

Chapter 4

A Research Synthesis of Validation Practices Used to Evaluate the Satisfaction with Life Scale (SWLS)

Mary L. Chinni and Anita M. Hubley

Researchers are faced with a vast number of studies reporting a variety of results when exploring any topic. Olkin (1996) identified a roughly ten-fold increase in the number of research publications between 1940 and 1990 from 2,300 to 25,000 biomedical journals, 91 to 1,100 journals in psychology, and 91 to 920 journals in mathematics. The amount of information available creates a formidable challenge to researchers and practitioners needing to gather, assimilate, and critically assess the volume of scientific information available to them. Moreover, Cooper et al. (2009) suggested that the increasing volume of knowledge has led to a narrowing of specialties within scientific fields and thus an increasing reliance by researchers on literature reviews to stay current with developments in their fields.

The terms ‘research synthesis’, ‘literature review’, and ‘systematic review’ are often used interchangeably (Cooper 2010). A research synthesis can be thought of as a type of literature review whose primary intention is to assess the quality of information available, to determine whether research findings are consistent and generalizable across populations, and to determine the extent to which findings vary across studies and populations (Mulrow 1994). Manten (1973) adds that literature reviews are “not based primarily on new facts and findings, but on publications containing such primary information whereby the latter is digested, sifted, classified, simplified, and synthesized” (p. 75). What further distinguishes a research synthesis from a literature review is the specific identification of what is to be examined within a literature, and a methodology for examination that can be replicated. Key elements of a research synthesis include: (1) a clearly stated set of objectives, (2) pre-set eligibility criteria for articles used in the study, (3) a methodology that can be replicated, (4) a systematic search to identify studies that meet the eligibility criteria, (5) an assessment of the soundness of all findings, and

M.L. Chinni • A.M. Hubley (✉)

Department of Educational and Counseling Psychology, and Special Education,
The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: anita.hubley@ubc.ca

(6) a systematic presentation of the results of all studies included in the analysis (The Cochrane Collaboration 2002).

A research synthesis of validation practices seeks to examine the methods and procedures that researchers use to evaluate measures and determine whether inferences made about respondents based on those measures are appropriate. Validity is a fundamental concern to measurement specialists and practitioners who use tests to inform and justify social policy decisions, medical and psychological assessments, or an individual's placement, training, and licensing within educational and professional contexts. The *Standards for Educational and Psychological Testing*¹ (AERA et al. 1999) assert that validity is "the most fundamental consideration in developing and evaluating tests" (p. 9).

The Standards (AERA et al. 1999) describe validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Zumbo (2009) argues that it is important to make a distinction between validity evidence and the process of validation. He argues that "validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation" (p. 69). He further explains that the "process of validation involves consideration of the statistical methods, as well as the psychological and more qualitative methods of psychometrics, to establish and support the inference to the explanation" (i.e., validity) (Zumbo 2009, p. 70). Validation practices include such methods as: indicators and descriptors of subject matter expert agreement/disagreement over (content) elements, factor analytic and structural equation modeling approaches to internal structure and measurement invariance, reliability and validity coefficients, and response times and descriptions of response option choices. Validation practices can be thought of as the tools that researchers use to build their argument and justification for the test score inference or explanation.

Research syntheses of test validation practices are still relatively new with little agreed-upon methodology. Based on an examination of ten available research syntheses of test validation practices, we found that syntheses may focus on articles in one or more journals or one or more reviews or entries in a single source (such as the *Directory of Unpublished Experimental Mental Measures* or *Mental Measurements Yearbook*). One measure or multiple measures may form the basis of the synthesis. Some authors choose to include all articles, reviews, or entries or a random or systematic selection. Articles, reviews, or entries may come from a single year, cover a range of years, or focus on particular years. None of the syntheses we examined appeared to use search terms to select articles.

Most research syntheses of test validation practices have focused on reporting the frequency that reliability and validity are reported for measures (e.g., Barry et al. 2013; Hogan and Agnello 2004; Meier and Davis 1990; Qualls and Moss 1996; Slaney et al. 2009, 2010; Whittington 1998). Some studies examined whether these frequencies differed for established, new, or modified measures (Barry

¹ Henceforth referred to as *The Standards*.

et al. 2013; Qualls and Moss 1996; Slaney et al. 2009, 2010; Whittington 1998) or by type of measure or journal (Cizek et al. 2008; Qualls and Moss 1996; Slaney et al. 2009, 2010). Many studies examined the types of reliability (e.g., internal consistency, test-retest, alpha) or validity (based on the present sample or previous research, content evidence, construct evidence, internal structure) evidence presented (e.g., Barry et al. 2013; Cizek et al. 2008; Hogan and Agnello 2004; Jonson and Plake 1998; Qualls and Moss 1996; Slaney et al. 2009, 2010). Whittington (1998) examined whether sample characteristics were taken into account when reporting reliability and validity evidence and Slaney et al. (2009, 2010) examined the extent to which researchers followed a logical order in their presentation of reliability and validity evidence. Cizek, Bowen, and Church (2010) narrowed their focus to the frequency that consequences of testing, one of the five sources of validity identified by *The Standards* (AERA et al. 1999), are reported. Several syntheses explicitly focused on the extent that testing practices may be influenced by testing standards (e.g., Jonson and Plake 1998; Qualls and Moss 1996) or validity theory (e.g., Cizek et al. 2008).

The findings of previous research syntheses of test validation practices suggest that (a) the frequency of reporting reliability and validity evidence each seems to have increased generally over time, although this may vary by journal or field of study, (b) reporting of both reliability and validity evidence seems to have increased generally over time but is much less frequent than reporting either type of evidence on its own, (c) there is a failure to take into account sample characteristics when reporting reliability and validity evidence based on previous research, (d) there is mixed evidence as to whether and how the status of a measure (as pre-existing or new/modified) is related to the frequency of presenting reliability and validity evidence, (e) internal consistency estimates of reliability, which almost always consist of Cronbach's alpha, are reported far more frequently than test-retest reliability estimates, (f) validity evidence is often not reported for all measures in a study, tends to be limited in terms of the amount of evidence provided, and is typically poorly reported, (g) some forms of construct validity evidence tend to be reported more often, (h) there is mixed evidence as to the relative frequency of validity evidence such as factor structure and content evidence, (i) validity evidence such as developmental changes, effect of experimental variables, response processes, and consequences of testing is rarely reported, and (j) there is a disconnection between validity theory, test standards, and validation practice.

Only one research synthesis appears to have examined validation practice with a single measure over time (i.e., Jonson and Plake 1998) and the focus of that study was to use MMY reviews over five periods of validity history to examine the relationship between test standards and validation practices. The purpose of the present synthesis was to examine validation practice with a single, well-known, and widely used measure, the *Satisfaction with Life Scale* (SWLS; Diener et al. 1985). Specifically, we aimed to examine a comprehensive list of validation studies of the SWLS to determine the sources of evidence provided and report, in more detail, the kinds of evidence provided for each source, the rationale for steps taken, criteria used, and the logic adopted for the process involved for each procedure. This study

will contribute to the small but growing literature on validation synthesis by (a) exploring validation practice in more detail, and (b) providing a foundation upon which further validation evidence for the SWLS can be built.

Method

Data Source and Collection

We conducted a literature search for articles on the SWLS containing psychometric or validation evidence using the PsycINFO database. Because the SWLS is used in a variety of disciplines and cultural contexts, and has been translated into several languages, PsycINFO was considered the optimal data source. It is the largest resource devoted to peer-reviewed literature in behavioral science and mental health and includes roughly 2,500 international periodicals, publications from more than 50 countries and journals in 20 languages (American Psychological Association 2013). The search history included publications from 1985 (publication date of the SWLS) to July, 2012. A literature search using the search terms “Satisfaction with Life Scale” and “valid*”, “reliability”, “psychometrics”, “factor analysis” “measurement”, or “measurement invariance” was used to capture studies whose purpose was to provide validity and reliability evidence for the SWLS. Because ‘satisfaction with life’ is a general and widely used term, “Satisfaction with Life Scale” was used as a title search term alongside the other terms listed above. Reference sections of identified articles were also used to identify relevant articles. All studies were screened to determine that: (a) the intent of the study was to provide reliability or validity evidence for the SWLS (as opposed to it being used as a comparison measure or assessment tool in differing research contexts), (b) no modified versions (with the exception of translated versions) of the scale were used, and (c) studies were peer-reviewed.

