A RESEARCH SYNTHESIS OF VALIDATION PRACTICES USED TO EVALUATE THE SATISFACTION WITH LIFE SCALE

by

MARY L. CHINNI

B.A., The University of Michigan, 1993

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES AND POSTDOCTORAL STUDIES

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2013

Abstract

This thesis had two primary goals. The first was to examine methods and procedures that researchers use in the process of validating the Satisfaction With Life Scale (SWLS) (Diener, Emmons, Larsen & Griffen, 1985). The second was to suggest a framework to organize and examine validation procedures presented in support of measures found across disciplines and journals. A literature search using the PsycINFO database from 1985 through July 2012 was conducted to capture all published validation studies of the SWLS. Each article was coded for reliability and validity evidence (i.e., test content, response processes, internal structure, relations to other variables, and consequences of testing) as described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Each area was further broken down into elements specific to each area to account for the rationale for theoretical decisions made, procedures used, and the completeness of the reporting of validation procedures presented. Results indicate that validation studies for the SWLS focused on internal structure and relations to other variables. Relations to other variables evidence consisted mainly of convergent evidence. Where statistical analyses were conducted, criterion values for interpretation of results were rarely provided. A greater understanding is needed of what constitutes evidence of relations to other variables; how to conceptualize this evidence and provide a rationale for constructs, measures and variables used; and how to describe expected relationships and subsequently evaluate the evidence.

Preface

This thesis is an original, unpublished, independent work by Mary L. Chinni. It was conducted and completed under the supervision of Dr. Anita M. Hubley, Professor, Measurement, Evaluation, & Research Methodology (MERM), Department of Educational and Counselling Psychology and Special Education, at The University of British Columbia.

Table of Contents

List of Tables

List of Abbreviations

CFA = Confirmatory Factor Analysis

EFA = Exploratory Factor Analysis

FA = Common Factor Analysis

PCA = Principal Components Analysis

$\chi 2$ = Chi Square

$\Delta\chi 2$ = Chi-Square change

$\chi 2/df$ = Chi-Square change/Degrees of freedom

Fit Indices:

GFI = Goodness of Fit Index

PGFI = Parsimony Goodness-of-Fit Indicator

AGFI = Adjusted Goodness of Fit Index

IFI = Incremental Fit Index

NFI = Normed Fit Index

NNFI = Non-normed Fit Index

RMSEA = Root Mean Squared Error of Approximation

SRMR = Standardized Root Mean Square Residual

RMSR = Root Mean Square Residual

CFI = Comparative Fit Index

TLI = Tucker Lewis Index

AIC = Akaike's Information Criterion

CAIC = Consistent Akaike Information Criterion

RMR = Root Mean Square Residual

CN = Hoelter's Critical N

Glossary

Definitions provided by the author except where noted.

Consequences of Testing – Consequences of testing are when claims are made regarding the benefits of testing that lie beyond the direct interpretation of test scores. The social consequences and unintended side effects of legitimate test interpretation (Hubley & Zumbo, 2013) are "relevant to validity when [they] can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" (AERA et al., 1999, p. 44). This is one of the five sources of validity evidence as outlined in *The Standards.*

Convergent Evidence – Convergent evidence for validity is gathered by examining the pattern of relationships between the variable of interest and comparison variables that are considered to be conceptually similar.

Criterion-related Evidence – Evidence that demonstrates "the degree to which scores obtained on a measure are related to a criterion. A criterion is an outcome indicator that represents the construct, diagnosis, or behavior that one is attempting to predict using a measure" (Hubley & Zumbo, 2013, p. 14).

Discriminant Evidence – Discriminant evidence for validity is gathered by examining the pattern of relationships between the variable of interest and comparison variables that are considered to be conceptually unrelated.

Factor Analysis – This is "a statistical method used to (a) discover how many factors (representing the latent variables) are being tapped by the items in a test (i.e., exploratory factor analysis) or (b) confirm whether the test items measure the factors as intended (i.e., confirmatory factor analysis)" (Hubley & Zumbo, 2013, p. 12).

Fit Index – This is an index used to determine how well a statistical model, specified a priori, represents the sample data being analyzed. Examples of fit indices referred to in this thesis include: the Comparative Fit Index (CFI) and the Goodness of Fit Index (GFI). There are a number of fit indices available and each is statistically tailored to different aspects of the model being examined. Generally, researchers use a number of fit indices in their analysis. The greater the number of indices demonstrating a good model fit, the more confident one can be that the model is a good representation of the data.

Internal Structure – This refers to "evidence that explores or confirms the structure or dimensionality of a measure. The structure of a measure may be posited to be unidimensional, multidimensional, or hierarchical in structure; this structure may be examined using analyses such as exploratory, confirmatory, or higher order factor analytic techniques. If the items or components of a test are designed to be of increasing difficulty, this may be examined using item response theory modelling. Whether total scores, subscale scores, or items are meant to function similarly or differently across groups, such theoretical assumptions can be tested using measurement invariance or differential item functioning techniques" (A. M. Hubley, personal communication, August 30, 2013).

Known-groups Evidence – This type of validity evidence is gathered by choosing two or more groups that are expected a priori to respond differently to the measure being evaluated based on theory or previous empirical evidence. One evaluates the measure based on whether the expected differences are found.

Life Satisfaction – This is an assessment of feelings and attitudes about one's life at a particular point in time. It is considered to be one of three components of the construct of

subjective well-being (positive affect, negative affect, and life satisfaction) (Diener, 1984).

Relations with Other Variables – One of the five sources of validity evidence as outlined in *The Standards,* this includes a variety of types of evidence discussed in this thesis, including convergent evidence, discriminant evidence, criterion-related evidence, and known-groups evidence.

Reliability – Reliability refers to "the degree to which test scores are repeatable or consistent, or the extent to which test scores are free from measurement error" (Hubley & Zumbo, 2013, p. 13).

Research Synthesis – This is 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.

Response Processes – One of the five sources of validity evidence as outlined in *The Standards,* response processes refer to an examination of the cognitive processes that individuals use when responding to test items.

Subjective Well-being – This is an individual's overall affective and cognitive evaluation of their life (Diener, 2008).

Test Content – One of the five sources of validity evidence as outlined in *The Standards,* this refers to evidence based on judgments about "the adequacy with which the test content represents the content domain" (AERA et al., 1999, p. 11).

Validation – Validation is "the process or method used to support validity and to explain variation in test scores" (Hubley & Zumbo, 2013, p. 11).

Validity – Validity is an "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13).

Acknowledgements

To the cast, in order of appearance:

Dr. Anita M. Hubley, who continues to see something in me that I do not...

Judy Globerman for her insight and guidance.

Bruno Zumbo because at heart you are a skeptic. And you are Italian.

Connie Ng (Conrad) for being exactly who you are.

Mihaela Launeanu, for making me feel normal in the times that I did not.

Ayumi Sasaki (Bad Apple) for all the coding, technical assistance, and being up and ready for pretty much anything.

Sophie Zhu (Dragon) for putting up with my panic, your technical assistance, and for your voracious curiosity - so inspiring to be around.

Shawna Goodrich for reminding me that I am not normal, and…cows.

Tavinder Ark for her gummy bear art, video sharing, and her willingness to drop whatever she may be doing, at my every cry for help, to guide me in minutes through a technical problem I would have spent decades trying to sort out.

Juliette Lyons-Thomas for, well, pretty much everything.

Dallie Sandilands for her great feedback in the final hour.

Angena Kalhar for your very presence and demented sense of humor.

Dedication

For Theresa – without you completion of this thesis would not have been possible.

Chapter 1: Introduction

When exploring any topic, researchers face a vast number of studies reporting a variety of results. 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 from 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. As well, Cooper (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) 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 6) a systematic presentation of the results of all studies included in the analysis (Cochrane Institute, 2012).

A research synthesis of validation practices seeks to examine the methods and procedures researchers use to evaluate measures and determine whether inferences made about respondents 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, and/or an individual's placement, training, and licensing within educational and professional contexts. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 1999) assert that validity is ''the most fundamental consideration in developing and evaluating tests'' (p. 9).

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). Inherent in this definition is the distinction between validity evidence and the process of validation. Zumbo (1999) states that 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" (p. 11). Validity evidence can be thought of as the tools that researchers use to build their argument and justification for the appropriateness and use of the measure being examined.

The debate as to what researchers believe, understand, or identify to be sufficient evidence to justify the use of a given measure remains. As Cizek, Bowen, and Church (2008) note,

> Although broad agreement exists about the importance of validity and major tenets of modern validity theory, disagreements persist regarding the definition and boundaries of the concept, and regarding what sources of validity evidence are desirable or necessary for sustaining defensible inference based on test scores (p. 398).

The validity evidence required for a given test is context specific and depends on the inferences to be drawn from the test administered and the assumptions implicit in the subsequent interpretations (Kane, 2008). For example, the intention of some employment tests, used during the hiring process, is to predict future job performance (predictive validity) while the intention of other employment measures may be to evaluate job performance on persons presently employed (concurrent validity) (Biddle, 2010). Face validity (the degree to which a measure appears to be related to a specific construct) may be important in the context of achievement tests, but not necessarily wanted in measures designed to detect malingering. Furthermore, for achievement tests, content validity is critical to establish an individual has mastered the skill being evaluated. However, Kane (2008) points out that achievement tests often have an additional predictive component in that they are "assumed to reflect readiness for something (e.g., college, work); if they have no implications beyond achievement on a domain of test items, why bother?" (p. 79).

There is still disagreement regarding what sources of validity evidence are "desirable or necessary for sustaining defensible inferences based on test scores" (Cizek et al., 2008, p. 398). Kane (2009) suggests that "different interpretations/uses will require different kinds and different amounts of evidence" (p. 49). For example, many employment measures aim to predict a potential applicant's future job performance and thus the process of validation may focus on establishing a given measure's predictive validity. For educational measures, it is suggested that content validity is of primary importance, e.g. that the items generated for an algebra test are

relevant and representative of knowledge and understanding of algebraic procedures. At the same time, if that same algebra test is being used to dictate an individual's access to an advanced course, then one would need to determine whether the skills being assessed are indeed necessary prerequisites, that test scores remain uninfluenced by other extraneous variables, that success in the advanced course is accurately assessed, and that those individuals with high scores do perform better in advanced courses than those who score lower on the test (AERA, APA, & NCME, 1999). Thus, judgments "about the appropriate sources and quantity of evidence for validation efforts depend on the nature, breadth, and complexity of the intended inference; the relative seriousness of inaccurate inferences; and the resources available for the validation effort" (Cizek et al., 2008, p. 11). A research synthesis of validation practices can determine in what contexts a given measure is being used, what evidence is presented in support of its use, and how researchers conceptualize and conduct the process of validation within the context that the measure will be used.

