Creating Short Forms for Construct Measures: The role of exchangeable forms

Knut A. Hagtvet[1] and Kornel Sipos

Abstract: *A popular trend has invaded applied psychometrics in a broad range of social science, in particular in research fields of educational psychology, in terms of creating short forms for construct measures. There seems to be a paucity of developing methodologies for creating short forms based on complete forms that meet psychometric standards related to the reliability of scores and valid inferences. The present article suggests a methodology that rests on the fundamental assumption that the concept of a short form attains meaning when derived from valid scores of a complete form. A pivotal construct for assessing the status of a short form is the concept of exchangeable forms, which incorporates two types of measurement invariance; a) invariance across groups, frequently exercised in studies applying confirmatory factor analysis, and b) invariance across random facets, as estimated in generalizability theory. The two types of measurement invariance involve two types of generalizations relevant for inferring constructs; generalizing from a sample of persons to a population of persons, and generalizing from a sample of construct indicators to a universe or domain of construct indicators. In addition, structural invariance is required; exchangeable short forms should relate equivalently to external reference variables. The Hungarian version of the State-Trait Anxiety Inventory for Children (STAIC-H) was used to illustrate the suggested short form methodology.*

Keywords: *valid complete form; exchangeable forms; generalizing to population of persons and universe of construct indicators; two types of measurement invariance; structural invariance.*

During recent years there has been an increasing tendency to create short forms for construct measures. This tendency has been driven by the need to reduce the burden on respondents to work through lengthy complete forms. Shortening measures may also allow more scales to be administered within a given period of time and thereby as many constructs as possible to be measured within the given time span. Some researchers argue that many scales may be redundant and should therefore be shortened. It may also be noted that a preset time span may also preclude the use of the full form of the instrument.

[1] Corresponding author: Knut A. Hagtvet; k.a.hagtvet@psykologi.uio.no. The authors are grateful to Professor Emeritus Ellen J. Hartman, University of Oslo, for guiding us to English translations of the quotation from Henrik Ibsen used in the present article.

DOI: 10.14712/23362189.2016.346

One may easily understand the reserved attitude to lengthy complete forms with regard to redundant items, response burden, and fewer constructs measured in a constrained time span, among other factors. However, do we pay a price for creating short forms?

This trend can easily be observed within a broad range of social sciences, in particular in education and educational psychology. An ERIC online search in late October 2016 using the key words "short form of academic aptitude test" returned 7823 references, of which 2311 were articles in refereed journals. When the alternative keyword "performance tests" was used, the corresponding returns were 21,588 and 7630, respectively. Irrespective of the validity of this search, the interest in short forms in educational research is considerable.

The present paper is confined to developing a methodology for creating and assessing short forms within the frameworks of classical test theory, confirmatory factor analysis/structural equation modelling, and generalizability theory, which are frequently applied in educational measurement (Bollen, 1989; Brennan, 2001a; Cardinet, Tourneur, & Allal, 1981; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability theory is, however, not commonly applied for assessing short forms. Other methodologies may also be applied, such as IRT. However, because of the mathematical and conceptual inconsistencies between our methodological framework and IRT, as recently discussed by Brennan (2001a, 2004), including IRT and related methodologies would easily go beyond the scope of the present paper.

The status of analytical approaches to selecting items for short scales

Among the most frequently used approaches to create short forms are: a) items with the largest item-total (remainder) correlations, b) items with the largest item discrimination parameters as estimated by the IRT methodology, c) items with large factor loadings on the focal factor and small factor loadings on the other factors, as estimated by exploratory factor analysis, and d) items with the largest factor loadings, smallest cross-loadings, and uncorrelated error variables as estimated by confirmatory factor analysis. Modification indices have been used to identify cross-loadings and correlated error variables.

It should be noted that these approaches appear mainly to have been applied on an unelaborated empirical basis. Often, the use of these methodologies is never or rarely justified. Their actual practice falls short of ideal or even reasonable standards. The short forms created by these approaches are often ad hoc or one-shot endeavours (Marsh, Ellis, Parada, Richards, & Heubeck, 2005). Their applications are not often soundly based on theory or not systematically evaluated, revised, or improved. This state of affairs has created a need not only to adhere to an adequate methodology for creating complete forms, but also to create short forms that are adequately based on the complete forms. These two objectives may be considered as two ways to fulfill the same purpose, as will be elaborated below.

The need to assess the methodological state of affairs of short forms has also been echoed in methodologically-oriented research journals. Psychological Assessment invited three researchers to elaborate the methodological challenges raised by this development. This invitation resulted in the article "On the sins of short-form development" (Smith, McCarthy, & Anderson, 2000). Smith et al. concluded at that time that "Short forms are continuously constructed with such methodological weaknesses that it is tempting to argue for a halt to the process" (p. 109). Smith et al. suggested a set of recommendations to follow when creating short forms. Short forms are still being created by means of ad hoc procedures or according to suboptimal methodological standards (Marsh et al., 2005; Widaman, Little, Preacher, & Sawalani, 2011). There seems to be a paucity of developing methodologies for creating short forms based on complete scales that meet psychometric standards related to the reliability and validity of scores. One noticeable exception is the elaborate methodology worked out by Marsh et al. (2005). Their methodology is mainly based on the perspective of confirmatory factor analysis. They assessed the recommendations suggested by Smith et al. (2000) and mostly agreed with their viewpoints, although a few disagreements were noted.

The present paper presents a methodology for creating and assessing short forms that are generally based on equivalent principles to those defended by Smith et al. (2000) and Marsh et al. (2005). However, we extend the current methodology for creating short forms by requiring measurement equivalence both a) across groups (Marsh et al., 2005; Vandenberg & Lance, 2000) as well as b) across samples of construct indicators as accomplished within the framework of generalizability theory.

Basic considerations of the concept of short forms

There is strong agreement that a fundamental focus of psychometrics is to study the nature of a domain or a universe of indicators on the basis of a sample of indicators from the domain or universe (Brennan, 2001a; Cornfield & Tukey, 1956; Cronbach, Rajaratnam, & Gleser, 1963; Guttman, 1953; Kaiser & Michael, 1975; Lord & Novick, 1968; McDonald, 1999; Nunnally & Bernstein, 1994; Tryon, 1957). Thus, a basic issue appears to be that of generalizing from a sample of indicators to the construct domain or universe of construct indicators. The test score itself is not the actual focus of attention, but is considered a platform in order to make a valid inference about the construct. This applies to a short form too. The emphasis on the universe or construct domain is most clearly expressed by Nunnally and Bernstein (1994), "…there is no way to know how to test the adequacy with which a construct is measured without a well-specified domain" (p. 88). By emphasizing the importance of a clear definition of a universe, it follows that the universe is as least as important than the test. We suggest that the universe or domain has logical priority over the test itself. This requirement is rarely taken into account when short forms are being created.

