Abstract

The Rosenberg Self-Esteem scale (RSES) is a commonly used measure, cited over 3,000 times in the past five years. The aim of this study was to produce a synthesis of the available sources of reliability and validity evidence for the RSES as classified by the Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014). Despite the popularity of the RSES, only 27 articles have examined reliability and validity evidence for the scale. This study showed that the most prevalent source of reliability is based on internal consistency and the most prevalent validity evidence is based on internal structure, followed by relations to other variables. The latter source of evidence primarily consisted of convergent validity evidence. Evidence based on response processes is seldom examined and no studies examined validity evidence based on content or consequences of testing. When examining reliability, internal structure, and relations to other variables, studies tended to overlook the implications of the order in which these factors are studied. There is also a need for researchers to clearly state assumptions and criteria to interpret findings as well as more clarity in the reporting of results. The implications of these findings for researchers interested in the use of the RSES and for measurement experts will be discussed.
Preface

This dissertation is original, unpublished, independent work by the author, M. Coronel Villalobos.
# Table of Contents

Abstract .......................................................................................................................... ii
Preface ............................................................................................................................ iii
Table of Contents ............................................................................................................ iv
List of Tables ................................................................................................................... vii
List of Figures ................................................................................................................ viii
Acknowledgements ....................................................................................................... ix
Chapter 1 – Introduction ............................................................................................... 1
Chapter 2 – Literature Review ....................................................................................... 3
  Validity ........................................................................................................................... 5
    Content ....................................................................................................................... 6
    Response processes .................................................................................................. 8
  Internal structure ....................................................................................................... 11
  Relationships with other variables .......................................................................... 13
  Consequences of testing .......................................................................................... 17
Research and Validation Synthesis ............................................................................ 19
  Reliability synthesis ................................................................................................. 23
  Validation synthesis ................................................................................................. 25
Study Purpose ............................................................................................................... 30
Chapter 3 – Article ....................................................................................................... 31
  Abstract ....................................................................................................................... 31
  Background ................................................................................................................. 32
    Self-esteem and the Rosenberg Self-Esteem Scale ................................................... 32
  Sources of reliability and validity evidence ............................................................. 33
  Reliability and validation synthesis synthesis ......................................................... 35
Study Purpose ............................................................................................................... 38
Method ......................................................................................................................... 38
  Database search ........................................................................................................ 38
  Reliability and validation synthesis coding ............................................................. 40
  Evidence based on content ....................................................................................... 41
  Evidence based on response processes ................................................................... 41
| Contributions to Reliability and Validation Syntheses | 82 |
| Contributions to Researchers Using the Rosenberg Self-Esteem Scale | 84 |
| Contributions to Measurement | 86 |
| References | 88 |
| Appendices | 104 |
| Appendix A: List of Articles Included in the Study | 104 |
| Appendix B: Rosenberg Self-Esteem Scale | 109 |
| Appendix C: Coding Sheet | 110 |
List of Tables

Table 1  Translation and sources of reliability examined ........................................46
Table 2  Sources of validity examined .......................................................................47
Table 3  Sources of reliability and validity by language version ............................56
Table 4  Order in which evidence was examined ................................................58
List of Figures

Figure I Flowchart for selection of relevant articles .............................................40
Acknowledgements

I would like to thank my supervisor Dr. Anita Hubley for her constant guidance and advice, and helping me on every step of the way.

I also wish to extend my thanks to the rest of my thesis committee: Dr. Bruno Zumbo and Dr. Amery Wu, for the insight and comments during my examination as well as their questions that challenged and incentivized me to push my limits.

My sincere thanks also go to Sneha Shankar, who assisted with the coding check.

I thank Dr. Sheila Marshall for her invaluable insight and suggestions.

I also would like to thank my family, without whom I would not have been able to make it this far. They have supported me morally, financially, emotionally and spiritually both through my thesis and my life.

I also thank Ilyana Mansfield and Jason Riles, who helped with some proofreading and make me feel at home.