Research syntheses of approaches to test validation are, at present, relatively new with no clear methodology identified. To date, there are eight notable published studies. Meijer and Davis (1990) focused on the *Journal of Counseling Psychology* to examine the reporting of validity evidence in all articles published in 1967, 1977, and 1987. They found that researchers rarely provided psychometric data about measures and a subsequent lack of reporting of validity evidence and validation practices. Hogan and Agnello (2004) investigated validity reporting practices in a sample of 696 research reports, encompassing a wide range of disciplines, listed in the APA's *Directory of Unpublished Experimental Mental Measures.* They found that just over half of the reports documented included validity evidence with no report providing more than two sources of evidence. The majority of those reports (approximately 90%) simply provided

correlations with other variables without identifying the comparative variables or providing a context in which to interpret the correlations presented and rarely used approaches to validation practices described by test standards. As well, they found no reporting of response processes or content validity evidence (which was particularly striking in the case of achievement tests examined). Cizek et al. (2008) focused on the lack of validity reporting regarding test consequences and proposed a restructuring of validity theory whereby consequences of testing not be considered "an integral part of validity theory and practice" (p. 410). In a follow-up article, Cizek, Bowen and Church (2010) examined, over a 10-year period, articles published in applied educational assessment and education policy journals. Similar to previous research, they found that validity evidence regarding test consequences was non-existent and further proposed that examination of test consequences be treated rigorously but distinct from the process of validation used to support the inferences drawn from test scores.

Other studies examined the relationship (or lack thereof) between validity theory and validation practice. Jonson and Plake (1998) examined the relationship between changes in validity theory and validation evidence presented for a single achievement test over a 50-year period (1954 through 1985). They compared the developments in validity theory, as expressed through changes outlined in the AERA, APA, and NCME published test standards and recommendations, to the evidence presented in test reviews by measurement professionals. In their aim to determine whether changes in theory caused changes in practice, changes in practice informed validity theory, or if a relationship between theory and practice even existed, they found that while test standards inform professionals' conceptual knowledge of validity, in practice "they are not as influential in determining the actual validity requirements that should be applied" (p. 751). Cizek et al. (2008) examined sources of validity evidence for all tests

contained within a single edition of the *Mental Measurements Yearbook* using the types of validation approaches outlined in the 1999 *Standards (AERA, APA & NCME)*. They also found that, within their sample tests, developers and measurement specialists rarely incorporated these perspectives into their test evaluations. In addition, Slaney, Tkatchouk, Gabriel, and Maraun (2009) found a distinct disconnect between validity theory and validation practices in their examination of 2004 articles, published in the journals *Educational and Psychological Measurement*, *Psychological Assessment*, *Journal of Personality Assessment*, and *Personality and Individual Differences*. Barry, Chaney, Piazza-Gardener, and Chavarria (2013), in their study analyzing seven journals in the area of health education and health behavior, examined articles involving scales used in both primary data collection and secondary data analysis to determine what reliability and validity statistics were presented and how they were assessed. They also found there was a disconnect between current practice and recommended testing practices, and asserted the "need for reporting of psychometric properties to be explicitly outlined as a requirement for publication" (Barry et al., 2013, p. 6). Qualls and Moss (1996) sought to determine whether researchers presenting reliability and validity evidence conduct their practices according to those procedures proposed by *The Standards*. They examined all articles in 22 out of 25 APA journals published in 1992 and found that "a disconcerting number of authors are not complying with various standards dealing with testing" (Qualls & Moss, 1996, p. 214).

Beyond the published literature, a Validity Symposium presented at the 2012 American Education Research Association Conference presented the empirical findings of seven research syntheses of validity evidence reported across areas in education, psychology, and the health sciences. Following the *Standards* (AERA, APA, &NCME, 1999), the papers focused on the

reporting of five sources of validity evidence (content-related, response processes, internal structure, associations with other variables, and consequences) and sought to determine whether validity theory informed validation practice (Shear & Zumbo, 2012). A meta-synthesis of the presented studies indicated that, amongst additional findings, reliability indices and other internal consistency analyses are cited as validation evidence; evidence of relationships and comparisons with other variables are reported although there is confusion regarding terminology; there is large variation in the use of content-related evidence in validation research; evidence related to internal structure has increased, and validity evidence based on response processes and consequences is essentially non-existent (Lyons-Thomas, Liu, Olivera, & Zumbo, 2012).

Previous research has focused on the relationship between changes in validity theory and validation evidence presented over time, and sources of validity evidence reported in research studies and test reviews across a wide range of disciplines. The present thesis is a methodological study that aims to examine methods and procedures that researchers use in the process of validating a single, well-known measure, the *Satisfaction With Life Scale* (SWLS) (Diener, Emmons, Larsen & Griffen, 1985). The SWLS, as shown in Appendix A, is a widely used measure in that 1) it is used cross-culturally and has been translated into many languages, 2) data has been collected across a broad range of samples such as older adults, prisoners, individuals under inpatient care for alcohol abuse, abused women, psychotherapy clients, elderly caregivers of demented spouses, and persons with physical disabilities, and college student samples, 3) studies across cultures and samples have examined reliability, internal consistency, and correlations with other constructs and variables, 4) it is asserted that it shows promise for clinical applications and may be used in conjunction with economic and social indictors to inform public policy (Pavot & Diener, 1993). This study aims to examine all validation studies of the SWLS

contained in the database PsycINFO using psychometric search terms or identified in the reference sections of these articles. This study will contribute to the small, but growing, literature on validation synthesis; suggest a framework that can be used for future examination of published findings for measures used across disciplines and journals; and provide a foundation upon which further validation evidence for the SWLS can be built.

Chapter 2: Methods

2.1 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 to be 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 (The 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 as a data source to ensure all relevant articles were identified. All studies identified were screened to determine that: 1) 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), 2) no modified versions of the scale were used, and 3) studies were peer-reviewed.

2.2 Method

I developed a detailed coding sheet, as shown in Appendix B, 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 1999 *Standards*; (1) test content, (2) internal structure, (3) relations to other variables, (4) response processes, and (5) test consequences. As my 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. Reliability may not constitute evidence of validity on its own but it is a necessary condition for validity (Hubley & Zumbo, 2013). Therefore, it is relevant to examine whether a validation study provided any indication of the reliability of the SWLS within the context specific to the population. Translation methods were also considered given the relatively 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 in the creation of this measure. Details regarding the coding of each section are as follows:

2.3 Reliability

Two reliability estimates were coded dichotomously: internal consistency estimates and test-retest reliability estimates. Alternate forms reliability was not included as there are no alternate forms of the SWLS scale. Where an internal consistency estimate was provided, I noted what estimate was used, and dichotomously coded for the presence of a criterion for the estimate presented. In addition, I coded for whether item-total correlations and inter-item correlations were reported in the study. Where a test-retest reliability estimate was provided, I dichotomously

coded for whether the test interval used was reported and whether a rationale for the test interval was provided. If a test interval was provided, I recorded the length of the interval.

2.4 Sources of Validity Evidence

Each study was analyzed to determine the sources of validity evidence provided. I dichotomously coded according to the following sources of evidence outlined in the 1999 *Standards*: (a) test content, (b) internal structure, (c) relations to other variables, (d) response processes, and (e) test consequences. Each category was further subdivided into further categories as follows:

2.4.1 Test Content

The *1999 Standards* 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." (AERA et al., 1999, p. 44). To determine that inferences drawn from test scores are applicable across groups being tested, evidence must be presented to indicate that the construct being examined is clearly defined, the items chosen accurately represent the construct, the process used in generating and evaluating test items is documented and reported, and results of all empirical analyses conducted in the test development and review process are presented. I 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 experiential experts and/or a target population (i.e., experiential experts), and c) subject matter experts (SMEs) or experiential experts (EEs) were consulted to examine elements of the measure, and d) whether any reference was made to item representation (i.e., the degree that items in the measure

represent the full range of the construct of life satisfaction), construct underrepresentation, and construct irrelevant variance.

2.4.2 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). I identified the type of analysis conducted (CFA, EFA, other). Where EFA was conducted, I noted the type of EFA used and dichotomously coded for the following procedural steps: criteria stated for number of factors found, what criteria was reported (e.g., eigenvalues > 1, scree plot, percentage of variance explained), whether factor loadings were reported, and the criterion for factor loadings reported. Where CFA was the type of analysis conducted, I noted the type of software used and dichotomously coded to determine if researchers reported the number of factors expected, the fit indices used, and the rationale and criteria for the fit indices chosen. For those studies examining measurement invariance, the type of invariance examined (e.g., gender invariance) and the procedures and rationale for those procedures were descriptively recorded (i.e., structural equation modeling, fit indices used, rationale).

2.4.3 Relations to Other Variables

Where comparisons with constructs (variables) are presented as validity evidence, the theoretical rationales behind the selection of those constructs (variables) and "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). Where measures are chosen to

determine relationships between similar and dissimilar constructs (convergent and discriminant evidence), questions regarding the degree of association between the measure being examined and comparison measures must be addressed and shown to be consistent with theoretical expectations (AERA et al., 1999, p.14). Where evidence presented involves assessing relationships with criterion variables, "information about the suitability and technical quality of the criteria should be reported" (AERA et al., 1999, p. 21). For this section, I recorded: 1) how researchers described the validation process (e.g. relations to other variables, construct validity), 2) did they identify the purpose for the measures chosen and state their expectations clearly (e.g., convergent evidence/similarities to construct vs. discriminant evidence/dissimilarities to construct), 3) did they use technical terminology and, if so, was it used appropriately (e.g. confusing criterion evidence with convergent evidence), 4) was the rationale for their choice of measures clearly stated, 5) was reliability evidence (based on the current sample) reported for the measures chosen, and 6) how did researchers conclude that the evidence found supported validity (e.g. magnitude/direction/statistical significance of correlations).

2.4.4 Response Processes

For those tests involving interpretations that presume underlying psychological or cognitive processes used by individuals being examined, "empirical evidence in support of those premises should be provided" (AERA et al., 1999, p. 20). Similarly, if those same processes are used by observers or scorers involved in testing procedures, supporting evidence should also be provided. Descriptions of the following were recorded: questions and probing responses to items (e.g., think-aloud protocols, cognitive interviewing), documenting or recording responses to items, indication of time needed to complete questionnaire, and post-test questionnaires or interviews.

2.4.5 Test Consequences

The social consequences and unintended side effects (Hubley & Zumbo, 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. 44). When claims are made regarding the benefits of testing beyond the direct interpretation of test scores, evidence is also needed. Descriptions of the following were recorded: use of the words "consequences", "consequential validity", "effects of", "impact of", "implications", and "clinical implications". Instances where consequences were misunderstood as test misuse and any citations related to consequences (Messick, *The Standards*, Kane) were also recorded.