Coding of Studies

We developed a detailed coding sheet to identify and record validation procedures used in each study. The coding sheet was organized according to the sources of validity evidence as outlined in *The Standards* (AERA et al. 1999): (1) test content, (2) internal structure, (3) relations to other variables, (4) response processes, and (5) consequences of testing. As our intention was to provide a detailed account of the reasoning behind the evidence presented, each category was further broken down to document the rationale for steps taken, criteria used, and the logic adopted for the process involved for each procedure. Two additional sections were added to document reliability evidence and translation methods. Although reliability may

not be included in *The Standards* as evidence of validity, it is a necessary condition for validity (Hubley and Zumbo 2011, 2013). Therefore, it is relevant to examine whether a validation study provided any indication of the reliability of the measure's scores within the context specific to the population. Translation methods were also considered given the large number of translated versions of the SWLS that appeared in our search. As each translation is, in essence, a creation of a new measure, it is important that researchers identify the methods used to create the measure. Details regarding the coding of each section are as follows:

Translations and Adaptations of the SWLS. We first identified if the measure was previously translated or newly translated or if use of a translated measure was suspected but not identified (e.g., sample suggested use of a non-English version of the measure but this was not identified in the paper). Where a newly translated measure was used, we coded for the method of translation used, qualifications of the translators, and whether any pilot tests were conducted.

Reliability. We coded for the presence of internal consistency estimates and test-retest reliability estimates based on the present sample. Alternate forms reliability and inter-rater reliability were not recorded as, respectively, there are no alternate forms of the SWLS scale and no rater decisions in scoring the SWLS. When an internal consistency estimate was provided, we noted what estimate was used, and coded for whether a criterion was identified for the estimate presented. Finally, we coded for whether (corrected) item-total correlations and (average) inter-item correlations were reported. When a test-retest reliability estimate was provided, we coded for whether the test-retest interval was reported; if this was the case, we recorded the length of the interval and whether a rationale for the chosen test-retest interval was provided.

Sources of Validity Evidence. Each study was examined to identify and explore what sources of validity evidence, as outlined in *The Standards* (AERA et al. 1999), were provided. Each category was further subdivided as follows:

Test Content. *The Standards* (AERA et al. 1999) dictate that "item selection, response formats, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers" (p. 44). To determine that inferences drawn from test scores are applicable across groups being tested, evidence must be presented to show that the construct being examined is clearly defined, the selected items accurately represent the construct, the process used in generating and evaluating test items is documented, and results of all empirical analyses conducted in the test development and review process have been presented. We descriptively coded to determine if: a) the construct being examined was clearly defined, b) items were generated based on a literature search, other measures of life satisfaction or related constructs (e.g., well-being, quality of life), or feedback from the target population (i.e., experiential experts), c) content experts (e.g., subject matter experts or experiential experts) were consulted to examine elements of the measure, and d) whether any reference was made to item representation (e.g., of different aspects of life satisfaction), construct under-representation, and construct irrelevant variance.

Internal Structure. To demonstrate that the interpretation of a test reflects the construct it proposes to measure, evidence of its internal structure must be presented. Multivariate statistical techniques are used to examine whether “score variability attributable to one dimension was much greater than the score variability attributable to any other dimension scores obtained from one group” (AERA et al. 1999, p. 20). We first reported whether an exploratory factor analysis (EFA), confirmatory factor analysis (CFA), or both analyses were conducted. When an EFA was conducted, we noted the type of EFA used (i.e., principal components analysis (PCA) or true factor analysis (FA)) and coded for the following information: whether criteria were stated a priori for determining the number of factors, the criteria used to determine the number of factors (i.e., eigenvalues > 1 , scree plot, parallel analysis, percentage of variance explained), whether factor loadings were reported, whether the criterion (e.g., $>.35$) used for determining if an item loads on a factor was reported, and whether percentage of variance explained was reported. If more than one factor was identified, we also recorded information about the types of rotation methods used. When CFA was conducted, we noted the software used and coded for whether researchers reported the number of factors expected and the items expected to load on each factor (if more than one factor expected), the fit indices used, and the rationale and criteria reported for the chosen fit indices. For those studies that examined measurement invariance, we recorded the software used, the type of invariance examined (e.g., gender invariance), the fit indices used, the criteria reported for the chosen fit indices, and the procedures and rationale for the invariance procedures used.

Relations to Other Variables. When relations to other variables were presented as validity evidence, it is clear in *The Standards* (AERA et al. 1999) that the theoretical rationales behind the selection of those variables and “evidence concerning the constructs represented by the other variables as well as their technical properties, should be presented or cited” (p. 20). Questions regarding the degree of association between the measure being examined (e.g., SWLS) and measures representing similar and dissimilar constructs (i.e., convergent and discriminant measures) must be addressed and shown to be consistent with theoretical expectations. The same is true when quasi-experimental or experimental evidence is presented (e.g., known-group differences based on demographic variables or interventions). When evidence is presented that involves assessing relationships with criterion variables, *The Standards* (AERA et al. 1999) notes that “information about the suitability and technical quality of the criteria should be reported” (p. 21). We recorded (a) the terms that researchers used to describe the validation process (e.g., relations to other variables, construct validity, concurrent validity, convergent validity), (b) whether researchers stated their expectations a priori, (c) the types of measures they included (e.g., discriminant measures) and whether terminology used was incorrect (e.g., confusing criterion evidence with convergent evidence), (d) whether any theoretical or empirical rationale was presented for the measures or variables selected, (e) whether reliability evidence (based on the present sample) was reported for the measures chosen, and (f) how the researchers used the evidence

presented (e.g., magnitude, direction, statistical significance of validity coefficients) to make their conclusions about validity.

Response Processes. Whenever a test involves interpretations that presume underlying psychological or cognitive processes used by respondents, observers, or scorers, empirical evidence in support of those premises should be provided. For example, if the SWLS is meant to involve an overall cognitive judgment of one's life by respondents, then empirical evidence should be provided that such a process is taking place. As approaches to examining response processes can be quite varied and less prescribed than some other sources of validity evidence, we simply described any practices but kept in mind some typical approaches (e.g., think-aloud protocols, cognitive interviewing, completion times, documenting or recording responses to items).

Consequences of Testing. The intended social consequences and unintended side effects (Hublely and Zumbo 2011, 2013) of legitimate test interpretation are "relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" (AERA et al. 1999, p. 16). When claims are made regarding the benefits of testing beyond the direct interpretation of test scores (e.g., use of a measure will result in reduced costs or more efficient employee selection), evidence is also needed. Like the response processes section, approaches to consequences of testing can be quite varied and less prescribed than some other sources of validity evidence, so we simply described any use of the words "consequences", "consequential validity", "effects of", "impact of", "implications", and "clinical implications".

Results

Our literature search yielded 35 articles that fit the criteria for inclusion in our study. In several cases, the authors conducted multiple studies using different samples within a single article. For example, a single journal article may have included a group of university students to examine internal structure, a different group of university students to examine dimensionality, and a third group using adolescents to examine relations to other variables. In these cases, each study was treated as an independent study and coded accordingly. This resulted in 46 studies. Of those studies, 31 (67.4 %) involved translated versions of the SWLS. In terms of reliability evidence and the broad categories of sources of validity evidence as outlined in *The Standards*, 37 studies (80.4 %) conducted reliability analyses, 39 studies (84.8 %) examined internal structure, 20 studies (43.5 %) examined relations to other variables, and two studies examined response processes (4.3 %). No studies examined test content or consequences of testing (see Table 4.1). Three studies (6.5 %) provided only reliability evidence. Of the 43 studies providing validity evidence, 25 (58.1 %) examined one source of validity evidence and 18 (41.9 %) examined two sources of validity evidence ($M = 1.33$, $SD = 0.60$). A total of 33 studies (71.7 %) examined both reliability and validity evidence.