In line with this way of thinking, a complete form should be valid in the sense of representing the construct domain. This statement is considered to be a fundamental assumption when conceptualizing and creating short forms. If the complete form cannot be assumed to be valid, deriving short forms will not make much sense. Short forms need to be based on a sound conceptual footing or point of departure. If this fundamental assumption can be questioned, the inadequacies in the complete form will be transmitted to the short forms that are derived. The well-known quotation from the dramatic poem Peer Gynt by Henrik Ibsen (1867) expresses the basic idea:

> *"…where the starting-point is mostly fatal, the outcome is often highly original."*[2]

A short form may then be conceptualized as a representative or random selection of indicators from a valid complete form. It follows that the short form that is selected should be equivalent to or exchangeable (Shavelson & Webb, 1981) with other representative or randomly selected sets of indicators from the same valid complete form. Thus a short form does not exist in terms of being the one and only short form unless its existence is equivalent to, or exchangeable with, other selected sets of indicators (short forms) or the remaining items in the complete form. This notion is another associated fundamental assumption of the present methodology. If the short form is not exchangeable with other sets of indicators or the remaining set of indicators in the complete form, the conceptual status of both the short and complete form can be questioned. Even though the assumption of exchangeability is critically important for making sense of the concept of a short form and consequently for the entire process of creating short forms, this assumption is, however, not commonly justified in the applied research literature. In line with this practice it is commonly observed that the remaining set of indicators in the complete form is considered less valid and therefore does not attract any more attention. If the short form cannot be considered exchangeable with other selected sets of indicators, inferences to a latent construct are impaired.

[2] Different translations into English exist for the Norwegian wording of this quotation; "…hvor utgangspunktet er galest, blir tidt resultatet originalest" (Ibsen, 1867). Our preferred translation is a combination of our own translation (first line) and the adopted second line from Northam (1995, p. 109), which, we suggest, expresses the basic idea most strikingly for the present context.

Two types of studies

Internal domain studies

In line with the basic considerations outlined above, a first step in creating short forms would be to focus on **internal** domain studies. Emphasis is placed on internal domain studies before **external reference** studies are conducted. This recommendation is based on both conceptual and data-analytical considerations (Jøreskog, 1993; Anderson & Gerbing, 1988). Internal domain studies should involve how well the actual constructs are being measured before they are related to variables external to the construct domain.

Individual differences should be generalized across different forms, as well as different samples of persons. The individual differences of a construct should not be specific to one particular set of items or short form or to one peculiar or particular sample of persons. In other words, a short form should be generalizable across different sets of conceptually relevant sets of indicators and across different samples of persons. Testing measurement equivalence across groups is commonly performed in applied research. The psychometric concern of estimating the generalizability of short forms, however, is rarely, if at all, seen in applied research, in particular when short forms are applied. These concerns suggest that both statistical and psychometric inferences (Mulaik, 1972) within the framework of internal domain studies are required to construct and assess short forms when aiming at the construct validity of the scores.

An impending concern is that short forms can potentially represent a peculiar set of items as a result of their abbreviated nature. This abbreviated selection of items may not necessarily generalize across person samples. Many short forms appear too "test-specific" and not domain-oriented. Commonly, items in a complete form that are not selected to constitute a short form are either discarded or considered to be less valid, or do not attract any more attention. Such an approach does seem to oppose the fundamental assumption of the validity of the scores of a complete form. On the other hand, even when a short form may have been purposely selected to represent the complete form, it may happen that the factor structure of the complete form may not necessarily be replicated in the short form. Such an observation was reported in a large-scale assessment of PISA by Carstensen (2009). Other selections of short forms, however, were reported to replicate the factor structure, as expected in Carstensen's study.

If short forms cannot be considered conceptually equivalent or exchangeable, critical consequences may then be expected. For example, short forms selected by strong item-total (remainder) correlations derived from an invalid complete form will very probably produce a biased representation of the underlying construct. Therefore strong item-total (remainder) correlations, often routinely considered a proper method for selecting "good" items, may not necessarily suggest short forms with valid scores unless the scores of the complete form can be assumed valid. Likewise, applying factor analyses, the EFA or CFA of a questionable complete form may obtain models that adequately fit the data, but the items with the largest factor loadings may not necessarily provide a short form of the intended underlying construct. The notion of short forms as exchangeable subsets of items that all make up the total valid score of the complete form does not seem to have been explicitly addressed. One is reminded of the observation made by Cronbach et al. (1972) years ago that "Investigators often choose procedures for evaluating reliability that implicitly define a universe narrower than their substantive theory calls for." (p. 352). This procedure will underestimate the measurement error or inflate the reliability coefficient. This is a potential danger when constructing short forms. Often, two or three items constituting a short form are selected from a far larger complete form because they correlate substantially. The reliability of such a set of items is often reported to be .75 or far above. Such a reliability coefficient may indicate invalidity of the short form scores if the reduced set of items is an unacceptably narrow representation of the construct domain. The conceptual assumptions on which estimations are based are as important as the estimations themselves. In other words, even well-fitted models resting on dubious conceptual assumptions are not worth more than the assumptions themselves. The need to emphasize internal domain studies is also caused by the fact that researchers often go too quickly to external reference studies with unelaborated or poor measures.

External reference studies

A short form should not only meet the requirements of internal domain studies, but also be subjected to external reference studies where the short forms should relate equivalently to variables external to the internal domain, designated as *structural invariance* in the present exposition. As suggested by Smith et al. (2000), a high correlation between short forms does not guarantee that the forms will have similar correlations with other measures.

The present investigation

The present short form methodology will be illustrated by the 20-item trait anxiety scale of the Hungarian State-Trait Anxiety Inventory for Children (STAIC-H; Sipos & Sipos, 1979) as derived from the State-Trait Theory of Anxiety (Spielberger, 1972). The state anxiety scale of the STAIC-H and gender will be used as external reference variables in assessing short forms of trait anxiety. The trait-state conception of anxiety and its measures have played an extensive role in research in educational psychology and related fields to assess processes operating in a different variety of achievement contexts over several decades (Hagtvet, 1989; Zeidner, 1998), such as different forms of evaluation anxiety (examination stress, test anxiety, mathematical anxiety, anxiety in competitive situations) and evaluating applicants for demanding jobs and candidates for admission to highly competitive educational programmes or educational intervention programmes for the treatment of examination stress, among other scenarios. Short forms of the STAI have been created (Hanin & Spielberger, 1986; Marteau & Bekker, 1992; Sanderson, 1988) and used in research contexts related to those listed above.

The present assessment of short forms will be carried out by means of a series of ordered steps to assess the exchangeable status of short forms measuring latent constructs. It is believed that the present methodology is applicable to a large variety of latent constructs. The methodology will be demonstrated by means of the STAIC-H, as introduced above.