Last but not least, I would like to thank Victor Diaz for his support and companionship.
Chapter 1 – Introduction

The Rosenberg Self-Esteem Scale (RSES) has been used very commonly ever since its introduction in 1962 (Blascovich & Tomaka, 1991). A search in PsycINFO shows that, in the period between January 2010 and January 2015, it has been cited over 3,000 times. Reliability for the RSES scores and validity for inferences based on those scores have been examined over several studies, and the factor structure for the RSES has been under some debate. It is the purpose of this study to present a synthesis of the available findings for the various sources of reliability and validity evidence as classified by The Standards for Educational and Psychological Testing (AERA et al., 2014). This study aims to provide a useful contribution to researchers interested in the RSES by examining the available sources of evidence and gaps in the literature for reliability and validity evidence on the RSES. It also aims to contribute to measurement experts by presenting the trends in methodology for the examination of sources for reliability and validity evidence across the life of a commonly used psychological measure, the RSES. Finally, this study aims to expand upon the methodology of reliability and validity synthesis, which will be described in detail below.

Reliability and validity are an integral part of measurement, as they allow researchers to provide meaningful interpretations to observed scores in various contexts (AERA et al., 2014; Hubley & Zumbo, 2013; Messick, 1995a). Given the popularity of the RSES as a measure, it would be expected to find a large number of publications examining reliability and validity evidence. Given the 50+ years since
publications of the RSES and the amount of available information, a reliability and validation synthesis of findings becomes a helpful guide for researchers.

The type of research synthesis used in this study is known as a reliability and validation synthesis, as described in the book “Validity and Validation in Social, Behavioral, and Health Sciences” (Zumbo & Chan, 2014). This methodology is a recent development in the area of measurement that builds upon research syntheses with a focus on the various available sources of reliability and validity evidence, and an examination of trends in reporting practices in research.

In chapter 2, I will describe the RSES and the construct of self-esteem, followed by an explanation of different sources of reliability and validity as described by The Standards for Educational and Psychological Testing (AERA et al, 2014). Subsequent to that description, I will describe research syntheses and focus on a specific type of research synthesis, which is a reliability and validation syntheses.

Chapter 3 will consist of the thesis study presented in article format with a brief introduction, a description of the methods used for the reliability and validation synthesis of the RSES, a report of the findings, and a short discussion.

In chapter 5, I will present an extended discussion of the results that also includes a description of the contributions of this paper to researchers using the RSES and to measurement experts interested in the sources of reliability and validity evidence, the strengths and limitations of the study, and directions for future research.
Chapter 2 – Literature Review

The purpose of this study is to conduct a reliability and validation synthesis in which I will examine the sources of evidence that have been used to evaluate the validity of inferences made from the Rosenberg Self-Esteem Scale (RSES). The 10-item RSES was created for the purpose of measuring what was then the new construct of self-esteem in adolescents (Rosenberg, 1965). Rosenberg defined self-esteem as a person’s evaluation of their own worth, a concept that was viewed as mostly established in adults but fluctuated in adolescents. Self-esteem is the attitude toward oneself, or the affective evaluation of the self-concept (Blascovich & Tomaka, 1991; Rosenberg, 1965).

Originally, the RSES was developed using Guttman scaling; it was intended to be a unidimensional measure of self-esteem aimed at high school students. This version of the RSES was used by some researchers, such as Silber and Tippett (1965) and Crandall (1973). However, in other studies, researchers adapted the scale to be used with a Likert-type response format (Morgan, 1981; Wallace, 1988) and, in 1989, Rosenberg formally revised the scale using a Likert-type response format. The four-point Likert-type response format ranges from strongly disagree to strongly agree. The RSES also has been adapted to have a 5 or 7 point Likert-type response format (Blascovich and Tomaka, 1991; Darity, 2008), and has been translated into several languages (e.g. Baños & Guillén, 2000; Martin, Thompson & Chan, 2006; Shapurian, Hojat, & Nayerahmadi, 1987). The scale has been used with a variety of samples that differ in age and ethnicity (Baños & Guillén, 2000; Blascovich and Tomaka, 1991; Classen, Velozo & Mann, 2007; Halama, 2008;

Self-esteem has been used widely in social and psychological research (Darity, 2008). It is a useful construct that can be related to other constructs such as personality, anxiety, behavioral patterns, depression, distress, and productivity; it has also served as a predictor for coping ability, sociability, and chance for success (Heatherton & Wyland, 2003; Macarthur, n.d.).