Table 4.1 Reliability and validity evidence across studies

Article	Language	Sample	Reliability	Sources of validity evidence ^a			Number of validation sources
				Internal structure	Relations to other variables	Response processes	
Diener et al. (1985) (Study 1)	English	Psychology undergraduates in U.S.	x	x			1
Diener et al. (1985) (Study 2)	English	Psychology undergraduates in U.S.		x			1
Diener et al. (1985) (Study 3)	English	Elderly in U.S.		x			1
Arrindell et al. (1991)	Dutch	Adult clinical outpatients in the Netherlands	x	x			2
Pavot et al. (1991) (Study 1)	English	Elderly in U.S.	x	x			2
Pavot et al. (1991) (Study 2)	English	Undergraduates in U.S.	x	x			2
Neto (1993)	Portuguese	Adolescents in Portugal	x	x			2
Shevlin and Bunting (1994)	English	Psychology undergraduates in Ireland		x			1
Lewis et al. (1995)	English	Undergraduates in Ireland		x			1
Abdallah (1998)	Arabic	Undergraduates in the West Bank	x	x	x		2
Shevlin et al. (1998)	English	Undergraduates in Britain	x	x			1
Arrindell et al. (1999)	Dutch	Young Dutch community adults	x	x	x		2
Lewis et al. (1999)	Czech	Undergraduates in Czech Republic	x	x			1
Pons et al. (2000)	Spanish	High school students and elderly in Spain		x			1
Atienza et al. (2003)	Spanish	High school students in Spain		x			1
Westaway et al. (2003)	English	Adults in South Africa	x	x	x		2
Vautier et al. (2004)	French	Adults in France	x	x			1
Vitterso et al. (2005)	Norwegian/ Greenlandic	Adults in Norway and Greenland	x			x	1

Navrátil and Lewis (2006)	Czech	Psychology undergraduates in Czech Republic	x				0
Tucker et al. (2006)	Russian/English	Students and community adults in the U.S. and Russia	x				1
Wu and Yao (2006)	Taiwanese	Undergraduates in Taiwan		x			1
Květon et al. (2007)	Czech	Undergraduates in Czech Republic	x				0
Hultell and Gustavsson (2008)	Swedish	Teachers in Sweden	x				1
Siedlecki et al. (2008)	English	Community adults in the U.S.	x				1
Wu and Wu (2008)	Taiwanese	Community adults with schizophrenia in Taiwan	x		x		1
Wu and Wu (2008)	Taiwanese	Community adults with schizophrenia in Taiwan	x		x		2
Gouveia et al. (2009)	Brazilian/Portuguese	High school students, undergraduates, community members, teachers, and physicians in Brazil	x		x		2
Laranjeira (2009)	Portuguese	Students, patients, and health professionals in Portugal	x		x		2
Laranjeira (2009)	Portuguese	Patients in Portugal	x				0
Laranjeira (2009)	Portuguese	Students, patients, and health professionals in Portugal			x		1
Slocum-Gori et al. (2009)	English	Community adults in Canada			x		1
Swami and Chamorro-Premuzic (2009)	Malay	Adults in Malaysia	x				1
Wu et al. (2009)	Taiwanese	Undergraduates in Taiwan	x				1
Wu et al. (2009)	Taiwanese	Adolescent athletes in Taiwan	x				1

(continued)

Table 4.1 (continued)

Article	Language	Sample	Reliability	Sources of validity evidence ^a			Number of validation sources
				Internal structure	Relations to other variables	Response processes	
Anaby et al. (2010)	Hebrew	Working adults in Israel	x	x	x	2	
Durak et al. (2010) (Study 1)	Turkish	University students in Turkey	x	x	x	2	
Durak et al. (2010) (Study 2)	Turkish	Correctional officers in Turkey	x	x	x	2	
Durak et al. (2010) (Study 3)	Turkish	Elderly in Turkey	x	x	x	2	
Howell et al. (2010) (Study 1)	English	Undergraduates in U.S.	x	x	x	2	
Howell et al. (2010) (Study 2)	English	Undergraduates in U.S.	x	x	x	1	
Howell et al. (2010) (Study 3)	English	U.S. undergraduates and adults on social networking sites	x	x	x	2	
Bai et al. (2011)	Chinese	Adults in China	x	x	x	1	
Clench-Aas et al. (2011)	Norwegian	Adults in Norway	x	x	x	1	
Glaesmer et al. (2011)	German	Individuals ages 14–91 in Germany	x	x	x	2	
Athay (2012)	English	Caregivers of mentally ill youth in the U.S.	x	x	x	1	
Sancho et al. (2012)	Portuguese	Elderly in Southern Africa	x	x	x	2	
Totals:			36	39	20	2	
						M = 1.33 (SD = 0.60)	

^aNo studies provided evidence based on test content or consequences of testing

Translations and Adaptations of the SWLS

As noted above, 67.4 % of the studies sampled involved translated versions of the SWLS, which includes translations into Arabic, Brazilian-Portuguese, Chinese, Czech, Dutch, French, German, Greenlandic, Hebrew, Malay, Norwegian, Portuguese, Russian, Spanish, Swedish, Taiwanese, and Turkish. These studies come from 25 articles; 12 articles (48.0 %) involved newly translated versions of the SWLS and 9 (32.0 %) used a pre-existing translated version of the scale.² In five articles (20 %), no information was provided about the version used; in these cases, we assumed the test was administered in the sample population's dominant language and involved a translated version of the SWLS. Of the 12 articles involving newly translated versions of the SWLS, all but one essentially used forward and backward translation with multiple individuals involved in the translation process. In nearly all cases, very brief descriptions with little elaboration were provided of the translation procedures used. Only one article indicated the authors incorporated a cultural adaptation in their translation process. In six articles (50.0 %), translation guidelines were cited. In 10 articles (83.3 %), it was noted who did the translations but only half provided any information about the translators' qualifications and that was primarily limited to whether they were native speakers, bilingual, or independent translators. In four studies (33.3 %), pilot tests were conducted but little to no information was provided.

Reliability

Thirty-seven (80.4 %) of the 46 studies provided reliability estimates. Of those studies, 33 (89.2 %) provided an internal consistency estimate. The most commonly identified internal consistency estimate was Cronbach's alpha (27 studies; 81.8 %). Five³ (15.2 %) of the 33 studies provided an "internal consistency coefficient" but were not clear as to which estimate was used. In separate single studies, model-based omega and ordinal alpha were provided in addition to Cronbach's alpha. Finally, another study assessed reliability using parameters estimated from CFA models. In terms of citing criterion values for acceptable internal consistency, only one study clearly cited a criterion (i.e., .70 or higher). Another study made reference to "acceptable" or "satisfactory" alphas of .80 and cited Cronbach's (1951) article, but it is unclear whether a criterion was being listed or the obtained alphas were simply being described.

²The total number of articles do not sum to 25 as one article included a new translation in one language and presumably a pre-existing version in another language.

³Three studies were contained within a single article wherein the author conducted reliability analyses on three different samples.

Seven (18.9 %) out of 37 studies reported inter-item correlation information; average inter-item correlations were reported in four studies and inter-item correlation tables were provided in another three studies. Twenty (54.1 %) out of 37 studies reported item-total, or corrected item-total, correlations; only five (25 %) of these studies reported an acceptable value (of $>.40$ or $>.50$) for evaluating the obtained correlations.

Seven (18.9 %) out of 37 studies examined test-retest reliability, with all studies reporting the time interval between administrations. Intervals examined were 1–2 days, 1 week, 2 weeks, 1 month, 2 months, 3 months, and 6 months. One study examined both 2-week and 1-month intervals and another study examined both 3- and 6-month intervals. Three (42.9 %) of the seven studies provided a rationale for the time interval chosen.

Nine (21.7 %) out of 46 studies did not provide reliability evidence; all but two studies either focused on the internal structure of the SWLS using CFA or examined measurement invariance.

Internal Structure

Thirty-nine (84.8 %) out of 46 studies examined internal structure. Of those 39 studies, 12 studies (30.8 %) conducted exploratory factor analysis, 23 studies (59.0 %) used confirmatory factor analysis, 3 studies (7.7 %) used both methods, and 1 study (2.6 %) was not clear about which approach was used.⁴

Exploratory Factor Analysis. Of 15 studies (i.e., 12 studies using EFA plus 3 studies using both EFA and CFA), 10 studies (66.7 %) used principal components analysis (PCA), 4 studies (26.7 %) used common factor analysis (FA), and 1 study (6.7 %) did not identify the method used. Of the four studies using FA, three studies (75.0 %) used principal axis factoring and one study (25.0 %) used maximum likelihood (ML) as the type of extraction method. No studies stated any criteria a priori for identifying the number of factors. Of the 15 EFA studies, 7 (46.7 %) used ‘eigenvalues greater than one’ as a criterion, 4 (26.7 %) used scree plots, and 3 (20.0 %) used a combination of both criteria. All studies reported the amount of variance explained by the single factor found, but no studies used a criterion value for the amount of variance explained to decide the number of factors. All but one study (93.3 %) reported factor loadings, but no studies identified a criterion (e.g., $>.40$) to determine if an item loaded on the factor. No EFA study reported finding more than one factor so other EFA considerations, such as factor rotation, were not explored.