In the steps involving internal domain studies measurement equivalence will be assessed within both the framework of confirmatory factor analysis as well as generalizability theory. In line with the methodological considerations given above, the factorial validity of the present complete trait anxiety scale in a total person sample will be assessed in Step 1. On the basis of the State-Trait Theory of Anxiety, the trait anxiety scale is assumed to measure one unidimensional factor. Earlier studies (Dorr, 1981; Hedl & Papay, 1982) supported the one-factor interpretation of the trait anxiety scale. The very purpose with Step 1 is to assess the fundamental premise that the complete form of the trait scale supports validity. Also included in this purpose is the assessment of the content coverage of the construct. Even if a one-factor model fits the data, if the content coverage can be questioned, the validity of the one-factor interpretation may also be questioned.

Step 2 will assess the factorial validity of the complete form in two random samples of persons by performing an invariance analysis of the unidimensional factor structure. The two random samples were created by means of the SPSS software (IBM, 2013). If Step 1 provided support for validity, but the invariance test in Step 2 failed, validity may still be questioned because of the possible existence of person sample-specific evidence. The first two steps should focus on the complete form of the STAIC-H with respect to both conceptual and empirical support. If the two former steps have provided a valid basis, it appears reasonable to create short forms in the third step. In the present study two short forms were created in four different ways by selecting 10 items for each of the two forms from the complete 20-item form by: a) making random splits, b) selecting the first 10 and the last 10 items, respectively, c) selecting the 10 oddly and the 10 evenly numbered items, and d) selecting the "best" and the "worst" set of 10 items by item-remainder correlations. Assuming that the items in the complete form are all valid indicators of the same construct, each way of creating two short forms will, one assumes, create two conceptually exchangeable short forms. In Step 3 a two-factor model will be assessed separately in each of two random samples. This step will address whether: a) the unidimensional factor in the complete form is reproduced in each short form within each sample, b) the two short forms are closely related in each sample because they are assumed to be exchangeable forms by originating from the same parent form, and c) the factor variance is invariant across the short forms within each sample. If the factor model does not fit the data in one or both samples, the reason may be attributed peculiarities in a short form and/or a sample.

In Step 4 the invariance of factor loadings and measurement residual variances across the samples of persons will be tested, while the invariance of the factor variances across *both* forms and samples will also be assessed.

Related to the models in Steps 1–4, different generalizability analyses will be performed to indicate to what extent individual differences in trait anxiety are generalizable across short forms and/or items.

So far, the exchangeability of short forms has been considered within the framework of internal domain studies. It is of equal importance to examine if the different short forms relate equivalently to variables outside the internal domain to claim their status as short forms. Structural invariance is the focus of Steps 5 to 8. Steps 5 and 6 will examine whether the short forms relate equivalently to the external variables of present and absent state anxiety, while Steps 7 and 8 will perform equivalent assessment with respect to gender. Based on the State-Trait Anxiety theory, the short forms of trait anxiety should relate positively to the state anxiety variables.

Method

Subjects

The Hungarian version of the STAIC (Sipos & Sipos, 1979) was administered to 1580 students (605 boys and 975 girls) in 12 schools. The sample consisted of students from 10 to 15 years of age.

The Hungarian STAIC

The children's form of the A-trait of the STAIC applies a 3-point scale for each of its 20 items (1= rarely; 2=/sometimes; 3=/often). The State Anxiety scale consisted of 20 items; 10 negatively and 10 positively worded items. These items also applied a 3-point scale.

The STAIC-H was administered under standard conditions. When one or two item responses were missing in the scale, their scores were replaced according to instructions provided in the STAIC manual (Spielberger, 1973).

Statistical estimation

Because of the three scoring points applied to all items, the statistics were estimated by ordered categorical variable estimation (Jøreskog, 2005) by means of "robust maximum likelihood" (Jøreskog & Sørbom, 2013; Satorra & Bentler, 1988). When estimating latent variables based on observed ordered categorical outcome variables by means of LISREL9.20 (Jøreskog & Sørbom, 2013), four types of parameters were estimated in the present internal domain studies; factor loadings (FL), factor variances (FV), factor covariances (FCov), and measurement error variances (EV). In the present external reference studies regression effects (RE) were estimated as well.

Reliability/generalizability was estimated by the relative generalizability coefficient, E_{ρ}^{2}, (Brennan, 2001a) and the coefficient omega – ω (McDonald, 1999). Because of the ordered categorical indicators of the STAIC-H, the parameters of the generalizability coefficient E_{ρ}^{2} as well as omega were estimated by means of the robust maximum likelihood available in LISREL9.20. The model applied for estimating variance components was described by Jøreskog (1978), while different related applications within the framework of generalizability theory are reported by Marcoulides (1996) and Hagtvet (1998). Omega is a reliability coefficient of a set of items fitting the general factor in a person (p) by item (i), p x i design, and will be estimated for this design in Steps 1–4. Assuming homogeneity or unidimensionality, omega may be called the coefficient of generalizability from a given item set to the domain/universe (McDonald, 1999). E_{ρ}^{2} is estimated on the basis of different random effects two-facet measurement designs implicit in Steps 2–4 in Table 1. For Step 2 the persons within samples (s) by items, (p : s) x i, design will be applied in order to estimate generalizability across items *within* a single randomly selected group and persons *over* groups, respectively. Generalizing over both items and forms will be assessed by the person by items within forms (f), p x (i : f), design in Steps 3A and 3B. This type of generalization will also be estimated in Step 4 but will now be based on the (p : s) x (i : f), design for persons *within* a single randomly selected group and persons *over* groups, respectively. The estimation formulas are presented in the Appendix. For detailed descriptions of the different designs, the reader is referred to Shavelson and Webb (1991).

Two types of inference concerning measurement equivalence

Measurement equivalence can be approached by two types of inference of equal importance. One type is typically obtained within the framework of classical confirmatory factor analysis, where the measurement equivalence of the parameters is commonly specified by constraining the parameters to be invariant across groups, occasions, or points in time (Marsh et al., 2005; Vandenberg & Lance, 2000). Generalizations are commonly made from a sample to the population of persons and measurement invariance is inferred across (typically fixed) groups, occasions, or points in time.

A different type of inference that is applied more rarely involves generalizing from a sample or set of indicators to the construct domain or universe of indicators. This type of inference is typically made within the framework of generalizability theory (Brennan, 2001a; Cronbach et al., 1972). Generalizing from a sample of indicators to the construct domain is critically important for making inferences about the construct in question, and in particular when assessing short forms, because of their abbreviated nature.

Results

The matrices of the parameters for assessing measurement and structural equivalence are summarized in Table 1 in the order of the analytical steps outlined above. The present short form methodology will be illustrated in detail by applying the short forms created by a random split through all eight steps shown in Table 1. The remaining three types of short form will be briefly commented on in the Discussion section on the basis of their results from the internal domain studies only.