2.5 Translations and Adaptations of Measures

As many studies examined in this validation synthesis of the SWLS involved translated versions of the scale, a section examining information about translation methods and procedures was included. I devised a coding system consisting of 'Yes', 'No', 'Partially' and 'Unclear' to identify whether a previously translated measure or a newly translated measure was used. Where a newly translated/adapted measure was used, I coded for the method of translation used, qualifications of the translators, and whether any pre-tests or pilot tests were conducted.

Chapter 3: Analysis and Results

3.1 General Summary

Our literature search yielded 36 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 an overall number of studies examined of N = 46. As shown in Table 1, of those studies, 31/46 (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*, 36/46 (78.3%) conducted reliability analyses, 39/46 (84.8%) examined internal structure, and 21/46 (45.7%) examined relations to other variables. Only one study out of the 46 studies (2.2%) made reference to response processes. No studies examined test content or consequences of testing.

Table 1 – Reliability and Validity Evidence Across Studies

	Article	Language	Reliability	Sources of Evidence		
				Internal Structure	Relations to Other Variables	Response Processes
1985	Diener et al. [Study 1]	English	✓	✓		
1985	Diener et al. [Study 2]	English			✓	
1985	Diener et al. [Study 3]	English			✓	
1991	Pavot et al. [Study 1]	English	✓	✓	✓	
1991	Pavot et al. [Study 2]	English	✓	✓	✓	
1991	Arrindell et al.	Dutch	✓	✓	✓	
1993	Neto	Portuguese	✓	✓	✓	
1994	Shevlin & Bunting	English		✓		
1995	Lewis et al.	English		✓		
1998	Shevlin et al.	English	✓	✓		
1998	Abdallah	Arabic	✓	✓	✓	
1999	Lewis et al.	Czech	✓	✓		
1999	Arrindell et al.	Dutch	✓	✓	✓	
2000	Pons et al.	Spanish		✓		
2003	Atienza et al.	Spanish		✓		
2003	Westaway et al.	English	✓	✓	✓	
2004	Vautier	French	✓	✓		
2005	Vitterso et al.	Norwegian/Greenlandic	✓			✓
2006	Tucker et al.	Russian/English		✓		
2006	Wu & Yao	Taiwanese		✓		
2006	Navratil & Lewis	Czech	✓			
2007	Kveton et al.	Czech	✓			
2008	Gouveia et al.	Brazilian/Portuguese	✓	✓	✓	
2008	Siedlecki	English	✓	✓	✓	
2008	Hultell & Gustavsson	Swedish	✓	✓		
2008	Wu & Wu [Study 1]	Taiwanese	✓	✓		
2008	Wu & Wu [Study 2]	Taiwanese	✓	✓	✓	
2009	Slocum-Gori et al.	English		✓		
2009	Wu et al. [Study1]	Taiwanese	✓	✓		
2009	Wu et al. [Study 2]	Taiwanese	✓	✓		
2009	Swami et al.	Malay	✓	✓		
2009	Laranjeira [Study 1]	Portuguese	✓	✓		✓
2009	Laranjeira [Study 2]	Portuguese	✓			
2009	Laranjeira [Study 3]	Portuguese			✓	
2010	Anaby et al.	Hebrew	✓	✓	✓	
2010	Durak et al. [Study 1]	Turkish	✓	✓	✓	
2010	Durak et al. [Study 2]	Turkish	✓	✓	✓	
2010	Durak et al. [Study 3]	Turkish	✓	✓	✓	
2010	Howell [Study 1]	English	✓	✓	✓	
2010	Howell [Study 2]	English	✓	✓		
2010	Howell [Study 3]	English	✓	✓	✓	
2011	Clench-Aas	Norwegian	✓	✓		
2011	Bai et al.	Chinese	✓	✓		
2011	Glaesmer et al.	German	✓	✓	✓	
2012	Sancho et al.	Portuguese	✓	✓	✓	
2012	Athay	English	✓	✓		
	TOTAL		**36**	**39**	**21**	**1**

3.2 Rater Discrepancy

I examined rater agreement to evaluate how well the coding sheet served as a tool to identify and record validation procedures used in each study. The coding sheet was designed to serve as both a detailed checklist for the kinds of evidence, rationales, and criteria that might be provided and as guide that even those with a moderate understanding of validation procedures could follow. This was particularly challenging in that the level of detail I felt needed to be addressed often required a fairly high level of understanding of measurement and statistical methods. As well, the utility of a document involving the accounting of more subtle concepts, particularly relations to other variables, an area that appears to be poorly understood by even seasoned researchers, needed to be evaluated.

To assure that each area of evidence was sufficiently represented by the studies chosen, I coded each study according to the broad categories of validation evidence as reflected in the coding sheet. Reliability was further divided into two categories (internal consistency and test-retest); internal structure was divided into the categories of EFA, CFA, and measurement invariance; and relations to other variables was further divided into convergent and discriminant evidence as these were the primary types of relationships examined. I then found the area with the least number of studies reporting that evidence (i.e., only five studies examined test-retest reliability and only two studies provided discriminant evidence). As there were only two discriminant validity studies, I automatically included those studies in the inter-rater agreement analysis. I randomly selected three out of the five test-retest reliability studies. For the remaining areas, I randomly selected studies until we achieved a sample where each category was reviewed by the second rater a minimum of four times. My goal was to sample anywhere from 10-15 articles overall to be coded by a second rater. As a result of the selection procedure, I

implemented, a second rater recoded a total of 12/46 (26%) articles from the original sample. The second rater is presently pursuing a Master's program in Counselling Psychology at The University of British Columbia. She completed a graduate level course in Basic Principles of Measurement and is also involved as a research assistant in another validation study. The articles sampled, and the areas of evidence reported per article, are shown in Table 2.

Within each broad category of evidence, I coded for a number of elements, i.e., each area of evidence was further broken down into elements to document the rationale for steps taken, criteria used, and the logic adopted for the process involved for each procedure as shown in Appendix A. For example, coding for reliability involved an examination of 14 elements related to reliability, coding for EFA involved coding for the reporting of 24 elements, etc. My intention was to identify areas where our coding sheet may not have clearly indicated to the coder what precisely constituted as adequate reporting for any given element. For example, if I found there were many discrepancies regarding providing a rationale for fit indices used in CFA, I sought to determine whether this was due to poor reporting by study authors, lack of knowledge on the part of the coder, and/or whether the coding sheet itself was the source of confusion. Where consensus between coders could not be reached, my supervisor, an expert in measurement and validation, was consulted to determine the final decision.

Table 2 – Articles Sampled for Rater Discrepancy Analysis and Topics Covered

Study	IC	Rel Test-Retest	IS EFA	IS CFA	IS MI	RV
1985 Diener 1	1	1	1			
1985 Diener 2						1
1985 Diener 3						1
1998 Abdallah	1	1	1			1
1999 Arrindell et al.	1		1			1
2000 Pons				1	1	
2006 Tucker et al.				1	1	
2009 Swami et al.	1			1	1	
2008 Gouveia et al.	1		1	1	1	1
2009 Laranjeira 2		1				
2010 Durak et al. 1	1			1		1
2011 Bai et al	1			1	1	
Total	**7**	**3**	**4**	**6**	**5**	**6**

IC = Internal Consistency, Rel = Reliability, IS = Internal Structure, EFA = Exploratory Factor Analysis, CFA = Confirmatory Factor Analysis, MI = Measurement Invariance, RV = Relations to Other Variables

There were no discrepancies in coding any of the elements involved in test-retest reliability, the presence of measurement invariance, or the two studies providing discriminant validity evidence. Of the seven reliability studies, the only discrepancies occurred within two studies when coding for the interpretation of item-total correlations as reliability evidence. Additionally, there was one case where coders disagreed as to whether researchers reported average inter-item correlations and one case where there was disagreement as to what type of internal consistency estimate was provided. Generally, discrepancies regarding numerical values reported occurred where reporting was unclear (e.g., inter-item correlations reported in a table but not referenced or discussed in the body of the article, and a numerical value for a reliability estimate was provided but not clearly identified as such). For the four studies involving EFA, in one study there was one disagreement regarding the reporting of a criterion for factor loadings, and another disagreement regarding the criterion for the number of factors found. The source for

this discrepancy seemed to be due to both a lack of technical knowledge on the part of the coder and poor reporting by study authors. As was indicated in our study overall, there is general confusion regarding the distinction between reporting variance explained and providing a criterion for variance explained as a decision point for deciding the number of factors to retain. Disagreements in CFA coding occurred in five out of 13 studies and involved whether researchers stated the number of factors expected (two studies) and whether criteria for fit indices were reported (three studies). Regarding the number of factors, it was often assumed, but not stated, that one factor was being explored. Disagreements regarding criteria occurred when researchers provided a list of fit indices but criteria were not identified for all fit indices. For the seven studies involving translations, discrepancies occurred regarding what constitutes clear reporting of methodology (two studies), and what qualifies as translator qualifications (i.e., describing who translated the SWLS versus what their qualifications were; three studies). These discrepancies regarding methodology perhaps occurred because of a lack of clarification on my part regarding what qualifies as a clear reporting of methodology. In an additional case, the discrepancy was directly the result of poor reporting in that the article did not clearly state whether a new or previously translated version of the SWLS scale was used.

For the 'Relations to Other Variables' section, inconsistency in reporting by researchers made efforts to code this section in a systematic manner highly problematic. As discussed earlier, confusion regarding terminology, along with the lack of a clear rationale indicating why measures or constructs were used, was the source of many discrepancies. Rarely was a clear rationale presented or expectations stated, results often consisted of interpreting the meaning of correlations post-analysis, and a discussion regarding the relative magnitude of the correlations found was lacking. Therefore, it was not possible to, for example, simply record the presence or

absence of any value or criteria as we were able to do in the other areas above described. In general, it was difficult, regardless of the information provided, to discern what type of analysis was conducted and what forms of evidence were being pursued to provide support for the interpretations of SWLS test scores.

To summarize, it appears that where a straightforward accounting of numerical values was involved (e.g., what type of internal consistency was used or what fit indices were reported), the coding sheet was adequate. Where an argument, hypothesis, rationale, or interpretation of relative comparisons were required (e.g., discussing convergent and discriminant evidence in relation to each other) or might be assumed or implied, discrepancies were greater. To the extent the discrepancies were a reflection of our coding system versus poor reporting or comprehension of validation procedures on the part of researchers, is difficult to determine.