Currently, there are other alternative measures for self-esteem, such as the Coopersmith Self-Esteem Inventory (Coopersmith, 1967), the Janis-Field Feelings of Inadequacy Scale (Janis & Field, 1959), Texas Social Behavior Inventory (Helmreich & Stapp, 1974), the Social Self Esteem Scale (Ziller, Hagey, Smith & Long, 1969), Piers-Harris Children's Self-Concept Scale (Piers, 1976), Self-Perception Profile for Children (Harter & Pike, 1983), and Tennessee Self-Concept Scale (Fitts & Warren, 1996); however, these scales have not seen as much use or tend to be focused on a more narrow population or construct, and are not as popular as the RSES (Blascovich & Tomaka, 1991).

Due to the widespread use of the RSES, it is important to have strong evidence of validity and an understanding of the sources of validity that have been gathered to support inferences made from this measure. Validity is required to give scores on a measure significant meaning (Messick, 1995a) and determine whether degree or level of self-esteem is the most reasonable interpretation of scores from the RSES. Moreover, validity evidence can provide an explanation for score variation beyond descriptive values (Zumbo, 2009).
Validity

Validity is defined by *The Standards for Educational and Psychological Testing*\(^1\) as “the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests” (AERA, APA & NCME, 2014, p. 11). *The Standards* describes this process as accumulating evidence from relevant sources in a conceptual framework. The process of validity, that is, validation, has been noted to strengthen and improve the process of assessment and to help one to have a better understanding of a hypothesis’ fit to the predicted model, original theory, and observations (Smith, 2005).

*The Standards* describes the criteria that should be met to ensure proper testing, such as the need for the test developer to be clear about the intended use and interpretation of the measure, reporting the conditions of data gathering for statistical analysis of test results in detail, or, if the use of a test results in unintended consequences, an attempt should be made to investigate if it happened because of characteristics of the particular sample or failure of the test to correctly represent the construct (AERA et al., 2014).

*The Standards* (AERA et al., 2014) describes validity based on its source of evidence: content, response processes, internal structure, consequences of testing, and relationship to other variables; the latter source is further subdivided into convergent and discriminant relationships, test-criterion relationships, and validity generalization.

---

\(^1\) Henceforth referred to as “*The Standards*”
**Content.** Validity based on the test content refers to how the construct intended to be measured is related to the scale that is assessing the said construct. Content validity was among the first sources of validity to be defined and recognized by *The Standards* in 1954 (APA, 1954). The definition and the methods used to support content validity have changed as validity theory has been revised (Jonson & Plake, 1998; Sireci & Faulkner-Bond, 2014). The 1985 version of *The Standards* regarded content validity as evidence of a unified validity concept along with criterion and construct evidence. By 1999, *The Standards* presented validity based on test content as a separate category of validity evidence with specific standards. Validity evidence based on content deals with the meaning and interpretation of a test, its representation, and the extent of its pertinent domain - which may be seen as the definition, relevance and appropriateness of the measure’s construct (Sireci & Faulkner-Bond, 2014).

Validity evidence based on content can be found in the operationalization of the construct, the theoretical basis for the creation of the measure to evaluate the construct and its adequate adaptation from construct to measure. It is supported by the relevance and representativeness of a test’s items to its intended construct or constructs, and is strengthened by clear and consistent construct definitions (Haynes, Richard & Kubany, 1995). It can be observed in the themes, wording and formatting presented, although it also is reflected in the administration and scoring process such as the adequacy of using a particular method to observe the intended construct. Haynes et al. (1995) noted various elements of content validity such as the sequence of items, scoring method, the precision of wording of items, and the
instructions given to participants; depending on the method of assessment (e.g. self-report, behavioural observation, psychophysiology and self-monitoring) these elements differ in relevance—for example, the sequence in which items are presented will be highly relevant in self-report scales but not as relevant in behavioural observations.