Internal domain studies

Table 1 should be read row-wise. The subscripts of the matrices indicate short forms. The same matrix *without* a sub-

Table 1. Assessing Invariance of Complete and Short Forms

	Internal domain studies			
	Complete form (N=1580)			
Step 1[1)]:	FL FV EV			
	Sample A (N=790)		Sample B (N=790)	
Step 2:	FL FV EV		FL FV EV	
	Short forms			
	F1	F2	F1	F2
Step 3A:	FL_1 FV EV_1	FL_2 FV EV_2		
Step 3B:			FL_1 FV EV_1	FL_2 FV EV_2
Step 4:	FL_1 FV EV_1	FL_2 FV EV_2	FL_1 FV EV_1	FL_2 FV EV_2
	External reference studies			
	Anxiety State factors			
Step 5A:	FL_1 FV EV_1	FL_2 FV EV_2		
Step 5B:			FL_1 FV EV_1	FL_2 FV EV_2
Step 6:	FL_1 FV EV_1	FL_2 FV EV_2	FL_1 FV EV_1	FL_2 FV EV_2
	Gender effects			
Step 7A:	FL_1 FCov FV EV_1	FL_2 FCov FV EV_2		
Step 7B:			FL_1 FCov FV EV_1	FL_2 FCov FV EV_2
Step 8:	FL_1 FCov FV EV_1	FL_2 FCov FV EV_2	FL_1 FCov FV EV_1	FL_2 FCov FV EV_2

Note. FL = factor loadings; FV = factor variance(s); FCov = factor covariance; EV = measurement error variances; RE = regression effects; FL1 = factor loadings for short form F1; See text how to read the table.

script in the same line indicates invariant parameters of that matrix across short forms and/or samples of persons. Repeated identical subscripts for the same matrix in the same line indicate invariant parameters of that matrix. For example, in Step 4 the factor loadings of form 1, FL_1, are assumed to be invariant across the samples A and B. Likewise, measurement error variances of items of F_1, EV_1, are assumed to be invariant across the samples A and B. The factor variance, FV, is assumed to be invariant across the forms and samples in Step 4. A matrix with identical designation including subscripts located in *different* lines is not an invariant or identical matrix.

Step1: The fit of the CFA model of the 20 items of the complete trait anxiety scale in the total sample of N=1580 was χ^2_{170}=666.33; RMSEA= .059 (.056; .062); CFI=.976; TLI=.973). On the basis of the p x i design for the complete form, the generalizability estimate of omega was .87. The standardized factor loadings are presented in Table 2.

As is shown in the line for the Step 2 factor loadings, factor variances and error variances were constrained to be invariant across the two random samples of N=790 for the complete form of the trait anxiety scale. The fit of the multi-sample model was χ^2_{380}=969.54; RMSEA = .064 (.061;.067); CFI=.971; TLI=.971. Omega for the complete form, based on the invariant matrices, FL and EV, assuming the random effects p x i design, was .87. The common metric completely standardized invariant factor loadings across the two person samples are reported in Table 2. On the basis of the (p : s) x i design of Step 2, the estimated generalizability coefficient, E_ρ^2, was .823 for persons both within and across samples. This estimate indicates that the scores representing the individual differences of the complete form of trait anxiety generalize well to the construct domain. (See the Appendix for details). Steps 1 and 2 have then provided

Table 2. Factor loadings, FL, for the complete form in the total sample and two random samples A and B

	N=1580	A (N=790) B (N=790)
		Invariant loadings[1]
T1	.585	.585
T2	.484	.484
T3	.571	.572
T4	.509	.509
T5	.429	.428
T6	.542	.542
T7	.524	.523
T8	.539	.536
T9	.568	.569
T10	.525	.524
T11	.511	.511
T12	.467	.468
T13	.538	.540
T14	.454	.454
T15	.366	.366
T16	.318	.319
T17	.558	.557
T18	.444	.445
T19	.566	.566
T20	.577	.578

Note. The table results correspond to the estimated models in lines for Step 1 and 2, respectively, in Table 1. [1]*Common metric completely standardized factor loadings.*

ample evidence for supporting the validity of the scores of the complete form of the STAIC-H.

In Step 3 the factor loadings and error variances were obviously specific to the two short forms in each of the two samples, while the factor variances, FV, were constrained to be equal across the short forms within each of the two samples (Sample A: FV_1=FV_2=FV=.081; Sample B: FV_1=FV_2=FV=.067). The covariance between the factors was specific to each sample (Sample A: $FCov_{12}$=.081; Sample B: $FCov_{12}$=.067). On the basis of these estimates it can be shown that the correlation between the short forms was 1.0 in both samples. The fit measures that were obtained for Samples A and B were χ^2_{170}=436.05; RMSEA=.065 (.060; .070); CFI=.975; TLI=.972) and χ^2_{170}=472.24; RMSEA=.067 (.063; .072); CFI=.969; TLI=.965, respectively. The omega coefficients for the p x i design in each short form were ω_{F1}=.75 and ω_{F2}=.80 in Sample A and ω_{F1}=.76 and ω_{F2}=.78 in Sample B. The generalizability coefficients $E_\rho^{\,2}$ based on the implicit p x (i : f) design in Step 3 for Samples A and B were estimated to be .87 and .86, respectively. Further details are provided in the Appendix. Both the factor analyses and the reported generalizability coefficients supported short forms being exchangeable.

To assess the invariance of the *correspondence* between the short forms across random person samples in the two-sample analysis in Step 4 the factor loadings, error variances, and factor covariances were constrained to be invariant across the samples, while the factor variances were constrained to be invariant across *both* the forms and samples (Table 3). (Sample A and B: FV_1=FV_2=FV=.074; FCov=.075; χ^2_{380}=969.65; RMSEA=.064 (.061; .068); CFI=.971; TLI=.971). Omega was estimated for the invariant short forms, F1 and F2, across the samples; ω_{F1}=.75; ω_{F2}=.79. It can be shown that the estimated correlation between the short forms across samples is 1.0.

Table 3. Invariant[1] factor loadings, FL, across samples A and B

Short form			
F1		F2	
T13	.454	T1	.613
T12	.430	T10	.517
T15	.350	T18	.363
T8	.544	T9	.596
T19	.556	T4	.451
T14	.447	T5	.425
T16	.339	T20	.630
T3	.548	T17	.540
T7	.513	T2	.459
T6	.570	T11	.507

Note. [1] Common metric completely standardized factor loadings. FL corresponds to invariance constraints in line for Step 4 in Table 1.

The estimated generalizability coefficients, $E_\rho^{\,2}$, for persons within a random sample and persons over samples, respectively, for the implicit random effects (p : s) x (i : f) design was basically the same: .824. Further details are provided in the Appendix. The results reported so far for the internal domain studies provided ample evidence for supporting the exchangeability of short forms based on both types of inference.

External reference studies

Anxiety State factors

In Steps 5 and 6 state anxiety factors were included as external variables. A confirmatory factor analysis, based on the total sample size (N=1580), supported a two-factor model indicating a state anxiety present and a state anxiety absent factor, respectively (χ^2_{134}=324.91; RMSEA=.095 (.092;.091); CFI=.992; TLI=.993).[3] The correlation between the two factors was .32.