3.3 Translations and Adaptations of Measures

As noted, 31 out of 46 (67.4%) 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, Spanish, Swedish, Taiwanese, and Turkish. Of these, 15/31 (48.4%) studies involved newly translated versions of the SWLS and 10/31 studies (32.3%) used a pre-existing translated version of the scale. Five studies (16.1%) provided no information about the version used. In these cases, we assumed the test was administered in the sample population's dominant language. For example, if the study was conducted on "community members living in China", and no reference to a translation was provided, we assumed that a non-English version of the SWLS was used. Finally, one study (3.2%) described who translated the scale and provided a fairly detailed description of transcription process used, but then directed readers to a previously translated version of the

scale published eight years prior to their present study. Whether they used the previously translated version and were simply re-iterating the translation process used by the initial author, or actually provided a new translated version of the scale was unclear. In 11 of the 31 (35.5%) studies, it was described who did the translations but none of the studies specified the translator's qualifications. Seven of the 31 studies (22.6%) identified guidelines used in the translation process. Of these, the guidelines proposed by Brislin (1970, 1980, 1986) were the most commonly cited. One study referenced guidelines proposed by de Figueiredo and Lemkau (1980). Four of the 31 studies (12.9%) conducted pilot tests but provided only the sample size and a rudimentary description of the groups taking the test.

Overall, the process and methods used to create translated versions of the SWLS were briefly and poorly described. In 15/31 studies (48.3%), brief descriptions were provided of the translation procedures used although, in some cases, procedural terms were not clearly defined. In 11 of the studies (69%), it was described who did the translations but none specified their qualifications. Four studies (12.9%) conducted pilot tests but provided only the sample size and a rudimentary description of the groups taking the test.

3.4 Reliability

Thirty-six of the 46 (78.3%) studies provided reliability estimates. Of those studies providing reliability evidence, 33 studies (out of 36 studies; 91.7%) provided an internal consistency estimate. The most commonly identified internal consistency estimate was Cronbach's alpha (27/33 studies; 82%). Four of the 33 studies (12.1%) provided an internal consistency reliability coefficient but were not clear as to which estimate was used, e.g., "internal consistency coefficient" (Neto, 1993, p. 129; Durak et al., 2010, p. 417). It should be noted that three of these studies were contained within a single article wherein the author conducted

reliability analyses on three different samples. One study 1/33 (3%) provided, in addition to Cronbach's alpha, a second estimate identified as model-based omega. Finally, one study (out of 33 studies; 3.0%) assessed reliability using parameters estimated from CFA models and one study conducted an IRT analysis (3.0%). The three studies (out of 33 studies; 9.1%) that did not provide an internal consistency estimate intended to examine test-retest reliability only. We also examined whether criterion values for internal consistency were cited. One 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. (Navratil & Lewis, 2006). No other studies provided a criterion for what was deemed an acceptable reliability coefficient.

Three studies (out of 36 studies; 8.3%) reported inter-item correlations and another three studies (8.3%) reported average inter-item correlations. Twenty out of 36 studies (55.6%) reported item-total, or corrected item-total, correlations but none provided an acceptable value or general range of values for evaluating these.

Seven studies (out of 36 studies; 19.4%) examined test-retest reliability; all reported the time interval between administrations. Intervals examined were 1-2 days, 1 week, 2 weeks, 4 weeks, one month, and two months. One study examined both two-week and one-month intervals. Three studies examined a time interval of two months, with two of these studies contained within a single article. Three of the seven studies (42.9%) provided a rationale for the time interval chosen.

The studies that did not provide reliability evidence (10/46; 21.7%) either focused on the internal structure of the SWLS using CFA and/or examined measurement invariance.

3.5 Internal Structure

Thirty-nine out of 46 studies (84.8%) examined internal structure. Of those 39 studies, 12 studies (30.8%) conducted exploratory factor analysis, 23 studies (59.0%) used confirmatory factor analysis methods, three studies (7.7%) used both methods, and one study (2.6%) was unclear as to which approach was used[1].

3.5.1 EFA

Of the 15 studies conducting EFA (i.e., 12 studies using EFA + 3 studies using both EFA and CFA), 10 studies (66.7%) used Principal Components Analysis (PCA), four studies (26.7%) used common factor analysis (FA), and one study (6.7%) did not identify the method used. Of the four studies using FA, three studies (75.0%) used Principal Axis Factoring as the type of extraction method and one study (25.0%) used Maximum Likelihood (ML). No studies stated any criteria up front for identifying the number of factors. Seven of the 15 studies (46.7%) used 'eigenvalues greater than one' as a criterion, four used scree plots (26.7%), and three studies (20%) used a combination of both. 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 reported factor loadings (14/15; 93.3%) but no studies identified a criterion for factor loadings (e.g., >.30, >.35, >.40) to determine if each item loaded on the factor. Because no study reported more than one factor found, other EFA considerations, such as factor rotation, were not explored.

---

[1] The focus of this methodological article was on steps to identify essential unidimensionality that could be used with either EFA or CFA. SWLS data was used in the 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.

3.5.2 CFA

Of the twenty-six studies conducting CFA (i.e., 23 studies using CFA + 3 studies using both EFA and CFA), 24 studies (92.3%) specified the software used for analysis. LISREL 8.0 was the program predominantly used in studies published between 1985 and 2008 (11/14; 78.6%). From 2009 through 2012, software programs identified were Amos, M Plus, EQS 5.7, and SAS. Twenty-four studies (out of 26 studies; 92.3%) specified the number of factors expected. All but two studies (out of 26 studies; 7.7%) specified fit indices used to assess their respective models. For a detailed breakdown of fit indices used across studies, see Table 3. The most commonly used fit indices were the Comparative Fit Index (CFI), Chi-Squared test (χ^2), Root Mean Squared Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). For the fit indices identified, citing the criteria for the range of acceptable values per index varied across the 26 studies: 15 studies (57.7%) provided criteria for all indices used, five studies (19.2%) provided criteria for some fit indices but not others, and six studies (23.1%) provided no criteria. Only one study (3.8%) stated the rationale for the fit indices chosen. Seven out of 26 studies (26.9%) conducted a factor analysis to explore potential two-factor models, two studies (7.7%) examined a modified one factor model (with items 4 and 5 allowed to correlate), and one study (3.8%) examined both a two-factor model with items 1, 2, and 3 loading on one factor and items 4 and 5 loading on a second factor, and a hierarchical/second order model with two factors explained by a general factor.

Table 3 – Fit Indices Used Across Studies

	$\chi 2$	GFI	PGFI	AGFI	IFI	NFI	NNFI/ TLI	RMSEA	SRMR	RMSR	CFI	AIC	CAIC	RMR	CN	$\Delta\chi 2$	$\chi 2$/df	Software	Total per study
Shevlin & Bunting (1994)	1	1		1						1								Lisrel 7.0	4
Lewis, et al. (1995)	1	1		1						1								Lisrel 7.0	4
Shevlin, et al. (1998)	1																	Lisrel 8.0	1
Lewis, et al. (1999)	1	1				1												Lisrel 8.0	3
Pons, et al. (2000)											1					1	1	Lisrel 8.0	2
Atienza, et al. (2003)								1			1						1	Lisrel 8.0	3
Vautier (2004)	1							1	1		1							EQS 5.7	4
Tucker, et al. (2006)		1					1	1			1							not ident	4
Wu & Yao (2006)	1					1	1	1			1	1	1	1				Lisrel 8.0	7
Gouveia, et al. (2008)	1	1					1	1	1		1							Lisrel 8.0	6
Hultell & Gustavsson (2008)								1	1		1							Lisrel 8.0	3
Siedlecki et al. (2008)	1						1	1			1							Not ident	4
Wu & Wu [1] (2008)	1						1	1	1		1							Lisrel 8.0	5
Wu & Wu [2] (2008)	1						1	1	1		1							Lisrel 8.0	5
Wu, et al. [1] (2009)	1						1	1	1		1							M Plus	5
Wu, et al. [2] (2009)	1						1	1	1		1							M Plus	5
Swami, et al. (2009)	1	1	1					1			1	1		1	1			Amos 4.0	8
Anaby, et al. (2010)								1			1							EQS 6.1	2
Durak, et al. [1] (2010)	1				1		1	1	1		1					1	1	Amos 7.0	8
Durak, et al. [2] (2010)	1				1		1	1	1		1					1	1	Amos 7.0	8
Durak, et al. [3] (2010)	1				1		1	1	1		1					1	1	Amos 7.0	8
Glaesmer et al. (2011)		1				1	1	1			1							Amos 6.0	5
Clench-Aas (2011)			1	1			1	1			1							Amos 7.0	5
Bai et al (2011)							1	1			1							M Plus	3
Athay (2012)		1							1		1							SAS 9.2	3
Sancho et al (2012)	1	1						1	1		1							EQS 5.7	5

	χ2	GFI	PGFI	AGFI	IFI	NFI	NNFI/ TLI	RMSEA	SRMR	RMSR	CFI	AIC	CAIC	RMR	CN	Δχ2	χ2/df	Software	Total per study
Total per indices	18	9	2	3	3	2	14	18	11	2	21	2	1	2	1	4	5		

χ^2 = Chi Square, GFI = Goodness of Fit Index, PGFI = Parsimony Goodness-of-Fit Indicator, AGFI = Adjusted Goodness of Fit Index, IFI = Incremental Fit Index, NFI = Normed Fit Index, NNFI = Non-normed Fit Index,, RMSEA = Root Mean Squared Error of Approximation, SRMR = Standardized Root Mean Square Residual, RMSR = Root Mean Square Residual, CFI = Comparative Fit Index, TLI = Tucker Lewis Index , AIC = Akaike's Information Criterion , CAIC = Consistent Akaike Information Criterion, RMR = Root Mean Square Residual, CN = Hoelter's Critical N , $\Delta\chi^2$ = Chi-Square change, χ2/df = Chi-Square change/Degrees of freedom.

Measurement Invariance. Fifteen of 33 studies (45.5%) examined measurement invariance. The majority of comparisons were made across gender and age.

3.6 Relations to Other Variables

3.6.1 Providing a Rationale for Selection of Comparative Constructs, Measures, and Variables

Twenty out of the 46 studies (43.5%) included in this synthesis examined relations to other variables. *The Standards* 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, seven out of 20 studies (35%) provided a rationale for all constructs used and six of 20 studies (30%) provided no rationale. Seven studies (35%) provided a rationale for some constructs but not others. Where the rationale for constructs was not explicitly stated, it was implied in that those constructs were used in previous research or involved the construct of subjective well-being (SWB) (i.e., to explicitly state a rationale would be redundant). Regarding a rationale for measures, no studies provided a rationale as to why they selected the specific measures chosen. With respect to demographic variables used, there were six cases (out of 20 studies; 30%) in which the way numbers were assigned to a given variable could both influence the construct being examined (e.g., categorizing age across 10 year periods as opposed to 20 year periods) and statistical results. None of these studies provided a rationale as to why they structured their demographic variables as they did.