⁴The focus of this methodological article was on describing steps to identify essential unidimensionality that could be used with either EFA or CFA. SWLS data were used as an example. Because it was unclear as to whether the researchers actually used CFA or EFA analyses with this data, this study was not included in the base rate counts in subsequent internal structure sections.

Confirmatory Factor Analysis. Of 26 studies (i.e., 23 studies using CFA plus 3 studies using both EFA and CFA), 24 studies (92.3 %) specified the software program used for analysis. Between 1985 and 2008, LISREL was used predominantly (i.e., in 11 out of 14 studies; 78.6 %). From 2009 through 2012, a more varied array of software programs were used, including Amos, M Plus, EQS, and SAS. Of the 26 CFA studies, 24 (92.3 %) specified the number of factors expected. The studies used between one and eight fit indices ($M = 4.5$, $SD = 1.84$) to evaluate model fit. The most commonly used fit indices were CFI (21/26; 80.8 %), RMSEA (20/26; 76.9 %), χ^2 (16/26; 61.5 %), TLI/NNFI (13/26; 50.0 %), and SRMR (13/26; 50.0 %) (see Table 4.2). Citation of criteria for the range of acceptable values per fit index varied across the 26 studies: 15 studies (57.7 %) provided criteria for all of the indices, 5 studies (19.2 %) provided criteria for some fit indices but not others, and 6 studies (23.1 %) provided no criteria. Only one study (3.8 %) stated the rationale for the fit indices chosen.

Measurement Invariance. A total of 16 (34.8 %) out of 46 studies from 15 articles examined measurement invariance (see Table 4.3). Approximately half (53.3 %) of the articles included a reference to 'invariance' in the title of the article. Of the 16 studies, most ($n = 14$; 87.5 %) involved a SWLS translated into a language other than English. Generally, the description of measurement invariance was similar across all studies in terms of conducting multi-group CFA using ML estimation and covariance matrices. Thirteen studies (81.3 %) indicated the software package used to conduct analyses; prior to 2009, all studies used LISREL. After that, studies used either AMOS or MPlus. The studies used between one and eight fit indices ($M = 4.4$, $SD = 1.82$) to evaluate the fit of invariance models. The most commonly used fit indices were RMSEA (14/16; 87.5 %) and CFI (13/16; 81.3 %), followed by TLI/NNFI (8/16; 50.0 %), SRMR (5/16; 31.3 %), GFI (3/16; 18.8 %), and NFI (2/16; 12.5 %); a few other indices were used only once. Overall, age and gender were each examined in half of the studies. More specifically, out of 16 studies, 2 (12.5 %) examined age invariance alone, 3 (18.8 %) examined gender invariance alone, and 4 (25.0 %) examined both age and gender invariance. The age divisions varied considerably across the studies both in terms of the number of categories (ranging from two to four categories), the age range of the entire sample, and where age cut-offs were made. Four studies (25.0 %) only examined other types of invariance: two studies from one article examined longitudinal invariance (over 2 month intervals) and two studies examined invariance across different samples (e.g., students, correctional officers, and elderly). One study (6.3 %) examined gender invariance and ethnic invariance. Two studies (12.5 %) examined age invariance and other invariance (i.e., invariance across scattered versus successive item order, nationality). Finally, one study (6.3 %) examined age, gender, education, income, and residence (i.e., metropolitan, town, or rural) invariance.

Table 4.3 Measurement invariance studies

Article	Language	Sample	Invariance mentioned in title	Type of invariance			Software used
				Age	Gender	Other	
Shevlin et al. (1998)	English	Undergraduates in Britain	x		x		LISREL
Pons et al. (2000)	Spanish	High school students and elderly in Spain	x		x		LISREL
Atienza et al. (2003)	Spanish	High school students in Spain	x		x		LISREL
Vautier et al. (2004)	French	Adults in France			x	x	–
Tucker et al. (2006)	Russian/English	Students and community adults in the U.S. and Russia	x		x	x	–
Wu and Yao (2006)	Taiwanese	Undergraduates in Taiwan	x		x		LISREL
Hultell and Gustavsson (2008)	Swedish	Teachers in Sweden			x	x	LISREL
Siedlecki et al. (2008)	English	Community adults in the U.S.			x		–
Gouveia et al. (2009)	Brazilian/Portuguese	High school students, undergraduates, community members, teachers, and physicians in Brazil				x	LISREL
Swami and Chamorro-Premuzic (2009)	Malay	Adults in Malaysia			x	x	AMOS
Wu et al. (2009) (Study 1)	Taiwanese	Undergraduates in Taiwan	x			x	MPlus
Wu et al. (2009) (Study 2)	Taiwanese	Adolescent athletes in Taiwan	x			x	MPlus
Durak et al. (2010) (Study 3)	Turkish	Elderly in Turkey				x	AMOS
Bai et al. (2011)	Chinese	Adults in China			x	x	MPlus
Clench-Aas et al. (2011)	Norwegian	Adults in Norway	x		x	x	AMOS
Glaesmer et al. (2011)	German	Individuals ages 14–91 in Germany			x	x	AMOS
Totals:			8		8	8	8

– = not identified

Relations to Other Variables

Twenty studies (43.5 %) out of the 46 studies included in this synthesis examined relations to other variables (see Table 4.4). The vast majority (19; 95.0 %) of these studies examined convergent evidence, with far fewer including evidence using discriminant measures, although it should be noted that many studies never directly addressed what measures qualified as convergent or discriminant. One study provided what might be best referred to as known-groups evidence only but three of the other studies also attempted to provide known-groups type evidence (4 of 20 studies; 20.0 %).

With respect to the 19 studies examining convergent and/or discriminant evidence, the total number of demographic variables or measures used per study ranged from 2 to 19 ($M=7.16$, $SD=5.09$). *The Standards* (AERA et al. 1999) state that, when comparisons with other variables are presented as validity evidence, the rationale behind the selection of those variables and “evidence concerning the constructs represented by the other variables. . . should be presented or cited” (AERA et al. 1999, p. 20). This means that researchers need to clearly state the rationale for both the construct selected and any variables used to represent that construct. Regarding a rationale for *constructs* used, 13 (68.4 %) out of the 19 studies provided no rationale and 6 studies (31.6 %) provided some rationale. When the rationale for constructs was not explicitly stated, it was often implied, because the constructs were used in previous research or comprised some aspect of the construct of subjective well-being (SWB); thus, authors may have thought that explicitly stating a rationale would be redundant. Regarding a rationale for *measures* used, no studies provided a clear rationale as to why they selected the specific measures chosen.

The terminology used to describe convergent and/or discriminant validity evidence varied considerably both across and within the 19 studies and was, at times, incorrect. Only one study avoided using any terms and three studies referred only to “construct validity”. Otherwise, the following terms were used: “construct validity” (7 studies), “convergent validity” (6 studies), “discriminant validity” (4 studies), “divergent validity” (1 study), “criterion”-(related) validity (5 studies), “criterial validity” (1 study), “concurrent validity” (4 studies), and “predictive” validity or relationships (3 studies).

When it came to stating in advance the expected relationships among variables, only 2 (10.5 %) of the 19 studies clearly identified what they expected to find and 11 studies (57.9 %) were vague in that the expected findings were not explicitly stated by researchers but it was implied they were based on findings in previous literature; 6 studies (31.6 %) did not indicate any expected findings.