Two threats to validity are construct-irrelevant variance and construct underrepresentation (AERA et al., 2014; Furr & Bacharach, 2014; Messick, 1995b). As an example, in the RSES, the construct to be measured is self-esteem; adding items that measure depression such as "I feel sad all the time", or writing the items in complex language that participants have trouble understanding would add construct irrelevant variance, while removing the negative self-worth items would lead to construct underrepresentation because it would fail to include the entirety of the construct. As a way to provide validity evidence based on content, a group of experts can be asked to assess the appropriateness of the measure, and rate the representativeness and relevance of items as well as their possible interpretation. Subject matter experts (SMEs) are generally people who have training, professional experience, or academic education that qualifies them in a specific subject or area (Grant & Davis, 1997). For example, SMEs asked to judge the validity of inferences made from scores on the RSES based on content could be scholars that have studied self-esteem in depth. Another type of expert is an experiential expert, which is someone with personal life experience in the subject, such as a pregnant woman in a study of discomforts during pregnancy. The use of such experiential experts has become more common over time. Another option evaluating content validity is the
use of alignment methods to match items to a standard (AERA et al., 2014; Sireci & Faulkner-Bond, 2014).

Validity evidence based on content affects the analysis of causal relationships, diagnosis, estimations and parameters of behavior and estimates of the effects of treatment (Haynes et al., 1995, p. 240). That is, a scale whose construct is not properly represented and contains construct-irrelevant items could be measuring more than one construct such as when high scores on a depression scale are obtained by people with high levels of stress rather than depression; similarly, the same depression scale with an underrepresented construct could fail to detect an individual with signs of depression. Thus, if someone used this hypothetical depression scale in a correlational study examining depression and sleeping disorders, the inferences drawn from this study might be wrong because depression was not measured accurately. This is important for scales that can be used clinically (such as a personality or anxiety scale) as well as social, educational and developmental scales. The RSES has seen use as a clinical research measure in some populations such as people with disabilities or diseases like cancer (Curbow & Somerfield, 1991; Davis, Kellett & Beail, 2009; Martin, Thompson & Chan, 2006), and that adds to the need for this measure to have solid validity evidence based on content.

**Response processes.** The evidence for validity of inferences based on response processes is found within the moment of assessment. It is different from content validity as it does not pertain to the actual instrument per se but to the process of response within the respondent when interacting with the measurement
item. It was first included in *The Standards* in 1999; the prior version in 1985 included the individual process as part of construct validity (Padilla & Benítez, 2014). Validity evidence based on response processes can be seen in asking participants about their reasoning for responding in a certain way, evaluating their performance, observing strategies used, or timing their response rates. In recording the process of the participant’s performance and cognitive processes, one can find support for interpretation and function of the items beyond the process of constructing the instrument, and with interviewing or focus groups, a researcher can detect patterns that arise during the assessment. Tracking eye movement, recording response times, or conducting cognitive interviews are other methods of measurement that may be used, and can be interpreted as well. For example, long delays might indicate that participants are having trouble understanding the task, a subgroup of the sample could be misinterpreting an item, or the available responses do not match what the subjects want to answer. These methods support the validity of inferences when the recorded behaviour matches the expected behavior of the participants for a given construct (AERA et al., 2014). A similar practice can be applied to examiners or scorers by examining if the scoring method was followed properly or if the interpretation of behaviour reflected in scores was appropriate.