3.6.2 Convergent and Discriminant Evidence

One type of validity evidence is gathered by examining the pattern of relationships between the variable of interest and comparison variables that are considered to be conceptually similar (i.e., convergent evidence). A second type of evidence is gathered by examining the pattern of relationships between the variable of interest and comparison variables where "Discriminant measures may consist of theoretically unrelated constructs (e.g., depression and intelligence) or constructs between which one wants to distinguish (e.g., depression from anxiety)" (Hubley & Zumbo, 2013, p. 15); this is known as discriminant evidence. When providing convergent and discriminant evidence, researchers need to: 1) identify the type of evidence they seek to obtain, i.e., how comparison variables are related to the variable of interest, 2) indicate, in advance of any analysis, an expectation of the direction and relative strength of the relationship that theory or previous empirical research would suggest, 3) interpret both types of evidence in relation to each other, and 4) report reliability evidence for all comparison measures used.

The total number of measures used per study ranged from one to 20 (M=8.3, SD=6.5). Overall, researchers poorly articulated their intentions as to what evidence they intended to provide. Thirteen of 20 studies (65%) did not identify which variables/measures were to be used to provide convergent evidence, and which were to be used to provide discriminant evidence. Three out of 20 studies (15%) suggested all measures used were convergent measures but referred to them as providing concurrent validity. One out of 20 studies (5%) initially indicated that measures used were providing evidence of convergent validity but concluded that some of those measures provided evidence of discriminant validity. One study (5%) did not indicate convergent or discriminant measures/variables but referred to all of the measures generally as

measures of SWB, suggesting their intention was to provide convergent evidence only. Two out of 20 studies (10%) clearly stated their intention to provide convergent evidence only.

Regarding stating in advance an expected relationship among variables, 14 out of 20 studies (70%) did not indicate any expected findings. Five of 20 studies (25%) were vague in that the expected findings were not explicitly stated by researchers but implied in that they were based on findings in previous literature. One out of 20 studies (5%) study clearly identified the direction and relative strength of the relationship they expect to find among the variables chosen. Two of the 20 studies mentioned above appeared to use measures to provide both discriminant and convergent evidence. One of these studies clearly identified their discriminant measure, but mistakenly identified their convergent measures as being indicative of concurrent validity, and the other appeared to use both convergent and discriminant measures but do not make this clear. Neither study interpreted both types of evidence in relation to each other. Additionally, two of the above studies stated the intention to provide criterion validity but, in fact, were providing convergent evidence. In both cases, the studies were included in the overall count reflecting those studies providing convergent evidence.

Eleven out of 20 studies (55%) provided no reliability evidence for the comparative measures used; six studies (30%) provided reliability estimates for all measures used, and two studies (10%) provided reliability estimates for some measures used but not others. Finally, one study (5%) provided reliability estimates but it was unclear whether estimates were based on the study sample.

It should be noted that two studies stated their intention to assess discriminant validity but both studies actually intended to provide known-groups validity evidence. These studies were

not included in the summary counts. The issue of misidentifying evidence will be directly addressed in the Discussion section of this thesis.

3.7 Response Processes

Two of 46 studies (4.3%) examined response processes. One study (2.2%) analyzed participant responses using IRT methods, and one study (2.2%) examined the mean time to complete the scale.

Chapter 4: Discussion

The primary goal of this thesis was to examine methods and procedures that researchers use in the process of validating the SWLS. From this, we aimed to establish a foundation upon which further validation evidence for the SWLS can be built. On a broader level, I also aimed to suggest a framework to organize and examine validation procedures presented in support of measures found across disciplines and journals and also contribute to the small but growing literature on validation synthesis. Thus my intentions are aimed at 1) validation researchers and those individuals who use measures and wish to discern and understand the validation procedures used to support the inferences drawn from test scores, 2) researchers interested in the SWLS, and 3) measurement specialists.

4.1 Framework for Conducting a Validation Synthesis

I created a framework to examine published findings for measures used across disciplines and journals. Using *The Standards* and several resources as guides to good validation practice (e.g., Hubley & Zumbo, 2013; Furr & Bacharach, 2008), I devised a coding sheet that reflects this framework and can be used as a "check-list" of sorts to document the rationale for decisions made in the studies selected, the practices that are conducted, and the completeness of the reporting of that information.

The structure of the coding sheet outlines broad sources of validation evidence, and within each source area is a list of procedures specific to each area. The sources of validity evidence are those outlined in the *The Standards*: (1) test content, (2) internal structure, (3) relations to other variables, (4) response processes, and (5) test consequences. In applying the checklist to an individual study, it allows one to determine the gaps in procedures and reporting that may have occurred (e.g., where researchers may have in essence applied good validation

practice but may not have reported it well or where misunderstandings about validation practice may occur). When applied to more than one study, it allows one to readily synthesize information across a range of studies to glean an overall view of, for example, what areas need further investigation or what areas have been adequately addressed and are consistent (or not) over time, and how procedures used have evolved (or not) over time.

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, 2008; Hogan & Agnello, 2004; Jonson & Plake, 1998; Shear & Zumbo, 2012), the detailed documentation of procedures and rationales involved in validation practice provided in this thesis appears to be the first of its kind and adds to the debate about what researchers believe, understand, or identify to be sufficient evidence to justify a given interpretation and use of a particular measure. 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.

4.2 SWLS Translations and Reliability Evidence

Before examining each of the sources of validation evidence, we considered two other types of information: translations of the SWLS and reliability evidence for the SWLS scores.

Translation. Generally, the process and methods used to create translated versions of the SWLS are not well reported. Specifically, the reporting of guidelines and methodology used for translating the SWLS lack clarity and explanation of terms used to describe translation procedures. Of the studies that identified guidelines used in the translation process, those

proposed by Brislin (1970, 1980, 1986) were most commonly cited. Interestingly, the lack of definitions regarding terminology was often found where guidelines were proposed, suggesting that stating adherence to an established protocol was the only information researchers thought was necessary to report. Furthermore, even within the studies citing Brislin, different procedures were reported for different studies. For example, one study reported conducting only a "back translation" (Swami et al., 2009, p. 28) whereas another reported conducting "independent (blind), back and educated translation" (Neto, 1993, p. 127). Little information is provided about the individuals who conducted the translation and no study appeared to use professional translators. Little use of pilot testing appears to have taken place. With different studies using different methods of translation, and some studies not clearly specifying and defining the translation procedures used, it is difficult to determine whether there is consistency in translation methods across studies, or to comment on the quality of the translations being used.

Reliability. Reliability is a necessary condition for validity and thus worth examining before addressing validation practice. For the SWLS, reliability evidence is documented well and consistently reported across studies. Internal consistency is examined most often. The internal consistency estimate most commonly used was Cronbach's alpha, which shows that classical test theory approaches to reliability still dominate, at least with respect to the SWLS. The weakest area of reporting is that no study clearly stated a criterion for an acceptable reliability estimate. Test-retest reliability studies all provided a test 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 test-retest reliability results. Examining test-retest reliability involves specific decisions regarding the amount of time between the intervals being chosen. With the SWLS, it would be important to choose a time interval length not so short that one might recall one's

responses but, more importantly, not so long that one might anticipate changes to occur in the construct (i.e., satisfaction with life) being examined. Put another way, it is crucial to an assessment of test-retest reliability to be able to determine whether an obtained low stability estimate is more likely due to the measure demonstrating low reliability or the influence of change over time in an individual's satisfaction with life.

Another weakness in terms of reliability evidence is researchers' failure to discuss acceptable values for, the role of, or how to interpret, either inter-item correlations or (corrected) item-total 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 you 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). Very few studies examined inter-item correlations. Three studies provided inter-item correlations in a table with no discussion of their relationship to internal consistency or how to interpret them (i.e., they did not provide any values suggesting what value is needed for an item to be deemed a "good" or a "bad" item). Similarly, three studies provided average inter-item correlations. All concluded their results were acceptable, but none indicated why or what constitutes an acceptable value despite the availability of such guidelines. For example, Clark and Watson (1995) suggest that, for higher order constructs (such as the SWLS), a mean correlation between .15 -.20 is deemed acceptable. For those constructs that are more narrowly defined (e.g., talkativeness), a higher mean inter-correlation (i.e., .40-.50 range) would be needed.

It has been suggested by others (e.g., Netemeyer, Bearden & Sharma, 2003; Clark & Watson, 1995) that the little attention paid to inter-item/average inter-item correlations may be problematic, and that an 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: 1) having a large number of items, 2) having items that are highly correlated, or 3) a combination of the two. Similarly, Cortina (1993) suggests that coefficient alpha is problematic for scales using more than 40 items. In such cases, coefficient alpha may be driven by the number of items, not the correlations among items where the result can be a high internal consistency estimate for a test with items that may, in essence, correlate poorly with one another. Alternatively, because the small number of items comprising the SWLS limits their influence on the value of coefficient alpha, the latter will, in this case, be driven by the inter-item correlations and, therefore, can arguably be considered a more straightforward indicator of internal consistency. Thus, with measures consisting of a small number of items, there is little issue with using coefficient alpha as the number of items will not bias an internal consistency estimate. However, more attention should be paid to inter-item correlations or average inter-item correlations, particularly when examining measures with large numbers of items.

The other problematic area within reliability for reporting 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. Some indication of what values are considered acceptable would be helpful in interpreting the results

presented. As a general rule, low or near zero correlations indicate problematic items (Hubley & Zumbo, 2013). Generally, values of .50 and above are found to be acceptable values (Netemeyer et al., 2003).

The most information is provided 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 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.

4.3 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 content validity or consequences of testing.

Internal Structure. Internal structure is the most common type of evidence examined for the SWLS. The majority of the studies examining internal structure used CFA. The number of factors expected, fit indices used, the criteria for the range of accepted values per indices, and software used for analysis were, overall, well reported. All but two studies reported the type of software used for analysis. The number of fit indices used per study ranged from one to eight. 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 (Tabachnik & Fiddell, 2013). Though there

are a number of indices to choose from, as a general rule, consistency in results across indices indicate a good fitting model (Tabachnik & Fiddell, 2013). However, because what fit indices you use influence the results obtained, it is informative to report a rationale for those indices. Tabachnik and Fiddell (2013) note that where "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 used EFA. Of these, the predominant method used was principal components analysis (PCA) rather than common factor analysis (FA). There appeared to be no association between the time (e.g., in which decade) a study was conducted and the EFA method used. With so few FA studies, it is difficult to make any further conclusions about each of these EFA methods. Therefore, for the remainder of this discussion, we will consider EFA studies as a whole. All EFA studies conducted found evidence to support a one-factor model. Eigenvalues were most commonly used to identify the number of factors, followed by scree plots. A small number of studies provided both. One recommended criterion is to use loadings obtained from a parallel analysis as a standard against which obtained loading values can be compared (Hayton, Allen & Scarpello, 2004). 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 & 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 stated the criterion

used to determine if an item loaded on a factor. This information is important regardless of the number of factors identified as underlying scores on a measure. As well, all studies reported the amount of variance explained by the single factor found, but no studies used 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.