An important piece of information in understanding validity coefficients is to know not only reliability estimates for the SWLS scores but also for the other measures used. Out of 19 studies, 6 (31.6 %) provided reliability estimates for all

Table 4.4 Types of relations with other variables evidence

Authors	Language	Sample	Convergent	Discriminant	Known-groups
Diener et al. (1985) (Study 2)	English	Psychology undergraduates in U.S.	x		
Diener et al. (1985) (Study 3)	English	Elderly in U.S.	x		
Arrindell et al. (1991)	Dutch	Adult clinical outpatients in the Netherlands	x	x	
Pavot et al. (1991) (Study 1)	English	Elderly in U.S.	x		
Pavot et al. (1991) (Study 2)	English	Undergraduates in U.S.	x		
Neto (1993)	Portuguese	Adolescents in Portugal	x		x
Abdallah (1998)	Arabic	Undergraduates in the West Bank	x	x*	
Arrindell et al. (1999)	Dutch	young Dutch community adults	x	x	x
Westaway et al. (2003)	English	Adults in South Africa	x	x*	
Wu and Wu (2008) (Study 2)	Taiwanese	Community adults with schizophrenia in Taiwan	x		x
Gouveia et al. (2009)	Brazilian/Portuguese	High school students, undergraduates, community members, teachers, and physicians in Brazil	x		
Laranjeira (2009) (Study 3)	Portuguese	Students, patients, and health professionals in Portugal			x
Anaby et al. (2010)	Hebrew	Working adults in Israel	x		
Durak et al. (2010) (Study 1)	Turkish	University students in Turkey	x	x	

(continued)

Table 4.4 (continued)

Authors	Language	Sample	Convergent	Discriminant	Known-groups
Durak et al. (2010) (Study 2)	Turkish	Correctional officers in Turkey	x		
Durak et al. (2010) (Study 3)	Turkish	Elderly in Turkey	x		
Howell et al. (2010) (Study 1)	English	Undergraduates in U.S.	x		
Howell et al. (2010) (Study 3)	English	U.S. undergraduates and adults on social networking sites	x		
Glaesmer et al. (2011)	German	Individuals ages 14–91 in Germany	x		
Sancho et al. (2012)	Portuguese	Elderly in Southern Africa	x		

* = possibly, but not entirely clear

measures used, 1 study (5.3 %) provided reliability estimates for some measures used but not others, and 11 (57.9 %) provided no reliability evidence for the other measures. Finally, one study (5.3 %) provided reliability estimates but it was unclear whether estimates were based on the study sample or previous research.

It was often unclear what information researchers relied on when it came to interpreting the correlations providing convergent and/or discriminant validity evidence. In some cases (6 studies; 31.6 %), it was not mentioned what information was used to interpret correlations; in other cases, some reference was made to statistical significance (7 studies; 36.8 %), sign of the correlation (positive/negative) (9 studies; 47.4 %), magnitude of the correlation (8 studies; 42.1 %), and/or effect size (1 study; 5.3 %).

Four studies attempted to provide validity evidence by comparing groups in a way that is akin to known-groups validity, although none of the studies used this term; rather, researchers referred to this as either construct or discriminant validity evidence. With the exception of additional procedures to complement convergent/discriminant evidence (e.g., use of factor analysis, partial correlations, or multiple regression to examine contributions of different variables to SWLS scores), no other forms of evidence under the heading of ‘relations to other variables’ were examined.

Response Processes

Two studies provided some evidence related to response processes. In one study, mean time to complete the SWLS as well as ease of use reported by study participants and interviewers was recorded. In another study, a mixed Rasch model was used to identify four latent classes of respondents to the SWLS. The classes tend to reflect differences in the use of the response categories or extreme scores, the difficulty or discriminability of items, or the level of life satisfaction.

Discussion

The purpose of this study was to contribute to the small but growing literature on validation synthesis by (a) exploring validation practices in more detail, and (b) providing a foundation upon which further validation evidence for the SWLS can be built. Thus, our intentions are aimed at measurement and validation specialists, researchers interested in using the SWLS and further examining the validity of inferences made from it, and those individuals who use measures and desire to better understand the validation procedures used to support the inferences drawn from test scores.

While others may have used *The Standards* and the five sources of validation evidence as an inspiration or guide for conducting validation synthesis in the past (e.g., Cizek et al. 2008; Hogan and Agnello 2004; Jonson and Plake 1998), the detailed documentation of procedures and rationales involved in validation practices provided in this study appears to be the first of its kind. If the validation process “begins at the construct definition stage before items are written or a measure is selected, continues through item analysis (even if one is adopting a known measure), and needs to continue when the measure is in use” (Zumbo 1999, p. 11), then a detailed account over time of procedures used, specific to a given test, and within the areas outlined by *The Standards*, is needed.

It is important to note that the majority of studies (67.4 %) included in this synthesis involved translated versions of the SWLS and nearly 50 % of studies involved newly translated versions. Generally, the process that was used to create these translated versions is not well reported. Little information is provided about the individuals who conducted the translation and there is relatively little use of pilot testing reported. Previous research syntheses of validation practices provide no explicit discussion of translated versions of measures, although these measures may be included under ‘modified measures’ in some studies. Translated versions of the SWLS are essentially new measures and so it is critical that the process used to translate the measures is well documented.

Reliability Evidence

Reliability evidence for SWLS scores was consistently well documented across studies. Internal consistency was examined most often (89.2 % of studies). The internal consistency estimate most commonly used was Cronbach's alpha (81.8 % of the time), which shows that classical test theory approaches to reliability still dominate, at least with respect to the SWLS. No study clearly stated a criterion for an acceptable reliability estimate, although it may be viewed as common knowledge to expect estimates of .80 or higher.

In several cases, researchers reported (average) inter-item correlations or (corrected) item-total correlations but failed to identify or discuss acceptable values for, the role of, or how to interpret, these correlations. Inter-item correlations indicate the degree to which items correlate with one another. They are particularly useful in item and test construction to identify whether an item correlates poorly with other items in a test, or whether an item correlates strongly with some items but not others. Both patterns suggest that one may be tapping into another construct altogether (construct irrelevant variance) or that some items tap into another aspect of the construct that the other items are not tapping into (either construct irrelevant variance or construct underrepresentation). Three studies presented inter-item correlations in a table and three studies provided average inter-item correlations. All concluded their results were acceptable, but none discussed the relationship of these correlations to internal consistency or indicated what constitutes an acceptable value, despite the availability of such guidelines. For example, Clark and Watson (1995, p. 316) suggest that, for higher order constructs (such as the SWLS), a mean correlation of .15 to .20 is acceptable whereas for constructs that are more narrowly defined (e.g., talkativeness), a higher mean inter-correlation (i.e., .40 to .50) would be needed. It has been suggested by others (e.g., Netemeyer et al. 2003; Clark and Watson 1995) that the little attention paid to inter-item/average inter-item correlations may be problematic, and that the average inter-item correlation provides a more useful index of internal consistency than does coefficient alpha, the predominant estimate reported in the studies examined. Because coefficient alpha is a function of the number of items in a test and the average inter-correlation among test items, it is possible to achieve a high internal consistency reliability estimate by: (a) having a large number of items, (b) having items that are highly correlated, or (c) a combination of the two. Similarly, Cortina (1993) suggests that coefficient alpha is problematic for scales with more than 40 items. In such cases, the coefficient alpha value may be driven more by the number of items than the magnitude of the correlations among items. The result can be a high internal consistency estimate for a test with items that may correlate rather poorly with one another. Having said this, the small number of items comprising the SWLS limits their influence on the value of coefficient alpha. Thus, alpha will, in this case, be driven more by the magnitudes of the inter-item correlations and is arguably an adequate and more straightforward indicator of internal consistency. Still, more

attention should be paid to inter-item correlations or average inter-item correlations and relaying to the reader what values are acceptable.

The other problematic area of reporting with respect to reliability involved (corrected) item-total correlations. Item-total correlations are computed by correlating the score for a single item with the total score on a scale, and corrected-item total correlations are computed by correlating the score of a single item with the total score on a scale based on the remainder of the items. Researchers should provide some indication of what values are considered acceptable to aid in interpreting the results presented. As a general rule, low or near zero correlations indicate problematic items (Hubley and Zumbo 2011). Generally, values of .50 and above are found to be acceptable values (Netemeyer et al. 2003). It is valuable when both (corrected) item-total correlations and inter-item correlations are presented. One can think of (corrected) item-total correlations as a photograph and inter-item correlations as a sort of zoom lens allowing for a more detailed examination of the items in question. In the case of the SWLS, few studies provided either of these values, and no studies provided both.

Not surprisingly given evidence from previous syntheses, test-retest reliability estimates were provided less often (18.9 % of studies). These studies all reported the test-retest interval but, in a majority of cases, did not provide a rationale for the length of interval chosen. This rationale is an important element needed to assess the obtained estimate because the interval needs to make sense given the expected stability of the construct. With the SWLS, it would be important to choose a time interval length not so short that respondents might recall their responses to items but also not so long that one might anticipate changes to occur in their satisfaction with life. Put another way, it is critical to be able to assume that respondents are not simply trying to report their previous responses and that no real change in satisfaction with life has occurred in order to appropriately evaluate a given test-retest reliability coefficient.

Sources of Validity Evidence

In terms of the five sources of validity evidence as outlined in *The Standards*, only three sources of evidence have been presented for the SWLS. The two primary sources consisted of internal structure and relations to other variables; two studies examined response processes. No studies examined evidence based on test content or consequences of testing.