In Steps 5A and 5B the factor covariance between the two short forms of trait anxiety on the one hand, and the respective anxiety state factors on the other, were constrained to be invariant across forms within each sample of persons. The factor loadings and item error variances were specific to each short form. The short form variances, FV_1 and FV_2, were constrained to be invariant across short forms in each person sample, while the factor covariance, FCov, between forms was specific to each sample. Sample A: FV_1=FV_2=FV=.08; $FCov_{12}$=.08; Sample B: FV_1=FV_2=FV=.067; $FCov_{12}$=.068). The model showing the relation between the two short trait anxiety forms and the two state anxiety factors in each of two random samples of N=790 persons provided a reasonable fit to the data; Sample A: χ^2_{662}=1171.38; RMSEA=.086 (.084; .088); CFI=.986; TLI=.986; Sample B: χ^2_{662}=1249.64; RMSEA=.090 (.087;.092); CFI=.981; TLI=.979. The covariance between the forms on the one hand and the anxiety state factors on the other was constrained to be invariant across the forms within each sample (standardized parameters in parentheses); Sample A: $FCov_{F1,\ ABS}$=$FCov_{F2,\ ABS}$= .042 (.495); $FCov_{F1,\ PRE}$=$FCov_{F2,\ PRE}$=.110 (.591); Sample B: $FCov_{F1,\ ABS}$=$FCov_{F2,\ ABS}$=.032 (.446); $FCov_{F1,\ PRE}$=$FCov_{F2,\ PRE}$=.095 (.531).

The invariance of the *correspondence* between short forms and anxiety state factors was assessed by means of the model specifications in Step 6. For this purpose factor variances for the short forms were constrained to be invariant across forms and samples. Measurement error variances and factor loadings for the short forms and anxiety state factors were invariant across the samples. The covariance between short forms, on the one hand, and anxiety state factors, on the other hand, was constrained to be equal across the short forms and samples. The fit of this restricted multi-sample model was χ^2_{1399}=2543.38; RMSEA=.087 (.085; .088); CFI=.983; TLI=.983). The support for the invariant factor loadings and relationships between short forms and anxiety state factors is reported in Table 4.

[3] Items ST1 and ST15 were deleted to improve the model fit. The RMSEA does not satisfy conventional criteria, while TLI and CFI indicated an excellent fit. The two factors were, nevertheless, considered to serve as external variables for the present purposes.

Table 4. Invariant[1] factor laodings, FL, across samples A and B

Short form			
F1		F2	
T13	.458	T1	.624
T12	.469	T10	.519
T15	.359	T18	.443
T8	.527	T9	.565
T19	.566	T4	.510
T14	.449	T5	.419
T16	.308	T20	.578
T3	.579	T17	.560
T7	.516	T2	.483
T6	.536	T11	.521
Invariant relationships between short forms and anxiety state factors across samples A and B			
		F1	F2
FCov:	ABS	.473	.473
	PRE	.560	.560

Note. [1] *Common metric completely standardized factor loadings. FL and FCov correspond to invariance constraints in line for Step 6 in Table 1.*

Gender effects

To expand the assessment of the two forms of trait anxiety, it is of interest to examine the invariance of the gender effect on the two short forms.

Steps 7A and 7B focus on the invariance of the gender effect across the two forms within each random sample, respectively. Given the corresponding set of invariance assumptions of the short forms to that in Step 5 above, the invariant gender effect across the short forms was estimated to be (standardized parameters in parentheses) .165 (ρ=.284; p<.001) in sample A and .119 (ρ=.226; p<.001) in sample B. These findings suggest that the girls report higher trait anxiety scores than the boys for both short forms in both samples. The fit of the model was acceptable in both samples; Sample A: χ^2_{189}=555.25; RMSEA=.070 (.065; .074); CFI=.967; TLI=.964; Sample B: χ^2_{189}=606.97; RMSEA=.073 (.068; .077); CFI=.959; TLI=.954.

In Step 8 the invariance of the gender effect across *both* forms and person samples

Table 5. Invariant[1] factor loadings, FL, across samples A and B

Short form			
F1		F2	
T13	.456	T1	.621
T12	.470	T10	.523
T15	.360	T18	.443
T8	.538	T9	.566
T19	.564	T4	.504
T14	.453	T5	.427
T16	.312	T20	.579
T3	.571	T17	.552
T7	.520	T2	.495
T6	.540	T11	.506
Invariant effects of gender on short forms across sample A and B			
Gender			
RE:	F1	.257	
	F2	.257	

Note. [1] *Common metric completely standardized factor loadings. RE corresponds to invariant regression effects in line for Step 8 in Table 1.*

was tested under assumptions corresponding to those in Step 6. The fit of the model was χ^2_{419}=1221.80; RMSEA= .069 (.066; .072); CFI=.962; TLI=.962. The invariant factor loadings and gender effects are reported in Table 5, which provides support for the invariant gender effect on the two short trait anxiety forms; .257 (p<.001). Thus the short forms displayed equivalent relationships to the applied external variables of the presence and absence of state anxiety and gender.

Discussion

The concept of exchangeable short forms was defined and empirically assessed by means of eight ordered steps within two perspectives of measurement equivalence. The fundamental assumption for deriving short forms was supported in Steps 1 and 2, where the one-factor model for the complete form of trait anxiety fitted the data well in both the total sample and with regard to the invariance of the one-factor model across two random samples. Steps 3 and 4 of the internal domain studies supported a two-factor model in single randomly selected samples and factorial invariance across samples. Furthermore, exchangeable short forms were supported by generalizability analyses based on different measurement designs associated with each step. Omega provided ample evidence for generalizing the complete and short forms to the domain/universe based on the one-facet p x i design in Steps 1–4. The generalizability coefficient E_ρ^2 was estimated on the basis of the different two-facet random designs associated with Steps 2–4. Both omega and E_ρ^2 serve an important purpose in the present study. Because short forms are generally more narrowly defined than the complete form because of their reduced number of items, it is therefore important to assess their reliability/generalizability. As reported above, the omega coefficient for the complete form was .87 in Steps 1 and 2, while the omega for the different short forms varied from .75 to .80 in Steps 3 and 4, thus still staying within the acceptable range.

A related and important question is how well scores on individual differences of the actual construct will generalize across both items and forms. Since we are considering short forms to be exchangeable measures of the same construct, individual differences should generalize not only across items but also forms. This concern was assessed by applying different implicit measurement designs in Steps 3 and 4. In Steps 3A and 3B generalizability was assessed by applying the p x (i : f) design. The estimated generalizability in samples A and B was .870 and .802, respectively, indicating that generalizing across items and forms was strongly supported. An associated estimation of generalizability was based on the

(p : s) x (i : f) design in Step 4 when person *within* a single randomly selected sample and person *over* samples were alternatively applied as the objects of measurement. It turned out that the estimated generalizability coefficient was .824, no matter which objects of measurement were applied. The estimate that was obtained indicates that individual differences are convincingly generalized across forms and items.