Relations to Other Variables. Validity evidence based on relationships to other variables describes the extent to which there is a relationship between the variable of interest (i.e., scores on the SWLS) and other variables (whether demographic variables or scores from measures or other variables). Just under half of the studies addressed relations to other variables; the majority of comparisons were with conceptually related measures. There is a lack of clarity, however, in terms of what constitutes convergent versus discriminant variables, what is expected in terms of relationships, the appropriate terms to use, and how to evaluate this evidence appropriately or clearly. Although it was fairly clear that convergent evidence was most commonly examined, I could not confidently establish a precise count of how many studies included convergent versus discriminant evidence. Greater inclusion of conceptually unrelated variables is needed, however, as is more comparison between correlations with convergent versus discriminant variables when evaluating evidence.

A clearly stated rationale for why constructs and variables were chosen is generally missing or, at best, very unclear. *The Standards* 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. For

example, if you are examining the relationship between the SWLS and neuroticism, you need to provide a rationale for why you are using the construct of neuroticism as well as state a rationale for the specific measure of neuroticism you have 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 you selected that construct. 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 yrs. (young), 2=50+ yrs. (old) versus 1=20-49 yrs. (young), 2=50-69 yrs. (middle aged), and 3=70+ yrs. (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?).

As noted earlier, "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" (AERA, 1999, 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, 1999, p. 20). It is also noted that these variables "might include intended measures of the same construct or of different constructs" (AERA, 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, necessarily, if the researchers themselves had a clear reason for choosing them. For example, researchers would state the comparisons would be made with "psychiatric symptomology" or "personality factors" but not clearly indicate why or how those constructs are relevant to life satisfaction (Arrindell, 1999; Neto, 1993). Some researchers made mention of relationships to variables without discussing the constructs those variables were designed to capture. An example of a brief and explicit rationale for why a given construct was used for comparison can be found in Howell et al. (2010) and Siedlecki et al. (2008); the former study cites a meta-analysis to support the use of neuroticism as a construct of comparison and the

latter study succinctly cites prior literature and provides a clear argument in support of positive and negative affect as components of life satisfaction.

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* 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 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 encompassed sex, age, marital status, educational level, employment status, monthly income, health insurance, and sociocultural level. For variables common to many studies, it is important to know how and why researchers constructed the variable as they did to determine comparability across studies.

In the studies examined, the distinction between construct and variable was often blurred, making it difficult to discern arguments in support of a rationale for constructs from rationales in support of a variable. For many studies, the rationale for other measures used was implied in that they were, for example, measures of SWB or personality variables found in previous studies that researchers expected to be related. But the rationale behind the specific choice of measures was not identified and no indication was given as to which variables were meant to reflect which construct. Some studies provided a rationale for some variables but not for others and a few studies provided no rationale for the other measures. Only one study offered a rationale justifying all constructs and measures included in their respective studies.

Relations to other variables includes more than just convergent and discriminant evidence. Another type of evidence is gathered by examining the pattern of relationships between test scores and a criterion variable. A criterion may be defined as "an outcome indicator that represents the construct, diagnosis, or behavior that one is attempting to predict" (Hubley & Zumbo, 2013, p. 14). Criterion-related evidence can be described as either predictive or concurrent. Concurrent validity refers to the degree to which test scores are correlated with other relevant variables that are measured at the same time as the primary test of interest, and predictive validity is the degree to which test scores are correlated with other relevant variables that are measured at a future point in time as the primary test of interest (Hubley & Zumbo, 2013; Bacharach & Furr, 2008). Choosing a criterion can be challenging in that there may not be a strong or easily identifiable criterion against which to evaluate your measure (e.g., your construct may theoretically be the first and only one of its kind) (Hubley & Zumbo, 2013). For example, due to the complexity and subjective quality of the construct "life satisfaction", one would be hard pressed to come up with a standard "indicator" that could be applied to any given

individual's satisfaction with their life. It should be noted that where there is a lack of a criterion for a given construct, criterion evidence and convergent evidence are often confused. For example, one study claimed to establish criterion-related validity by correlating life satisfaction with constructs "theoretically linked to this factor in the literature," a definition that applies to convergent evidence (Sancho, 2012, p. 6). In another study, researchers asserted that a given measure, the brief World Health Organization Quality of Life Assessment (WHOQOL-BREF), was a criterion measure but offered no argument as to why it constituted a criterion; in fact, the measure used qualified as a convergent measure (Wu & Wu, 2008). Another study sought to examine the criterion-related validity of the SWLS but the measures used (Portuguese versions of both the General Health Questionnaire (GHQ-12) and the Positive and Negative Affect Scale, translated by the author) qualified as convergent measures (Gouveia, 2008).

One issue that arose with validation evidence for the SWLS was that two studies stated their intention to assess discriminant validity but both studies actually intended to provide known-groups validity evidence. In a known-groups validation study, researchers choose two or more groups that are expected a priori to respond differently to the measure being evaluated based on theory or previous empirical evidence. One evaluates the measure based on whether the expected differences are found. When it is known that two groups differ on a specific construct and these differences are not found in one's study, then the validity of inferences drawn from the measure of interest must be questioned. For example, Laranjeira (2009) examined differences in mean SWLS scores between arthritis patients and university students/health professionals. Though differences were found between the two groups, the researchers provided no theory or prior empirical evidence to support the assumption that each group can be expected to respond differently on life satisfaction.

In addition to researchers' choice of variables/measures and the rationale they provide for their use, a hypothesis should be provided regarding an expectation of how variables are related to the measure being examined and the relative strength of the relationship that theory or previous empirical research suggests. A hypothesis should include a stated expectation regarding both the direction and relative magnitude of the expected relationship between variables and should be stated in advance of analysis. Where the direction of the relationship between the variables chosen and the primary variable of interest provide a conceptual basis for the use of the measures chosen, indicating the relative magnitude that is expected is needed to provide values against which to evaluate the associations found. 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 criterion values as a means to interpret the correlations obtained. In essence, researchers in this area must 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.

One of the greatest obstacles in reviewing these studies was the confusion regarding terminology used to describe validation procedures, particularly regarding relations to other variables. We initially attempted to code according to what the researcher intended given the terms used to describe the evidence presented. But how the researcher defined their intentions did not always reflect how their study was ultimately conducted. For example, in one study the

researchers intended to examine concurrent validity, which they defined as the "the correlation of SWLS with other measures assessing conceptually related constructs", confusing concurrent validity with convergent validity (Durak et al, 2010. p. 419). Terminology often morphed throughout the study; for example, one study stated the intention to establish construct validity in its abstract, which became concurrent validity in its introduction, and then concluded with an analysis of convergent and discriminant evidence in its discussion section (Arrindell, 1991).

The confusion regarding terminology creates two problems. The first problem is that if one were to do a search regarding concurrent evidence of the SWLS, a close examination of the studies would indicate the evidence presented indicated relationships to constructs that are conceptually related, not the degree to which test scores are correlated with outcome or criterion variables that are measured at the same time as the primary measure of interest. Alternatively, if one were to seek studies examining convergent evidence, some studies may be overlooked due to the inaccurate use of terminology. Additionally, studies that accurately described the validation process in the absence of formal terminology (e.g., described the intention to examine variables that consist of theoretically unrelated constructs without specifically using the term discriminant validity) may be missed in a literature search attempting to capture studies relevant to a specific area of validation evidence (i.e., concurrent vs. criterion vs. convergent evidence). On a more pedestrian level, those readers who may not be measurement specialists (or are new to validation procedures) and are reading the literature for educational purposes (e.g., how one approaches test validation in an applied context as opposed to a conceptual framework found in textbooks), will likely find themselves confused by the procedures presented rather than informed by a clarification of subtle concepts that, until an applied example is presented, remain elusive.

Overall, a much greater understanding of what constitutes evidence of relations to other variables; how to conceptualize this evidence and provide a rationale for constructs, measures and variables used; how to describe expected relationships and subsequently evaluate the evidence is sorely needed.

Test content. It is not surprising that content validation is unexamined by researchers. The lack of evidence provided is an excellent way to address the debate as to what type of evidence is necessary in support of inferences generated from test scores. A content analysis of any measure addresses such areas as a clearly defined construct definition, item representation, and the process used to generate and evaluate test items. Evaluation of test content generally involves feedback from experiential experts and subject matter experts. This area of evidence is particularly relevant in the areas of large-scale assessment where, for example, it is imperative that a math test reflects the content it is intended to assess (Furr & Bacharach, 2008). For such a construct as life satisfaction, the procedure to evaluate test content is less clear. No matter how clearly the construct of life satisfaction is defined, how to demonstrate the question of whether the item reflects that construct is not so easily accomplished. It may be difficult to identify who qualifies as an expert on what constitutes life satisfaction. That the SWLS was designed to capture an individual's own judgment leaves the realm of content up to the individual's own perception. In the case of the SWLS, I suggest that establishing test content as proposed by *The Standards* is not relevant in this case; examining how people respond to SWLS items is perhaps a more appropriate indication of whether the SWLS adequately reflects the construct, i.e. the cognitive component of life satisfaction.

Response Processes. The SWLS is intended to capture the judgmental component of life satisfaction (Diener, et al., 1985). Where there is a presumption that individuals being examined

are using an underlying psychological or cognitive process when responding to test items, *The Standards* recommend that "empirical evidence in support of those premises should be provided" (AERA et al., 1999, p. 20). Though Diener (2013, p. 500) generally concludes that a high response rate regarding questions relating to how happy one feels indicates that "people understand the subjective well-being questions and can readily answer them," whether or not participants using the SWLS across samples use a similar cognitive process remains unexamined. Of the two studies that addressed response processes, one of those simply examined the mean time to complete the SWLS, important information that, nonetheless, does not attempt to capture the underlying process used when responding to the SWLS (Laranjeira, 2009). The second study used IRT methods and concluded that "the meaning of the SWLS items diverge both within and between cultures" (Vitterso, 2005, p. 345). For an excellent example for exploring responses processes, see Gaderman, Guhn, and Zumbo's (2011) examination of how children respond to the Satisfaction With Life Scale Adapted for Children (SWLS-C).