Internal Structure. Internal structure is the most common type (84.8 % of studies) of validity evidence examined for the SWLS. The majority (59.0 %) of the studies examining internal structure used CFA. The number of factors expected, fit indices used, and software used for analysis were, overall, well reported. The number of fit indices used per study ranged from one to eight, with less than five fit indices used on average. Information needed, but lacking, involves the rationale for fit indices chosen, and, in some cases, criterion values for the fit indices chosen.

When conducting CFA, a rationale for the fit indices used should be provided. Once a model is chosen and estimated, the “fit” of the model must be determined. The fit of a model is largely influenced by sample size and assumptions regarding score distributions and independence assumptions (Tabachnick and Fidell 2013). Although there are a number of indices from which to choose, as a general rule, consistency in results across indices indicate a good fitting model (Tabachnick and Fidell 2013). However, because what fit indices you use influence the results obtained, it is informative to report a rationale for those indices. Tabachnik and Fidell note that, “numerous measures of model fit have been proposed. In fact, this is a lively area of research with new indices seemingly developed daily” (p. 720). To provide a rationale for the selected fit indices not only indicates that the researcher has considered the influence of details specific to the sample being examined, it also provides a context for other researchers using or developing new indices.

Fewer, but still a significant number of, studies (30.8 %) used EFA. Of these, the predominant method used was principal components analysis (PCA; 66.7 %) rather than common factor analysis (FA; 26.7 %). There appeared to be no association between the time (e.g., in which decade) a study was conducted and the EFA method used. All of the EFA studies conducted found evidence to support a one-factor model. Eigenvalues greater than one (46.7 %) were most commonly used to identify the number of factors, followed by scree plots (26.7 %); few studies (20.0 %) used both criteria. One currently recommended criterion is to use loadings obtained from a parallel analysis as a standard against which obtained loading values can be compared (Hayton et al. 2004). Specifically, this procedure involves comparing the eigenvalues found against those eigenvalues that would be obtained from random numbers generated from a data set that is equivalent in sample size and consists of the same number of variables (Ledesma and Valero-Mora 2007). If the eigenvalues obtained exceed those that are randomly generated, then those components can be retained. None of the SWLS studies used this criterion. All but one study reported factor loadings. Surprisingly, no study appeared to use a criterion (e.g., factor loading $>.40$) to determine if an item loaded on a factor. As well, all studies reported the amount of variance explained by the single factor found, but no studies used this as a criterion value to decide the number of factors. For example, no one explicitly stated that a given factor must explain a minimum of 25 % of the variance explained in order for a factor to be retained or considered worthwhile. Given the small number of items on the SWLS, it probably makes more sense to use CFA and test the fit of a unidimensional structure in future studies. If, however, EFA is used, greater attention needs to be paid to the criteria used for (a) identifying whether items load on a factor and (b) the number of factors.

Measurement invariance of the SWLS across groups was examined in 34.8 % of studies, with most studies focusing on invariance across sex or age groups. Notably, however, 87.5 % of these studies examined invariance of a non-English language version of the SWLS so there is a relative gap in the literature on invariance studies with the English version of the SWLS. It would not be surprising for researchers to want to make comparisons in SWLS scores among different groups (e.g., to

examine sex, age, socio-economic status, ethnic or country differences). However, even if there is other validity evidence to support the inferences made from SWLS scores in the different groups, this does not guarantee that the SWLS functions the same way across groups as required for comparison purposes (Horn and McArdle 1992). Only through evidence for measurement invariance can SWLS total scores be deemed to measure the same attribute across groups. If no evidence is presented to support an adequate level of measurement invariance, any differences found among groups cannot be interpreted unambiguously. As Horn (1991, p. 119) has argued, “Without evidence of measurement invariance, the conclusions of a study must be weak”. Thus, if researchers want to compare life satisfaction levels among different groups using the SWLS, evidence of strong or scalar levels of measurement invariance must be shown for SWLS total scores among those groups.

Relations to Other Variables. Validity evidence based on relationships to other variables describes the extent to which there is a relationship between SWLS scores and other variables (whether demographic variables or scores from measures or other variables). Just under half of the studies (43.5 %) addressed relations to other variables; 95 % of studies examined convergent evidence. Many studies never directly addressed what measures qualified as convergent or discriminant, which often made it difficult or impossible to determine whether researchers included, or intended to include, discriminant measures. Moreover, there appeared to be considerable confusion and inconsistency across, or even within, studies as to the appropriate terms to use to describe evidence. Most commonly, criterion-related validity terms (including concurrent or predictive validity) were used to describe convergent evidence. There were three issues related to evidence based on relations to other variables that stood out in this validation synthesis: (a) lack of a clearly state rationale for the selection of constructs and variables, (b) lack of clarity in terms of precisely what researchers expected to find, and (c) poor evaluation of the obtained evidence. We will describe each of these issues in turn.

A clearly stated rationale for why constructs and variables were chosen is generally missing or, at best, very unclear. *The Standards* (AERA et al. 1999) state that, when comparisons with other variables are presented as validity evidence, the rationale behind the selection of those variables and “evidence concerning the constructs represented by the other variables. . .should be presented or cited” (p. 20). For example, if examining the relationship between scores on the SWLS and neuroticism, one needs to provide a rationale for why the construct of neuroticism is being used as well as state a rationale for the specific measure of neuroticism chosen (e.g., the Big Five Inventory subscale of neuroticism). When comparing measures representing the same construct (e.g., life satisfaction or even subjective well-being), there seems to be little point in providing a rationale for why that construct has been selected. However, a rationale for the variable(s) used to measure the construct is needed (e.g., why was a particular single-item measure of life satisfaction chosen for use as opposed to another measure of life satisfaction?). In the case of demographic variables, it is less clear whether a rationale is needed for why researchers have assigned the numbers the way they did. On the one hand, because gender, for example, tends to be clearly defined, it may not be necessary to

justify the variable once you have justified the construct. On the other hand, a variable such as age can have numbers assigned in many different ways (e.g., 1 = 20–49 years (young), 2 = 50+ years (old) vs. 1 = 20–49 years (young), 2 = 50–69 years (middle aged), and 3 = 70+ years (old)). Where the assigning of numbers can alter the construct being examined, the decision about how to categorize the variable may require justification (e.g., why is old = 50+ years in one case vs. 70+ years in another case?).

The Standards (AERA et al. 1999) noted that “when validity evidence includes empirical analyses of test responses together with data on other variables, the rationale for selecting the additional variables should be provided” (p. 20). However, *The Standards* do not explicitly articulate or provide a detailed explanation as to what constitutes a rationale. It is noted that the relationships between scores on the variable of interest and other variables “should be consistent with theoretical expectations” (AERA et al. 1999, p. 20). It is also noted that these variables “might include intended measures of the same construct or of different constructs” (AERA et al. 1999, p. 21). This implies that the rationale requires some theoretical explanation to support why the selected variable (or construct) should or should not be related to the variable (or construct) of interest. Alternatively, or in addition, the rationale could include consistently found empirical evidence of a relationship between the variable of interest and other variables.

The constructs most often used for comparison with the SWLS were subjective well-being (SWB) – including positive and negative affect, personality (particularly neuroticism and extroversion), and psychological constructs (e.g., self-esteem, depressiveness). Of these constructs, SWB was clearly and consistently defined, possibly because the definition is inherent when describing what the SWLS is designed to measure. Most researchers provided a rationale by virtue of explaining how the SWLS is designed to measure the cognitive aspect of life satisfaction. In further situating life satisfaction within SWB, the construct of SWB was fairly well described. Other constructs, such as psychological functioning, perceived health, personality traits, and mental health constructs such as depression and self-esteem, were commonly used but the rationale provided for their use was not clearly articulated. This leaves the reader to wonder why those constructs were chosen, and, by extension, if the researchers themselves had a clear reason for choosing them. Some researchers made mention of relationships to variables without discussing the constructs those variables were designed to capture.