The short forms were also supported by structural invariance, indicated by their invariant relationships to state anxiety factors and gender as external variables, respectively. In sum, the results of the present eight-step procedure supported the notion of exchangeable short forms of trait anxiety within both internal domain studies and external reference studies.

The concept of exchangeable forms

The notion of invariant properties of measurement parameters is well known from the extensive evaluation of measurement invariance across existing fixed groups or time within the framework of confirmatory factor analysis, as exhaustively demonstrated and elaborated by Vandenberg and Lance (2000) and illustrated within the research area of short forms by Marsh et al. (2005). What is far less known and recognized is how generalizability analysis can favourably add to the assessment of the measurement invariance of short forms by emphasizing a different perspective. Validating the scores of short forms may take advantage of both perspectives on measurement invariance.

Exchangeable short forms should exhibit invariant measurement properties that would justify substituting one form, A, for another form, B. In the present assessment of short forms involving confirmatory factor analyses ample evidence was generated with respect to invariant factor loadings, strong factor covariances, and measurement residual variances across random samples of persons. Additionally, short form factor variances were invariant across both samples and short forms. This pattern of invariant matrices may be characterized as strong measurement invariance. There may be situations, however, in which measurement residuals should be allowed to vary across samples, while keeping the remaining matrices invariant.

From the perspective of generalizability theory, short forms should support invariance laws associated with the construct in mind, as elaborated by Kane (2002). In the present context of short forms, which, one assumes, measure the same homogeneous trait anxiety construct, it would be expected that the D-study variance components for short forms and interactions involving short forms should exhibit small values which support the assumption of invariance over short forms in the universe of generalization. On the other hand, large values for the D-study variance components for short forms would provide strong evidence against invariance across short forms. The analyses of Steps 3 and 4 reported small values for D-study variance components for short forms in the two-facet designs or equivalently large estimates of the generalizability coefficients (see the Appendix). In this way generalizability coefficients indicated strong support for invariance properties for short forms. In other words, the rank order of individual differences did not change noticeably from one short form to another. This invariance property indicates that the measured construct does not vary across forms. This type of invariance facilitates the inference drawn from an actual sample of indicators to a larger universe or a domain of trait anxiety indicators. Thus, the invariant property of exchangeable forms facilitates universe score interpretations. On the other hand, in the case of a short form with an exchangeable status that is not known or not conceptually and empirically justified, the score interpretation of the trait anxiety measure would probably be impaired or biased. This would probably occur if only one short form has been selected from a complete form whose validity could be questioned. Because the exchangeable status of forms created from the perspective of generalizability theory does not seem to have been addressed in most studies of short forms, a relevant validity inference is missing in these studies. Generally speaking, short forms may suffer from insufficient validity inferences.

Methodological challenges

The present methodology for creating short forms is by no means exhaustive. The methodology was illustrated by an available source of data. The present data provided an opportunity to present an assessment procedure of both types of measurement invariance. However, issues still remain before an exhaustive assessment has been accomplished.

Both analytic traditions rest upon random sampling which is rarely satisfied in both traditions. This issue is beyond the scope of the present article. However, suffice it to say that the issue has been discussed over the years and some convenient sampling strategies have been offered. The reader is referred to writings by Brennan (2001a), Cronbach et al. (1972), Cornfield and Tukey (1956), Lord and Novick (1968), and Shavelson and Webb (1981), among others. More recently, Kane (2002) discussed this issue from the perspective of generalizability theory. He provided four general guidelines for the conduct of G-studies which may also partly be applied to studies based on CFA. Instead of relying on random sampling, he suggests applying "representative" sampling in the sense that selective forces that might bias the results have been identified and controlled.

a) In line with many others, he suggested that the universe of admissible observations or the domain of content should be defined with clarity. The present application of STAIC-H has been adapted from the Spielberger (1973) STAIC, which is conceptually founded on the unidimensional construct of trait anxiety (Spielberger, 1972). Factor analyses of the trait anxiety scale of the STAIC, including the present CFA results, have provided consistent support for a unidimensional or a homogeneous factor. These findings provide support for assuming that the actual different items represent the same construct domain. The present study includes three additional ways of selecting two forms from the complete form, which were carried out by i) selecting the first

10 and the last 10 items, respectively; ii) selecting the 10 oddly and the 10 evenly numbered items and, iii) selecting the " best" and the "worst" set of 10 items by item-remainder correlations. Since sampling from a homogeneous complete form is immaterial because all sets of the same size are supposed to be essentially the same, we therefore expected that each set of two short forms should fit a two-factor model representing two exchangeable forms. Each of the three two-factor models estimated in each of two random samples of persons was acceptably fitted to the data.[4] This may be interpreted as meaning that each set of the two forms created, including the present randomly selected forms, represents two exchangeable forms that each mirror the complete form. Consequently, all these forms represent the same intended underlying homogeneous construct domain. Generally speaking, by selecting two forms of equal size in different ways from the same unidimensional complete form, the assessment of the short forms will be more comprehensive.

b) Another concern involves the existence of representative samples of persons from the population to the extent that they are free from identifiable sources of selection bias. The present person sample included gender, 12 schools, and six age groups. A MIMIC model (Muthén, 1989) was estimated for examining measurement invariance across subgroups of schools and ages, coupled with a multigroup analysis assessing factorial invariance across gender (Hagtvet & Sipos, 2004). The findings suggested a high and acceptable degree of measurement invariance for the STAIC-H. The subgroups did not support any threat against a valid representation of the trait anxiety construct in the present population of students measured by the complete form. We do not know about additional facets of measurement that might represent potential biasing factors affecting generalization from the sample of short forms to the intended construct domain.

c) To check for potential idiosyncratic selection bias, replication in independent samples is especially important. A replication was, in principle, carried out in the present study by examining how the two randomly selected short forms behaved in two random samples of persons. However, the scores of randomly selected short forms may be affected by having been subjected to the same administration. Strong correlation between two exchangeable forms or low person by short form component would be expected. However, the present estimated correlation of unity between the two forms in Steps 3 and 4 is very probably an overestimate.

A related methodological challenge is to estimate the overlapping variance between the short form and the complete form. From a validity viewpoint it is important to show that the short form and the complete form do correlate substantially (Smith et al., 2000). The basic issue is to show that the short and complete

[4] Not reported in the present paper.

form not only correlate, but share equivalent psychometric properties (Smith et al., 2000; Marsh et al., 2005). This psychometric concern appears to have a solution within the present methodology that rests on the notion of exchangeable forms. Creating short forms by randomly or arbitrarily selecting items from the complete form should a) prevent idiosyncratic selection bias, and b) automatically create two equivalent short forms consisting of the selected set and the remaining set of items. If the two forms do correlate substantially and display invariant factor models, support for the reproduction of the psychometric properties of the complete form in each of the selected short forms has been established. The results derived from Steps 3 and 4 support these inferences.