As the RSES is a self-report measure that does not assess skill or does not involve some degree of subjectivity or decision-making on the part of the administrator or scorer, it is likely that possible reports of validity based on response process will not include evaluation of the rater, but they can include subject observation, the use of a think aloud protocol, or cognitive interviews to compare the
rationale or process that the subjects use when answering the RSES to the intended theory of self-esteem. Interview and think aloud methods can be also used to assess if the subjects understood the instructions easily and clearly and what strategy they used to respond, such as guessing, or reflecting on the item’s task (Padilla & Benítez, 2014). While it is also possible to measure response times and tracking eye movement, these methods might not yield as much information about this particular scale. Long response times and wandering eye movement could be caused by subjects having trouble understanding or performing the task, lack of interest or distractibility; unlike interview methods, the measurement of a person’s behaviour while responding to test items can be ambiguous and requires to be interpreted in order to draw conclusions (Padilla & Benítez, 2014). The biggest limitation to establishing validity evidence based on response process for the RSES is the small number of items and short average time it takes to complete the questionnaire, as the amount of observable behaviours during this period would be limited. A review of the literature did not find any behaviours or cognitive responses expected during the completion of the RSES and, as such, any research on validity based on response process for this scale will probably be based on the original theory proposed by Rosenberg. Rosenberg (1965) saw self-esteem as an affective self-evaluation of worth, a process that changed during adolescence but solidified at adulthood. The items he presents aim for the reader to understand the item, and then agree or disagree on how closely the item relates to their perception of themselves.
Internal structure. Validity evidence based on internal structure refers to the relationship between test items or components and the latent variable that they intend to measure (AERA et al., 2014). Reliability, invariance of the measure, and its dimensionality are the three main components of validity based on internal structure (Rios & Wells, 2014). The dimensionality of a scale determines how many latent variables are being measured; test developers often use factor analysis to evaluate the dimensionality of a measure and to observe the associations between multiple factors on a multidimensional test (Furr & Bacharach, 2014). Item invariance indicates whether some subgroups of a sample behave differently on particular items; differential item functioning can be observed on groups with comparable overall skill or status, such as men responding to items on a questionnaire significantly differently from women, or members of immigrant minority population responding differently than the majority population on an item that deals with a local custom. Similarly, scale invariance observes if the measure yields equal or similar scale or subscale results across groups with different characteristics such as race or sex (Rios & Wells, 2014). Measurement invariance can be classified as configural, weak, strong, and strict invariance (Wu & Zumbo, 2007). Configural invariance is when the same factor model shows the same good fit for both (or all) groups. This holds true in all levels of invariance; in addition, weak invariance constrains factor loadings to be equal, strong invariance constrains factor loadings and intercepts to be equal, and strict invariance constrains factor loadings, intercepts and residual variances to be equal. When a model shows fit at these levels, factor loading constraints imply that the factors should have a similar or equal weight in measuring
the construct for all groups, intercept constraints imply that all groups should have a similar or equal mean scores and should obtain similar values when all factors are the same, and residual variance constraints imply that random variation and measurement error should not significantly modify a subject’s score regardless of group membership.

Reliability is the degree to which measurement scores stay consistent or are repeatable (Hubley & Zumbo, 2013). Reliability of a measure is affected by the difference between a subject’s observed score and their true score (Furr & Bacharach, 2014); larger error in the measure will cause less consistent scores and thus lower reliability. Common methods of assessing reliability are internal consistency (Cronbach’s alpha coefficient), test-retest, and inter-rater reliability. Reliability, as a component of validity evidence based on internal structure, assesses the variability of scores that originate from differences among respondents on the construct of interest rather than random measurement error (Hubley & Zumbo, 2013); scores intended to measure the same construct should yield higher reliability estimates, systematic variation in scores that should perform similarly suggests problems with the internal structure of a test (Cook & Beckman, 2006).

There are some statistical approaches that can provide validity evidence based on internal structure, such as exploratory, confirmatory, and multi-group factor analysis. Other methods also include differential item functioning, and inter-item correlation (Rios & Wells, 2014). Whatever the method, the prevalent idea is finding how the measure can be broken down into basic components, understanding the
way it is assembled and provides consistency to the measurement, and examining the invariance of the measure or its items.

The RSES studies are likely to report reliability evidence along with item invariance. Additionally, they will be highly likely to report factor analyses and studies on inter-item correlation as the dimensionality of this scale has often been debated (Goldsmith, 1986; Huang, & Dong, 2012; Mannarini, 2010; Richardson, Ratner, & Zumbo, 2009; Tomás, & Oliver, 1999; Zimprich, Perren, & Hornung, 2005).

**Relationships with other variables.** For evidence based on the relationship with other variables, it is necessary to observe how the test scores relate to external elements such as scores from other scales or constructs. The presence of associations that should or should not exist as per theory supports validity evidence based on other variables (Furr & Bacharach, 2014).