Consequences of testing. No studies addressed consequences of testing. Shear and Zumbo (2013) suggest this could be due to the difficulty in finding an example of a framework or study that incorporates consequences of testing into a validation study. Cizek et al. (2008) and others argue that consequences of testing have no place in the process of validation. It is difficult to imagine a context where the SWLS would be used as an indication of something other than an individual's rating of their own life satisfaction or well-being (e.g., where scores on the SWLS would be used as an indication of an individual's compatibility, and thus placement, in a job setting). To our knowledge, the SWLS has not been used in this way. Diener et al. (2013) do suggest that measures of well-being are being considered as a source of information for policymakers, particularly in the areas of health and economics. However, they add the caveat

that there may be "specific instances where life satisfaction measures can help illuminate current policy debates, but being able to tie the scores to factors that bear on policy is essential" (Diener et al., 2013, p. 521).

4.4 Overarching Areas of Concern

Within each source of validation evidence as outlined in *The Standards*, researchers need to provide a rationale where an argument in support of theory has been asserted or where analytical choices are made.

Providing a Rationale for Decisions. The lack of a rationale was pervasive across all studies and within all areas of evidence. Rationales specific to areas of evidence are as follows: for reliability, the area of concern is test-retest reliability where a rationale for the time interval should be provided. EFA analyses require, if needed, a rationale for the extraction method chosen. CFA analyses require a rationale for the number of factors expected and, if more than one factor, which items load on which factor, and fit indices used. Relations to other variables need a rationale for constructs chosen, the specific measures used to represent those constructs, and, where the assigning of numbers to demographic variables potentially alters the construct being examined, why the variable was constructed and categorized as such.

Values Needed for Interpretation. Possibly the weakest area of reporting involved researchers not providing any information as to what are acceptable criterion values, information that is needed to interpret the results obtained. This occurred in all studies and across all areas of evidence. On a broad level, this raises the question as to whether researchers understand the relationship of these values to the area of evidence being examined. On a more specific level, without providing acceptable values for comparison (i.e., context) it is not possible to properly evaluate and interpret the obtained results.

For reliability, where an internal consistency estimate, inter-item or average inter-item correlations, and item total/corrected item total correlations are provided, the criterion for what constitutes an acceptable value is needed to interpret the obtained values. For CFA studies of internal structure, criteria for cut-offs for the fit indices used are needed. For EFA studies of internal structure, criterion values are needed for identifying the number of factors chosen (eigenvalues, % of variance explained) and what constitutes an acceptable factor loading. For relations to variables, the expected relationships between the measure of interest and comparison variables (measures) should be stated in advance of the analysis, and these values referenced when analyzing and discussing results.

4.5 Recommendations and Future Directions

This thesis was directed at researchers conducting validation studies, researchers interested in the SWLS specifically, and measurement specialists. What follows are suggestions relevant to each of these groups.

4.5.1 Researchers Conducting Validation Studies

Validation is "the process of developing and testing the explanation" for inferences drawn from test scores (Zumbo, 2009, p. 69). Without a guiding rationale indicating why you chose the methods you did, the quantitative validation tests you conduct are merely descriptive (Zumbo, 2009). Where statistical procedures are undertaken, criterion values for those procedures are necessary to provide a context in which to interpret obtained results. When examining relations to other constructs and variables, 1) provide a rationale for why you chose the comparison constructs and measures, and, where applicable, how and why you constructed demographic variables as you did, and 2) state the relationships you expect based on prior research or your own theory. Similar to the criterion values for statistical procedures, these are

the only values that provide a context and means to interpret the obtained results. Without a rationale, criterion values, and expected relationships presented a priori, there is no relationship between obtained results and conclusions drawn.

4.5.2 Researchers Examining the Satisfaction With Life Scale

If you are translating a version of the SWLS, more than a simple translation is required. A detailed reporting of translation methods used are needed, as well as input from professional translators. For researchers using a translated version of the SWLS, be aware that there appears to be little reliability and validity evidence in support of those versions and thus they should be used with caution. Response processes is an area that is under-researched. As well, research specifically and clearly examining the SWLS and comparison constructs and variables used to date is needed.

4.5.3 Measurement Specialists

There appears to be a great disconnect between validity theory and validation practices. This is particularly striking in that there is a widely accepted resource, *The Standards,* which, over time, is written to reflect changes in validity theory and the process of validation. Presently, *The Standards* defines five sources validation evidence and provides guidelines to demonstrate what constitutes the validity of inferences drawn from test scores pertaining to each area of evidence. I would suggest that *The Standards* do not provide clear guidelines for researchers and thus may be the reason that practices are poorly reported and techniques appear poorly understood. It is easy to demand that journal editors specify what constitutes adequate reporting of validity evidence, and simple to assert that the lack of evidence presented reflects either a misunderstanding of the procedures required to demonstrate the quality of evidence presented or disagreement as to what constitutes validation evidence overall. If the source itself, i.e. *The*

Standards, written by measurement specialists is unclear, then perhaps measurement specialists need to direct their attention to articulating a clearer understanding of validation procedures, and a user-friendly guide to conducting those procedures. Within measurement programs the importance of integrating quantitative test validation practices with a guiding rationale that supports the methods chosen needs to be emphasized. The results of this study suggest that, at best, this connection is not being made. As well, further instruction in measurement across disciplines, not just within measurement programs, would be useful for all professions and researchers involved in the development, evaluation, and use of tests and measures.

4.6 Strengths and Limitations

Strengths of this study are that first, we sought to examine all peer-reviewed and published validation studies regarding the SWLS. Where other studies used a random sample of studies to examine validation procedures, we have 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. Secondly, we sought to ground our analysis in procedures proposed by *The Standards* (AERA et al., 1999), a widely accepted resource for validation procedures. Third, where past studies have coded according to the broad areas of evidence, as outlined in *The Standards,* we sought to examine each of those areas in detail to identify the specific methods and procedures that researchers use in the process of validating the SWLS. In applying a coding sheet structured on specific procedures proposed by *The Standards* across each area of evidence, we have laid the groundwork for future examination of any given measure across publications. Specifically, we provided a thorough summary of validation procedures used to date regarding the SWLS.

Limitations are that the chosen search criteria rule out any studies that did not use the above stated search terms or those studies where researchers implicitly intended to conduct a validation study but did not explicitly identify it as such. Another limitation is that the level of detail addressed required a fairly high level of understanding of statistical methods used in the analysis of measures may have affected the accuracy of coding those areas. There was a great deal of subjective judgment when determining what types of evidence were provided in regards to relations to other variables, particularly due to the confusion with terminology and lack of a clear framework presented by researchers.

Chapter 5: Conclusion

Reliability is well and consistently reported across studies and populations, suggesting that researchers understand its relationship (necessary but not in and of itself sufficient) to validity. Where statistical results are provided, criterion values for interpretation of these values are lacking. Validation studies for the SWLS focused largely on internal structure and relations to conceptually related variables. Procedures used to examine internal structure consisted of mostly CFA and EFA. Whereas statistical methods and technical procedures are well reported, providing criterion values to interpret results was lacking. Relations to other variables evidence consisted mainly of relationships to conceptually related variables. However, a clear accounting of procedures used and evidence examined is difficult to determine due to: 1) lack of a clear rationale provided for the selection of comparison constructs, measures and variables, 2) confusion regarding terminology, and 3) lack of values/expected values needed for interpreting results.

Evidence regarding response processes is inadequate. Messick (1995) suggested that empirical evidence regarding how and why individuals respond to a measurement task is needed to assure that respondents are actually engaged in the processes one presumes the task is capturing. It is argued that test content validation as outlined in *The Standards* (AERA et al., 1999) is not directly applicable to the SWLS and that examining response processes is a more appropriate means to explore the extent to which the SWLS captures an individual's cognitive judgment of their life satisfaction. While there is no evidence regarding consequences of testing, it is suggested that, at present, the SWLS is not used in contexts other than to determine an individual's judgment of their own level of life satisfaction, and thus that consequences of testing need not be explored. It is further suggested that *The Standards* needs to provide clearer

guidelines regarding procedures needed to present evidence in support of the inferences drawn from test scores.

Bibliography

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

American Psychological Association (2013). *PsycINFO Quick Facts.* Retrieved May 14, 2013 from http://www.apa.org/pubs/databases/psycinfo/index.aspx

An Introduction to meta-analysis. Retrieved July 17, 2012 from http://www.cochrane-net.org/openlearning/html/mod3.htm.

Anaby, D., Jarus, T., & Zumbo, B.D. (2010). Psychometric evaluation of the Hebrew language version of the Satisfaction With Life Scale. *Social Indicators Researcher, 96,* 267-274. DOI 10.1007/s11205-009-9476-z

Arrindell, W., Heesink, J., & Feij, J. (1999). The Satisfaction With Life Scale (SWLS): Appraisal with 1700 healthy young adults in The Netherlands. *Personality and Individual Differences, 26,* 815-826.

Arrindell, W., Meeuwesen, L., & Huyse, F. (1991) The Satisfaction With Life Scale (SWLS): Psychometric properties in a non-psychiatric medical outpatients sample. *Personality and Individual Differences, 12,* 117-123.

Athay, 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.

Atienza, F., Balaguer, I., & Garcia-Merita, M. (2003). Satisfaction With Life Scale: Analysis of factorial invariance across sexes. *Personality and Individual Differences, 25,* 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*. DOI: 10.1177/1090198113483139

Biddle, D. (2010). Should employers rely on local validation studies or validity generalization (VG) to support the use of employment tests in title VII situations? *Public Personnel Management, 39,* 307-326. DOI: 10.1177/009102601003900402

Cizek, G., Rosenberg, S., & Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397-412.

Cizek, G., 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.

Clark, A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Clench-Aas, J., Nes, R., Dalgard, O., & Aaro, L. (2011). Dimensionality and measurement invariance in the Satisfaction With Life Scale in Norway. *Quality of Life Research, 20,* 1307-1317.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Sage.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step by step approach*. Thousand Oaks, CA: Sage Publications Inc.

Diener, E., Emmons, R., Larsen, R., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49,* 71-75.

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research, 112,* 497-527. DOI 10.1007/s11205-012-0076-y

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

Fleishman, J., & Benson, J. (1987). Using Lisrel to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47,* 925-939.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Gaderman, A., Guhn, M., & Zumbo, B.D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction With Life Scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100,* 37-60.

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., Milfont, T., Nunes da Fonseca, P., & Pecanha de Miranda Coelho, J. (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., Allen, D., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7,* 191-205.

Hogan, T., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64,* 802-812.

Howell, R., Rodzon, K., Kurai, M., & Sanchez, A. (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. (2013). Psychometric characteristics of assessment procedures: An overview. In K.F. Geisinger (ED.), *APA handbook of testing and assessment in psychology.* Washington, DC: American Psychological Association.

