTRY ZIP

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

____
YOUR NAME AND PHONE NUMBER

ISSUES: ____MISSING ____DAMAGED

TITLE	VOLUME OR YEAR	NUMBER OR MONTH

Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4–6 weeks.

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: __________________
ACTION TAKEN: __________________
STAFF NAME: __________________

DATE OF ACTION: __________________
INV. NO. & DATE: __________________
LABEL NO. & DATE:__________________

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.