However, the potential overestimate of the correlation between the two forms and the convergence of factor models may partly be caused by the present measurement design, which allowed all items or both forms to be administered in the same test occasion. By allowing the two short forms, however, to be administered in different test occasions separated by a sufficient timespan, unbiased findings may be expected.

d) Finally, Kane reminded us about the ultimate concern to provide assurance that the error in estimating the universe score is not so large as to invalidate the inferences and decisions being made in a D-study. It should be remembered that the D-study variance components in the present study were relatively small compared to the size of the universe score variance components. This was especially noticeable for the person by form components, which were practically zero (see the Appendix). By having included an occasion facet with independent administrations as conditions, a less biased estimation of the generalizability coefficients may have been obtained. However, hidden or implicit facets may always be a potential threat to unbiased estimation. With this precaution in mind, we may assume that the present estimations are adequate for their intended use.

Concluding comments

The present short form methodology has focused on the notion of exchangeable forms. This concept has been linked to two types of measurement invariance. In the tradition of covariance structure analysis and structural equation modelling persons are commonly treated as random, while the possibility of a random facet of construct indicators does not seem to have been considered (Brennan, 1992, 2001a). In the framework of generalizability theory both persons and indicators can be treated as random. This feature allows generalization from random samples of indicators to universes of indicators for persons randomly sampled from a population of persons. The invariance notion in CFA allows a factor model to be generalized to a population of persons, while generalizability theory allows invariance laws associated with facets of measurement to be assessed. In this way the concept of exchangeable forms includes both types of generalization. This bilateral meaning of invariance does seem to be consistent with the very notion of a *construct*, in which the intended inference or generalization goes beyond the actual sample of persons and short form.

This way of conceptualizing and performing assessments of short forms appears to contrast with mainstream approaches, which are restricted to generalizing to a population of persons only. As noted in the introductory part of this paper, it is commonly observed that the remaining set of indicators in the complete form after the selection of only one set of items to constitute the short form is considered less valid and therefore does not attract any more attention. Such an approach may often imply that conceptually relevant indicators are deleted in order to improve the fit of the factor model. Alternatively, the remaining set of indicators may have been given a conceptually relevant status at the time it was included in the original complete form. The reason for its exclusion may be linked to different circumstances, such as ambiguous item formulations, underrepresenting other factors in the domain, or narrowing the construct domain. Generally speaking, their exclusion may be caused by an insufficient description of the domain. If these considerations reflect common ways of creating short forms, the only form selected may not attain the status of an exchangeable form.

The choice of exchangeable forms as a unifying notion for defining, creating, and assessing short forms should remind us of the challenges that will be faced when aspiring to make inferences to a population of persons and a domain of construct indicators. As stated in the introductory section, most approaches to the creation of short forms have often been applied on an unelaborated empirical basis and not subjected to adequate methodological standards. To create short forms does not seem to be a "quick fix" procedure.

The present methodology mistrusts observed scores. The intended inference goes beyond the observed scores to properties in both a population of persons, as well as a universe or domain of indicators. No matter how necessary and desirable the inference is supposed to be, it does not come into being without paying a price. In generalizability theory random sampling of items or short forms is a fundamental property but nevertheless untestable. Likewise, a clear definition of the construct domain does not seem to be easily attainable. The domain of many constructs is considered fuzzy, which makes the full meaning of the constructs unsure (Nunnally & Bernstein, 1994). Strictly parallel forms as required in classical test theory represent an ideal situation. However, as in much scientific work, the researcher needs to rely on untestable assumptions. The solution to such challenges is to apply reasonable approximations to the ideal assumptions. Kane (2002) considered representative sampling as a promising vehicle to approach random sampling. Shavelson and Webb (1981) considered exchangeability a reasonable approximation to random sampling. Random sampling was considered by Lord and Novick (1968), Brennan (1992, 2001a), and McDonald (1991) to be a useful idealization of situations encountered by actual educational and psychological measurement operations. Approximations are then approached by a commitment to a comprehensive set of issues, including careful conceptual consideration and a reasonable choice of measurement operations and data analytical models that, all together, would facilitate validity inferences. The price to pay for the temptation to follow a short cut in creating short forms is very probably less valid construct inferences.

The trait anxiety scale of the STAIC-H has been applied to illustrate different requirements in order to attain the status of a short form. The fundamental assumption in attaining the status of a short form is a complete form from which valid construct inferences can be drawn. Ample evidence was provided for the complete form of the trait anxiety scale of the STAIC-H to be considered as a dependable starting point for creating exchangeable short forms.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411–423.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brennan, R. L. (1992). *Elements of generalizability theory*. Revised edition. Iowa City: American College Testing.

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

Brennan, R. L. (2001b). *Manual for urGENOVA. Occasional papers no. 49*. Iowa Testing Programs: University of Iowa.

Brennan, R. L. (2004). *Some perspectives on inconsistencies among measurement models. CASMA*. Research Report No. 8. University of Iowa, Iowa City: CASMA.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its application in educational measurement. *Journal of Educational Measurement, 18*, 183–204.

Carstensen, C. H. (2009). Linking PISA competencies over three cycles – results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA* (pp. 199–213). Dordrecht, Netherlands: Springer.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27*, 907–949.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137–163.

Dorr, D. (1981). Factor Structure of the State-Trait Anxiety Inventory for Children. *Personality and Individual Differences, 2*, 113–117.

Guttman, L. (1953) A special review of Harold Gulliksen: Theory of mental tests. *Psychometrika, 18*, 123–130.

Hagtvet, K. A. (1989). *The construct of test anxiety. Conceptual and methodological issues.* Bergen/London: Sigma Forlag/Jessica Kingsley Publishers.

Hagtvet, K. A. (1998). Assessment of latent constructs: a joint application of generalizability theory and covariance modeling with an emphasis on inference and structure. *Scandinavian Journal of Educational Research, 42*, 41–63.

Hagtvet, K. A., & Sipos, K. (2004). Measuring anxiety by ordered categorical items in data with subgroup structure: The case of the Hungarian version of the trait anxiety scale of the State-Trait Anxiety Inventory (STAIC-H). *Anxiety, Stress, and Coping, 17*, 49–67

Hanin, Y. L., & Spielberger, C. D. (1986). The development and validation of the Russian form of the State-Trait Anxiety Inventory. In C. D. Spielberger & R. Diaz-Guerrero (Eds.), *Cross-cultural anxiety* (Vol. 2, pp.15–23). Washington, DC: Hemisphere.

Hedl, J. J. & Papay, J. P. (1982). The factor structure of the State-Trait Anxiety Inventory for Children: kindergarten through the fourth grades. *Personality and Individual Differences, 3*, 439–446.

IBM Corp. Released 2013. *IBM SPSS Statistics for Windows. Version 22.0.* Armonk, NY: IBM Corp.

Ibsen, H. (1867). *Peer Gynt. Et dramatisk digt.* København: Gyldendalske Boghandel.

Jøreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43*, 443–477.

Jøreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. Scott Long (Eds.), *Testing structural equation models* (294–316). Newbury Park, CA: Sage.

Jøreskog, K. G. (2005). *Structural equation modeling with ordinal variables using LISREL*. Revised version 10 February, 2005. http://www.ssicentral.com/lisrel/corner.htm

Jøreskog, K. G., & Sørbom, D. (2012). *Some new features in LISREL9*. Document available in Jøreskog and Sørbom (2013). Chicago, IL: Scientific Software International

Jøreskog, K. G., & Sørbom, D. (2013). *LISREL (Version 9.10)* [Computer software]. Chicago, IL: Scientific Software International.

Kaiser, H. F., & Michael, W. B. (1975). Domain validity and generalizability. *Educational and Psychological Measurement, 35*, 31–35.

Kane, M. (2002). Inferences about variance components and reliability – generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement, 39,* 165–181.

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

Marcoulides, G. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling, 3*, 290–299.

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17,* 81–102.

Marteau, T. M., & Bekker, H. (1992) The development of a six-item short form of the state scale of the Spielberger State-Trait Scale Anxiety Inventory (STAI). *British Journal of Clinical Psychology, 31*, 301–305.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Muthén, B. O. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Northam, J. (1995). *Henrik Ibsen. Peer Gynt. A dramatic Poem*. Oslo, Norway: Scandinavian University Press.

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

Sanderson, F. H. (1988). Analysis of anxiety levels in sport. In D. Hackfort & C. D. Spielberger (Eds.), *Anxiety in sport: An international perspective* (Chap. 4). Washington, DC: Hemisphere.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308–313.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133–166.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizabilty theory. A primer.* Newbury Park: Sage.

Sipos, K.. & Sipos, M. (1979). The development and validation of the Hungarian form of the State-Trait Anxiety Inventory for Children (STAIC-H). *Magyar Pediater, 13*(6), 47.

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

Spielberger, Ch. D. (1972). *Anxiety: Current trends in theory and research* (481–493). New York: Academic Press.

Spielberger, Ch. D. (1973). *STAIC Preliminary Manual for the State-Trait Anxiety Inventory for Children*. Palo Alto, CA: Consulting Psychologist Press.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*, 229–249.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011), On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, M. B. Donnellan, & Lucas, R. (Eds.), *Secondary data analysis* (pp. 39–61). Washington, DC: American Psychological Association.

Zeidner, M. (1998). *Test anxiety; The state of the art.* New York: Plenum.

Knut A. Hagtvet,
Department of Psychology, Faculty of Social Sciences, University of Oslo, Norway

Kornel Sipos,
Department of Psychology, Faculty of Physical Education and Sport Sciences, Semmelweis University, Budapest, Hungary

Appendix

Generalizability coefficients and D-study variance components

Coefficient omega (ω) is estimated on the basis of factor loadings (λ_j) and measurement error variances (θ_j). Both the factor loadings and error variances for omega and the variance components entering the generalizability coefficients reported below are estimated by the robust maximum likelihood provided by LISREL9.20 for observed ordered categorical variables. In addition, variance components were estimated with the urGENOVA software (Brennan, 2001b) for four models, as noted below, with continuous observed variables being assumed.

$$\omega = \frac{(\Sigma\lambda_j)^2}{(\Sigma\lambda_j)^2 + \Sigma\theta_j}$$

Step 2: Random effects (p:s) x i design

The objects of measurement are persons within a single randomly selected group: p:s

$$E_\rho{}^2 = \frac{\sigma^2_{p:s}}{\sigma^2_{pi:s}/n_i + \sigma^2_{p:s}} = \frac{.1277}{.1438} .888 \text{ (.823 estimated by urGENOVA)}$$

D-study comp. (n_i = 20): .3211/20 + .1277 = .1438
.0161 + .1277 = .1438

The objects of measurement are persons *over* group: s and p.s (Estimated by urGENOVA)

$$E_\rho{}^2 = \frac{\sigma^2_s + \sigma^2_{p:s}}{\sigma^2_{pi:s}/n_i + \sigma^2_{si}/n_i + \sigma^2_s + \sigma^2_{p:s}} = \frac{0 + .0748}{.0909} = .823$$

D-study comp. (n_i = 20): .3210/20 + .0001/20 + 0 + .0748 = .0909
.0161 + .0 + 0 + .0748 = .0909

Step 3: Random effects p x (i:f) design
Sample A:
The objects of measurement are persons, p.

$$E_{\rho}^{2} = \frac{\sigma^{2}_{p}}{\sigma^{2}_{pi:f}/n_{i}\,n_{f} \;+\; \sigma^{2}_{pf}/\,n_{f} \;+\; \sigma^{2}_{p}} = \frac{.1156}{.1329} = .870$$

D-study comp. (n_i=10; n_f=2): 3458/10x2 + 0/2 + .1156 = .1329
.0173 + 0 + .1156 = .1329

Sample B:
The objects of measurement are persons, p.

$$E_{\rho}^{2} = \frac{\sigma^{2}_{p}}{\sigma^{2}_{pi:f}/n_{i}\,n_{f} + \sigma^{2}_{pf}/\,n_{f} + \sigma^{2}_{p}} = \frac{.1069}{1240} = .862$$

D-study comp. (n_i = 10; n_f = 2): .3412/10x2 + .0/2 + .1069 = .1240
.0171 + .0 + .1069 = .1240

Step 4: Random effects (p:s) x (i:f) design
The objects of measurement are persons within a single randomly selected group, p:s.

$$E_{\rho}^{2} = \frac{\sigma^{2}_{p:s}}{\sigma^{2}_{pi:sf}/n_{i}\,n_{f} \;+\; \sigma^{2}_{pf:s}/\,n_{f} \;+\; \sigma^{2}_{p:s}} = \frac{.1112}{.1284} = .866 \text{ (.824 estimated by urGENOVA)}$$

D-study comp. (n_i=10; n_f=2): .3435/10x2 + .0/2 + .1112 = .1284
.0172 + .0 + .1112 = .1284

The objects of measurement are persons *over* groups, s and p (Estimated by urGENOVA).

$$E_{\rho}^{2} = \frac{\sigma^{2}_{s} \;+\; \sigma^{2}_{p:s}}{\sigma^{2}_{pi:sf}/n_{i}\,n_{f} + \sigma^{2}_{pf:s}/\,n_{f} + \sigma^{2}_{si:f}/\,n_{i}\,n_{f} + \sigma^{2}_{sf}/\,n_{f} + \sigma^{2}_{s} + \sigma^{2}_{p:s}} = \frac{.0753}{.0914} = .824$$

D-study comp.: .3216/10x2 + .0/2 + .0/10x2 + .0/2 + 0 + .0753 = .0914
(n_i = 10; n_f = 2): .0161 + .0 + .0 + 0 + 0 + .0753 = .0914