The sources that can influence an instrument’s validity based on other variables are divided into convergent and discriminant, test-criterion, and validity generalization by *The Standards* (AERA et al., 2014). The first sources are convergent and discriminant validity. Convergent validity is how closely scores on a measure are correlated to scores on another theoretically similar measure. Discriminant validity is the presence of little to no correlation between the scores on the measure and the scores of a second measure that evaluates a different construct. Validity evidence is supported when the theoretically expected correlations match the observed correlations from a scale. It is possible to measure both convergent and discriminant validity evidence separately, but when they are
observed together, it is easier to evaluate the pattern of relationships and judge the magnitude of the observed convergent and discriminant correlations (AERA et al., 2014; Furr & Bacharach, 2014). For example, a depression scale would be expected to correlate highly with other depression scales, highly with anxiety measures but lower than with other depression scales, and much lower with an intelligence test.

The purpose of measuring convergent and discriminant validity is to ensure that the test is measuring the targeted construct and not a similar construct. If the test is correctly measuring the intended construct, correlations should be stronger with tests that measure the same construct, lesser correlations with tests that measure different but related or similar constructs, and little to no correlation with tests that measure entirely different constructs (Messick, 1995a). The theory that supports the construct dictates the constructs that are relevant to compare (Messick, 1995b).

While it would be easy to find discriminant constructs completely unrelated to the measure of interest, observing the relationships among constructs on a continuum of relatedness can be compared more easily against the theory of the construct and can be more informative about what is being measured. It is possible to study these related measures as distinctly discriminant or convergent when a clear distinction is needed, or observed as a pattern along a continuum (Hubley & Zumbo, 2013). It is likely that convergent validity for the RSES may be found using measures of constructs such as depression, satisfaction with life, confidence, and social skills, among others. Evidence of discriminant measures is useful to show that the RSES is not measuring a construct that has similar traits, such as the previous
examples, and is also reinforced by comparing it with measures with which it should have little to no relationship, such as intelligence or gender identity. Discriminant evidence is also likely to be found for the RSES, although less so than convergent validity evidence. Studies that provide discriminant validity evidence are likely to pair it with convergent validity evidence.

For the purpose of observing convergent and discriminant validity along with reliability, as well as examining the method effect in these coefficients, Campbell and Fiske (1959) developed the Multitrait-Multimethod (MTMM) matrix. This matrix presents reliability and validity coefficients in an organized matrix to have a simple visual comparison of convergence, discriminance and reliability. These coefficients compare two or more methods of measuring a construct and two or more traits or constructs. Reliability is presented along the main diagonal of the matrix, convergent validity in the monotrait-heteromethod correlations and discriminant validity in the heterotrait-monomethod and heterotrait-heteromethod correlations – the latter being strongest evidence of discriminant validity because neither construct nor method is shared (Hubley & Zumbo, 2013). This method is unique in that it counts the method as a variable for validity and isolates it in the analysis. The MTMM matrix requires more than one method to measure a construct, which might not be easily accessible or even possible in some cases (Campbell & Fiske, 1959). This difficulty has worked against a wider adoption of its use. While it is possible for an MTMM method to be used for the validity of the RSES, the construct to be measured (self-esteem) is difficult to measure in a different format other than self-report, and thus it is unlikely that such methodology will be found in current validation studies.
Secondly, we have test-criterion relationships, which refers to how accurately a test score predicts or is related to criterion performance, as a source of evidence (AERA et al., 2014). A criterion is the standard that a measure is trying to approach, such as a result, an outcome, or a diagnosis. It is also the ultimate indicator of the construct of interest. Measures with strong test-criterion relationships are practical for when cost, time, or availability does not allow the use of the criterion itself, and the measure can act as a substitute (Hubley & Zumbo, 2013). For example, the Student Adaptation to College Questionnaire (SACQ) assesses adjustment to college to detect problems that cause dropout. In a test-criterion relationships study of this measure, the criterion could be student dropout and grades, and the SACQ would be used in situations where it would not be possible or practical to wait for these results. While this is similar to convergent validity, this comparison is not done with a scale that measures a similar construct, but rather with what is considered to be the criterion or the ultimate indicator that usually one cannot obtain due to time or financial constraints. Test-criterion relationships can be subdivided into predictive and concurrent validity. Predictive validity evidence is obtained when test scores are correlated to a criterion collected at a future point in time (Furr & Bacharach, 2014); concurrent validity is obtained when the test scores are correlated with a criterion that is obtained simultaneously.