Hultell, D., & Gustavsson, J. (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., & Plake, B. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736-755.

Kane, M. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37,* 76–82.

Kveton, P., Jelinek, M., Klimusova, H., & Voboril, D. (2007). Data collection on the internet: Evaluation of web-based questionnaires. *Studia Psychologica, 49,* 81-88.

Laranjeira, C. (2009). Preliminary validation study of the Portuguese version of the Satisfaction With Life Scale. *Psychology, Health & Medicine, 14,* 220-226.

Ledesma, R., & 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 & Evaluation, 12,* 1-11.

Lewis, C., Shevlin, M., Bunting, B., & 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., Shevlin, M., Smekal, V., & Dorahy, M. (1999). Factor structure and reliability of a Czech translation of the Satisfaction With Life Scale among Czech university students. *Studia Psychologica, 41,* 239-244.

Lyons-Thomas, J., Liu, Y., Olivera, O., & Zumbo, B.D. (2012, April). *Meta-synthesis of studies in 'When Validity Theory Meets Validation Practices' with an eye toward comparing and contrasting the seven research syntheses.* Poster presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia.

Manten, A. (1973). Scientific literature review. *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, 3*7, 113-115.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Mulrow, C. D. (1994). Systematic Reviews: Rationale for systematic reviews. *BMJ*, *309,* 597-599. DOI:10.1136/bmj.309.6954.597

Navratil, M., & Lewis, C. (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.* California: 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., & 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.

Pavot, W., & Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment, 5,* 164-172.

Pons, D., Atienza, F., Balaguer, I., & Garcia-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.

Sancho, P., Galiana, L., Gutierrez, M., Francisco, E., & Tomas, J. (2012). Validating the Portuguese version of the Satisfaction With Life Scale in an elderly sample. *Social Indicators Research.* DOI 10.1007/s11205-012-9994-y

Shear, B. & Zumbo, B. (2012, April). *What counts as evidence? An empirical review of validity studies in educational psychological measurement.* Poster presented at the annual

meeting of the American Educational Research Association, Vancouver, British Columbia.

Shear, B. & Zumbo, B.D. (2013). *What counts as evidence: A study in validity studies.* Unpublished manuscript, Department of Educational and Counselling Psychology and Special Education, University of British Columbia, Vancouver, Canada.

Shevlin, M. & Bunting, B. (1994). Confirmatory factor analysis of the Satisfaction With Life Scale. *Perceptual and Motor Skills, 79,* 1316-1318.

Shevlin, M., Brunsden, V., & Miles, J. (1998). Satisfaction With Life Scale: Analysis of factorial invariance, mean structures and reliability. *Personality and Individual Differences, 25,* 911-916.

Siedlecki, K., Tucker-Drob, E., Oishi, S., & Salthouse, T. (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.

Slocum-Gori, S., Zumbo, B.D., Michalos, A., & 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.

Tabachnik, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.) Boston, MA: Pearson Education.

Tucker, K., Ozer, D., Lyubomirsky, S., & Boehm, J. (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.

Vitterso, 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., & Maritz, C. (2003). Empirical testing of the Satisfaction With Life Scale: A South African pilot study. *Psychological Reports, 91,* 551-554.

Wu, C., & 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., & Wu, C. (2008). Life satisfaction in persons with schizophrenia living in the community. *Social Indicators Research, 85,* 447-460. DOI 10.1007/s11205-007-9136-0

Wu, C., Chen, L., & Tsai, Y. (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

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 ON: 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, NC: IAP Information Age Publishing, Inc.

Appendices

Appendix A - The Satisfaction With Life Scale

The Satisfaction With Life Scale

Below are five statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

- 7 - Strongly agree
- 6 - Agree
- 5 - Slightly agree
- 4 - Neither agree nor disagree
- 3 - Slightly disagree
- 2 - Disagree
- 1 - Strongly disagree

____ In most ways my life is close to my ideal.

____ The conditions of my life are excellent.

____ I am satisfied with my life.

____ So far I have gotten the important things I want in life.

____ If I could live my life over, I would change almost nothing.

Appendix B - Coding Sheet

Project: Validity Synthesis SWLS

Author:
Year:
Title:
Journal:

Sample Description:

Reliability

** based on this sample only*

Internal consistency:

R1	Reported an internal consistency reliability estimate	Y N Unclear
R2	▪ If yes, type of int. consist. estimate (e.g., Cronbach's alpha, split half (odd/even), KR-20, ordinal alpha)	Cronbach's alpha Other:
R3	▪ If yes, criterion used for int. consist. is clearly referenced	Y N Unclear
R4	▪ If yes, describe criterion (e.g., >.80, >.90; citation):	
R5	Reported inter-item correlations	Y N Unclear
R6	Reported average inter-item correlation	Y N Unclear

Test-retest reliability:

R7	Reported a test-retest reliability estimate	Y N Unclear
R8	▪ If yes, reported the retest interval used	Y N Unclear
R9	▪ If yes, describe retest interval (e.g., 1 week, 3-41 days): 1 week	
R10	▪ If yes, provided rationale for choice of retest interval	Y N Unclear

Other:

R11	Presented item-total (or corrected item-total) correlations	Y N Unclear
R12	▪ If yes, interpreted these correl. as int. consist./reliab. evidence	Y N Unclear

IRT reliability information:

R13	Reported test information function (TIF) – not IIFs	Y N Unclear
R14	Reported conditional standard error of measurement (CSEM)	Y N Unclear
R15	▪ Other; describe:	

Internal Structure

Internal Consistency as Dimensionality:

IS1	Internal consistency evidence interpreted as evidence of dimensionality	Y N Unclear

General Info:

IS2	Conducted a factor analysis	Y N Unclear
IS3	▪ If yes, type of factor analysis	EFA CFA Both

Exploratory Factor Analysis:

IS4	Type of EFA	PCA FA Both
IS5	▪ If FA, was the extraction method (see below) identified?	Y N Unclear
IS6	▪ Describe extraction method (e.g., PAF, ML, ULS, GLS):	
IS7	Criteria stated for identifying number of factors stated	Y N Unclear
IS8	▪ Used criterion of eigenvalues > 1	Y N Unclear

Author:
Year:
Title:
Journal:

IS9	▪ Used criterion of scree plot	Y N Unclear
IS10	▪ Used parallel analysis (PA)	Y N Unclear
IS11	▪ Describe software/program used:	
IS12	Used criterion of % of variance explained	Y N Unclear
IS13	Used another criterion; describe:	Y N Unclear
IS14	Reported % of Variance Explained	Y N Unclear
IS15	Factor loadings reported	Y N Unclear
IS16	Criterion for factor loadings reported	Y N Unclear
IS17	If yes, describe criterion used (e.g., >.30, >.35, >.40):	Y N Unclear
IS18	▪ More than 1 factor found	Y N Unclear
IS19	If yes, factors were rotated	Y N Unclear
IS20	▪ If yes, type of rotation method used	Y N Unclear
IS21	▪ If orthogonal, specific method used	Oblique Orthogonal Both Unclear
IS22	▪ If oblique, specific method used	Varimax Quartimax Other:

Higher Order / Hierarchical Factor Analysis:

IS23	Conducted a FA to examine if total score & subscale scores apply	Y N Unclear
IS24	▪ Describe:	

Notes:

Confirmatory Factor Analysis:

IS1	Specified software used	Y N Unclear
IS2	▪ Describe software:	
IS3	Specified the number of factors expected	Y N Unclear
IS4	▪ If >1 factor, specified which items load on which factors	Y N Unclear
IS5	Specified fit indices (e.g., NFI, RMSEA) used	Y N Unclear
IS6	▪ List fit indices used:	
IS7	▪ Provided criteria / cut-offs for fit indices	Y N Partially Unclear
IS8	▪ List criteria used for each:	
IS9	▪ Provided rationale for fit indices selected	Y N Partially Unclear

Measurement Invariance:

IS10	Examined measurement invariance	Y N Unclear
IS11	▪ Describe:	

Factor Structure and Scoring:

IS12	Factor structure informed / matched scoring	Y N Partially Unclear
IS13	▪ Comments:	

Author:
Year:
Title:
Journal:

Relations to Other Variables

• How did researcher describe this process (e.g., relations with other constructs/variables, use of terms convergent, discriminant, concurrent, construct validity)?
• Was it clear what the researcher expected (e.g., how the measures should correlate or what they saw as converg. vs. discrim. evidence/measures)?
• Whether they used the terms or not, did they seem to include both converg.-like and discrim.-like measures?
• Did the researchers provide any rationale for why particular constructs or measures were selected?
• Did the researchers report reliability evidence based on this sample for the (conv./discrim./other) measures used?
• How did the researchers know if the evidence supported validity or not (e.g., they didn't really explain it, they seemed to base this on the stat. signif. (or not) of correlations, on the magnitude of the correl., on the sign (pos./neg.) of the correl.)?
• Did they use some other procedure like a multitrait-multimethod (MTMM) matrix or a FA to determine if converg. and discrim. evidence load on different factors?
• Did they provide some other kind of evidence (describe)?
• Notes/comments:

Response Processes

Things to look for:

- Questioning or probing responding to items (e.g., think-aloud protocols, cognitive interviewing)
- Documenting or recording responses to items
- Recording time to complete individual items or measures
- Post-test questionnaire or interview
- Tracking eye movements (e.g., on computer)

Describe:

Author:
Year:
Title:
Journal:

Consequences

Things to look for:

- Use of the words "consequences", "consequential validity", "effects of", "impact of", "implications", "clinical implications" (summarize or copy/paste relevant info)
- Consequences misunderstood as test misuse
- Citations related to consequences (e.g., AERA Standards, Messick, Kane)

Describe:

Test Content

Things to look for:

- construct was clearly defined
- 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 experiential experts and/or a target population (i.e., experiential experts)
- subject matter experts (SMEs) or experiential experts (EEs) were consulted to examine elements of the measure
- reference was made to item representation (, i.e the degree that items in the measure represent the full range of the construct of life satisfaction), construct underrepresentation, and construct irrelevant variance.

Translation and Adaptation of Measures

T1	No information but suspect translated version of measure used	Y	N/A		
T2	Used previously translated version of measure	Y	N	Unclear	

T3	Used newly translated/adapted measure	Y	N	Unclear	
T4	▪ Citations provided for translation method or guidelines	Y	N	Unclear	
T5	▪ Described methodology (e.g., forward or double-back transl.) used	Y	N	Partially	Unclear
T6	▪ Describe:				
T7	▪ Described qualifications of translators	Y	N	Partially	Unclear
T8	▪ Describe:				
T9	▪ Conducted pre-tests/pilot tests	Y	N	Unclear	
T10	Other comments:				