The argument in support of the use of constructs is distinct from the rationale used in support of the variables representing those constructs. *The Standards* (AERA et al. 1999) state that “evidence concerning the constructs represented by the other variables as well as their technical properties, should be presented or cited” (AERA et al. 1999, p. 20). To demand that empirical evidence in support of every variable (measure) chosen be presented may be unmanageable due to page or word restrictions dictated by journals and their editors or place an unreasonable burden on researchers. As well, such information may overwhelm rather than inform the reader. However, some indication as to why the variable was chosen and what construct it was intended to represent is needed. Without some logic to

orient the reader as to where constructs and variables fit within existing literature and a nomological network for the construct and measure of interest, and without the distinction between the two clearly articulated, constructs risk being inconsistently defined. Measures are designed to capture specifically defined constructs. If the definition of the construct varies (or remains undefined) across multiple studies, then the validity of the specific inferences made from the variables (measures) cannot be determined and comparisons across studies cannot be evaluated. As well, information regarding the ability of a measure to consistently capture the intended construct is also compromised. The demographic variables used in the studies examined in this validation synthesis included sex, age, marital status, educational level, employment status, monthly income, health insurance, and sociocultural level. It is important to know how and why researchers constructed the variable(s) as they did to determine comparability across studies. In the studies examined here, the distinction between construct and variable was often blurred, making it difficult to discern arguments in support of a rationale for constructs from those in support of a variable.

In addition to providing a rationale for researchers' choice of constructs/variables, hypotheses or a description of how variables are expected to be related to SWLS scores should be provided based on theory or previous empirical research. Ideally, convergent and discriminant evidence should both be included in a study and these results should be interpreted in relation to each other. Hypotheses or description of expectations should provide information about both the direction and relative magnitude of the expected relationship between the scores and should be stated in advance of the analyses so the obtained evidence may be properly evaluated as supportive of the intended SWLS interpretations or not. Just as statistical procedures used in other areas of evidence (reliability estimates, factor loadings for internal structure) demand criterion values as a means to interpret results obtained, relations to other variables also demands some criterion as a means to interpret the obtained correlations. In essence, researchers provide their own criterion by stating a priori the relationships they expect to find. Without clearly stating this expectation, one is left with a series of correlations of varying magnitudes but no context in which to interpret the immediate study results, their relative standing in relation to a proposed theory, or to the results found in other studies examining similar variables. In the absence of expected values for interpretation, there is no link between results obtained and conclusions drawn.

The problem of not providing hypotheses or a description of how variables are expected to be related to SWLS scores is directly related to the final issue of researchers often presenting either a vague, or lack of, evaluation of the obtained evidence. It was often unclear what information researchers relied on when it came to interpreting the correlations providing convergent and/or discriminant validity evidence. There is a strong tendency for researchers to present correlations without interpreting some, or even all, of them and to then to assert that the findings support validity. In many cases, only a vague reference is made to the sign (positive/negative), magnitude, or statistical significance of the correlations. It is fairly

obvious when reviewing this evidence that researchers do not appear to have a clear sense of how to properly and thoroughly evaluate this kind of evidence.

In a few studies, researchers attempted to provide validity evidence by comparing groups on SWLS scores. Typically, researchers referred to this as either construct or discriminant validity evidence, although it might be better described as known-groups validity. It is notable that in the two studies using demographic variables to form groups, the researchers referred to differences found in previous research whereas in the two studies using pre-existing groups (e.g., clinical sample vs. community members), the differences were presumed rather than based on known differences from previous research or theory. The evidence presented was not particularly strong primarily because the foundation of a previously known difference was not firmly in place. It is also worth noting that it was sometimes a challenge to distinguish between cases in which group differences were simply being examined and those cases in which validity evidence was being presented.

Response Processes. The SWLS is intended to capture the judgmental component of life satisfaction (Diener et al. 1985). When there is a presumption that respondents are using an underlying psychological or cognitive process when responding to test items, *The Standards* recommends that “empirical evidence in support of those premises should be provided” (AERA et al. 1999, p. 20). Of the two studies that addressed response processes, one of those simply examined the mean time to complete the SWLS; while this is useful information, it does not contribute much to our understanding of the underlying process used when responding to the SWLS. The second study used a Rasch model to identify four latent classes of respondents to the SWLS. Further research is needed to better understand the processes used to respond to SWLS items by different groups. For an excellent example of exploring responses processes, readers are referred to Gaderman et al.’s (2011) examination of how children respond to the Satisfaction with Life Scale Adapted for Children (SWLS-C).

In summary, the findings of this synthesis of test validation practices suggest that (a) most psychometric studies on the SWLS are based on a wide range of non-English versions of the SWLS but little attention is paid to the language version of the SWLS or the sample characteristics when reporting reliability and validity evidence from previous research, (b) internal consistency (especially alpha) estimates of reliability are reported far more frequently than test-retest reliability estimates or other indicators such as (average) inter-item correlations or (corrected) item-total correlations, (c) sources of validity evidence for inferences made from the SWLS rely heavily on internal structure and relations to other variables with no evidence presented based on test content or consequences of testing, (d) relations to other variables evidence relies heavily on convergent evidence, and (e) validity evidence in the form of relations to other variables tends to be poorly evaluated and reported. Moreover, as has been reported in previous validation synthesis research, there continues to be a disconnection between validity theory, test standards, and validation practice. While some researchers assert that the lack of some sources of validity evidence reflects either a misunderstanding of the procedures required to demonstrate the quality of evidence presented or disagreement as to what

constitutes validation evidence (e.g., Cizek et al. 2008; Hubley and Zumbo 2013), it seems that much is to be gained by presenting more user-friendly guides to conducting and reporting different sources of evidence and different validation procedures, including a more user-friendly version of *The Standards* (AERA et al. 1999).

Strengths and Limitations

The first strength of this validation synthesis study is that we sought to examine a comprehensive set of peer-reviewed and published validation studies regarding the SWLS. Where other validation synthesis studies have used a random sample of studies to examine validation procedures, we sought to examine all published validation studies found in PsycInfo, one of the largest resources of peer-reviewed literature in behavioral science and mental health. Second, we sought to ground our analysis in procedures proposed by *The Standards*, a widely accepted resource for validation procedures. Third, where many validation synthesis studies have coded according to broad areas of evidence, we sought to examine, in detail, each source of validity evidence found in *The Standards* to identify the specific methods and procedures that researchers use in the process of validating inferences made from the SWLS.

Despite these strengths, there are a number of limitations that affect this study. First, the selected search criteria ruled out any studies that did not use our search terms or those studies where researchers implicitly intended to conduct a validation study but did not explicitly identify their study as such. A second limitation is that the level of detail addressed required a fairly high level of understanding of measurement and statistical methods used and thus some subjective judgment in the coding of evidence. This was particularly evident when coding information related to ‘relations to other variables’ and reflected the confusion and inconsistency in the terminology used, the lack of a clear framework presented by researchers, and poor evaluation and discussion of this type of evidence.

References⁵

- *Abdallah, T. (1998). The Satisfaction with Life Scale (SWLS): Psychometric properties in an Arabic-speaking sample. *International Journal of Adolescence and Youth*, 7, 113–119.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

⁵* indicates articles included in the validation synthesis

- American Psychological Association. (2013). *PsycINFO Quick Facts*. Retrieved May 14, 2013 from <http://www.apa.org/pubs/databases/psycinfo/index.aspx>
- *Anaby, D., Jarus, T., & Zumbo, B. D. (2010). Psychometric evaluation of the Hebrew language version of the Satisfaction with Life Scale. *Social Indicators Research*, *96*, 267–274. doi:10.1007/s11205-009-9476-z.
- *Arrindell, W. A., Meeuwesen, L., & Huyse, F. J. (1991). The Satisfaction with Life Scale (SWLS): Psychometric properties in a non-psychiatric medical outpatients sample. *Personality and Individual Differences*, *12*, 117–123.
- *Arrindell, W. A., Heesink, J., & Feij, J. A. (1999). The Satisfaction with Life Scale (SWLS): Appraisal with 1700 healthy young adults in The Netherlands. *Personality and Individual Differences*, *26*, 815–826.
- *Athay, M. M. (2012). Satisfaction with Life Scale (SWLS) in caregivers of clinically-referred youth: Psychometric properties and mediation analysis. *Administration and Policy in Mental Health and Mental Health Services Research*, *39*, 41–50. doi:10.1007/s10488-011-0390-8.
- *Atienza, F. L., Balaguer, I., & García-Merita, M. L. (2003). Satisfaction with Life Scale: Analysis of factorial invariance across sexes. *Personality and Individual Differences*, *35*, 1255–1260.
- *Bai, X., Wu, C., Zheng, R., & Ren, X. (2011). The psychometric evaluation of the Satisfaction with Life Scale using a nationally representative sample of China. *Journal of Happiness Studies*, *12*, 183–197. doi:10.1007/s10902-010-9186-x.
- Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2013). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior* [online]. doi:10.1177/1090198113483139
- Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397–412. doi:10.1177/0013164407310130.
- Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*, 732–743. doi:10.1177/0013164410379323.
- Clark, A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.
- *Clench-Aas, J., Nes, R. B., Dalgard, O. S., & Aarø, L. E. (2011). Dimensionality and measurement invariance in the Satisfaction with Life Scale in Norway. *Quality of Life Research*, *20*, 1307–1317.
- Cooper, H. (2010). *Research synthesis and meta-analysis: A step by step approach* (3rd ed.). Thousand Oaks: Sage.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- *Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment*, *49*, 71–75.
- *Durak, M., Senol-Durak, E., & Gencoz, T. (2010). Psychometric properties of the Satisfaction with Life Scale among Turkish university students, correctional officers, and elderly adults. *Social Indicators Research*, *99*, 413–429. doi:10.1007/s11205-010-9589-4.
- Gademann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction of Life Scale adapted for children: A focus on cognitive processes. *Social Indicators Research*, *100*, 37–60. doi:10.1007/s11205-010-9603-x.
- *Glaesmer, H., Grande, G., Braehler, E., & Roth, M. (2011). The German version of the Satisfaction with Life Scale (SWLS): Psychometric properties, validity, and population-based norms. *European Journal of Psychological Assessment*, *27*, 127–132.
- *Gouveia, V. V., Milfont, T. L., Nunes da Fonseca, P., de Miranda, P., & Coelho, J. A. (2009). Life satisfaction in Brazil: Testing the psychometric properties of the Satisfaction with Life Scale (SWLS) in five Brazilian samples. *Social Indicators Research*, *90*, 267–277.

- Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7, 191–205.
- Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, 64, 802–812. doi:10.1177/0013164404264120.
- Horn, J. L. (1991). Discussion of the issues of factorial invariance. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 114–125). Washington, DC: American Psychological Association.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Journal of Experimental Aging Research*, 18, 117–144.
- *Howell, R. T., Rodzon, K. S., Kurai, M., & Sanchez, A. H. (2010). A validation of well-being and happiness surveys for administration via the Internet. *Behavior Research Methods*, 42, 775–784. doi:10.3758/BRM.42.3.775.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103, 219–230. doi:10.1007/s11205-011-9843.4.
- Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 3–19). Washington, DC: American Psychological Association Press.
- *Hultell, D., & Gustavsson, J. P. (2008). A psychometric evaluation of the Satisfaction with Life Scale in a Swedish nationwide sample of university students. *Personality and Individual Differences*, 44, 1070–1079.
- Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, 58, 736–753. doi:10.1177/0013164498058005002.
- *Květon, P., Jelínek, M., Klimusová, H., & Vobořil, D. (2007). Data collection on the internet: Evaluation of web-based questionnaires. *Studia Psychologica*, 49, 81–88.
- *Laranjeira, C. A. (2009). Preliminary validation study of the Portuguese version of the Satisfaction with Life Scale. *Psychology, Health & Medicine*, 14, 220–226.
- Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research and Evaluation*, 12, 1–11.
- *Lewis, C. A., Shevlin, M. E., Bunting, B. P., & Joseph, S. (1995). Confirmatory factor analysis of the Satisfaction with Life Scale: Replication and methodological refinement. *Perceptual and Motor Skills*, 80, 304–306.
- *Lewis, C. A., Shevlin, M. E., Směkal, V., & Dorahy, M. J. (1999). Factor structure and reliability of a Czech translation of the Satisfaction with Life Scale among Czech university students. *Studia Psychologica*, 41, 239–244.
- Mantel, A. A. (1973). Scientific literature reviews. *Scholarly Publishing*, 5, 75–89.
- Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology*, 37, 113–115. doi:10.1037/0022-0167.37.1.113.
- Mulrow, C. D. (1994). Systematic reviews: Rationale for systematic reviews. *British Medical Journal*, 309, 597–599. doi:10.1136/bmj.309.6954.597.
- *Navrátil, M., & Lewis, C. A. (2006). Temporal stability of the Czech translation of the Satisfaction with Life Scale: Test-retest data over one week. *Psychological Reports*, 98, 918–920.
- Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. London: Sage.
- *Neto, F. (1993). The Satisfaction with Life Scale: Psychometrics properties in an adolescent sample. *Journal of Youth and Adolescence*, 22, 125–134.
- Olkin, I. (1996). Meta-analysis: Current issues in research synthesis. *Statistics in Medicine*, 15, 1253–1257.

- *Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the Satisfaction with Life Scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, *57*, 149–161.
- *Pons, D., Atienza, F. L., Balaguer, I., & García-Merita, M. (2000). Satisfaction with Life Scale: Analysis of factorial invariance for adolescents and elderly persons. *Perceptual and Motor Skills*, *91*, 62–68.
- Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement*, *56*, 209–214. doi:[10.1177/0013164496056002002](https://doi.org/10.1177/0013164496056002002).
- *Sancho, P., Galiana, L., Gutierrez, M., Francisco, E.-H., & Tomás, J. M. (2012). Validating the Portuguese version of the Satisfaction with Life Scale in an elderly sample. *Social Indicators Research* [online first]. doi:[10.1007/s11205-012-9994-y](https://doi.org/10.1007/s11205-012-9994-y)
- *Shevlin, M. E., & Bunting, B. P. (1994). Confirmatory factor analysis of the Satisfaction with Life Scale. *Perceptual and Motor Skills*, *79*, 1316–1318.
- *Shevlin, M., Brunsten, V., & Miles, J. N. V. (1998). Satisfaction with Life Scale: Analysis of factorial invariance, mean structures and reliability. *Personality and Individual Differences*, *25*, 911–916.
- *Siedlecki, K. L., Tucker-Drob, E. M., Oishi, S., & Salthouse, T. A. (2008). Life satisfaction across adulthood: Different determinants at different ages? *The Journal of Positive Psychology*, *3*, 153–164.
- Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting practices: Incongruence between theory and practice. *Journal of Psychoeducational Assessment*, *27*, 465–476. doi:[10.1177/0734282909335781](https://doi.org/10.1177/0734282909335781).
- Slaney, K. L., Tkatchouk, M., Gabriel, S. M., Ferguson, L. P., Knudsen, J. R. S., & Legere, J. C. (2010). A review of psychometric assessment and reporting practices: An examination of measurement-oriented versus non-measurement-oriented domains. *Canadian Journal of School Psychology*, *25*, 246–259. doi:[10.1177/0829573510375549](https://doi.org/10.1177/0829573510375549).
- *Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research*, *92*, 489–496.
- *Swami, V., & Chamorro-Premuzic, T. (2009). Psychometric evaluation of the Malay Satisfaction with Life Scale. *Social Indicators Research*, *92*, 25–33.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Allyn & Bacon.
- The Cochrane Collaboration. (2002). *An introduction to meta-analysis*. Retrieved July 17, 2012, from <http://www.cochrane-net.org/openlearning/html/mod3.htm>
- *Tucker, K. L., Ozer, D. J., Lyubomirsky, S., & Boehm, J. K. (2006). Testing for measurement invariance in the Satisfaction with Life Scale: A comparison of Russians and North Americans. *Social Indicators Research*, *78*, 341–360.
- *Vautier, S., Mullet, E., & Jmel, S. (2004). Assessing the structural robustness of self-rated satisfaction with life: A SEM analysis. *Social Indicators Research*, *68*, 235–249.
- *Vittersø, J., Biswas-Diener, R., & Diener, E. (2005). The divergent meanings of life satisfaction: Item response modeling of the Satisfaction with Life Scale in Greenland and Norway. *Social Indicators Research*, *74*, 327–348.
- *Westaway, M. S., Maritz, C., & Golele, N. J. (2003). Empirical testing of the Satisfaction with Life Scale: A South African pilot study. *Psychological Reports*, *92*, 551–554.
- Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, *58*, 21–37. doi:[10.1177/0013164498058001003](https://doi.org/10.1177/0013164498058001003).
- *Wu, C.-H., & Wu, C.-Y. (2008). Life satisfaction in persons with schizophrenia living in the community. *Social Indicators Research*, *85*, 447–460. doi:[10.1007/s11205-007-9136-0](https://doi.org/10.1007/s11205-007-9136-0).

- *Wu, C.-H., & Yao, G. (2006). Analysis of factorial invariance across gender in the Taiwan version of the Satisfaction with Life Scale. *Personality and Individual Differences, 40*, 1259–1268.
- *Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the Satisfaction with Life Scale. *Personality and Individual Differences, 46*, 396–401. doi:[10.1016/j.paid.2008.11.002](https://doi.org/10.1016/j.paid.2008.11.002).
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte: Information Age Publishing.