Thirdly, we have validity generalization. This refers to how test scores vary in populations with different characteristics. It can be evaluated by observing the validity coefficients obtained from different studies (AERA et al., 2014; Furr & Bacharach, 2014). Validity generalization can show how generalizable a test can be.
to a specific group; it can give an idea of the predictive validity of the test, the variability of validity evidence among smaller groups, and can help in detecting sources of variability among studies (Furr & Bacharach, 2014).

The RSES is a popular measure and it has been compared with other similar measures so evidence for convergent and discriminant validity is expected (Blascovich & Tomaka, 1991; Robins, Hendin & Trzesniewski, 2001). It is unlikely to find a criterion related to the RSES given that self-esteem does not have a clinically diagnosed gold standard approach to its measurement and is not really a stand-in for a specific outcome. Based on prior research (Chinni & Hubley, 2014; Gunnell, Schellenberg et al., 2014), it is expected that instances may be found in which the RSES has been incorrectly dubbed as a criterion measure because of incorrect use of the term. As mentioned before, the RSES has been used in various groups with differences in age, language and ethnicity, and thus it is possible to find validity generalization evidence.

Consequences of testing. Finally we have validity based on consequences of testing, which is a more recent addition to the sources of validity evidence (AERA et al., 2014; Messick, 1995a; Cizek, Bowen & Church, 2010), and is related to the intended use and consequences for the testing. According to Messick (1995a, p. 746), it evaluates “the intended and unintended consequences of score interpretation and use in both the short- and long-term”. Validity based on consequences of testing considers if the test is being used for its intended purpose and the effects of that. The uses and interpretations of the test have to remain coherent to the original design, fitting the model without causing potentially negative
circumstances (AERA et al., 2014). It is important to note that validity evidence based on consequences should not be confused with test misuse or issues of social policy, ethics, and other similar areas (Messick, 1995a; Messick, 1995b; Cizek et al., 2010).

Validity based on consequences of testing can be observed in two aspects: value implications and social consequences (Messick, 1995a; Hubley & Zumbo, 2011). Value implications of a measure are the influences of values on the scale; social and personal values are reflected in the choice of construct, the theory that supports the development of the scale, as well as the naming and intended use of said scale (Hubley & Zumbo, 2011). These values are also an incentive to use a measure to make decisions and can show differing scores in a different social context (Messick, 1995a). Social and personal values also influence findings and formation of theories, what studies are considered important to study, or the focus of studies, and by extension the theory that is formed or confirmed by it (Hubley & Zumbo, 2011).

Social consequences are the effects caused by the use of a measure when utilized for its intended use (Messick, 1995a); this includes intended and unintended results, both positive and negative. Consideration should also be given to the relevance and usefulness of the scores of a test when utilized with a social group different from the original development sample, the roles and impact that the use of this scale has in this context, and consequences that might arise from the particular use of the test (Hubley & Zumbo, 2011).
Because the original scale was developed for adolescents in American schools, there are implied social characteristics for the intended use such as an age range, language of the scale, and literacy level – in this case, adolescents that have more than elementary-level education between 14 and 18 years of age in a school setting in which English is spoken. The RSES has been adapted and used with a variety of samples of different ages, ethnicities and nationalities, and contexts, and has been a popular measure (Blascovich & Tomaka, 1991); as such, consequences of testing could certainly be an area of examination. However, Cizek, Rosenberg, and Koons (2008) showed that evidence for the consequences of testing is rarely present in validation studies, and even other works of research synthesis focused on validity rarely mention it (e.g., Barry, Chaney, Piazza-Gardner & Chavarria, 2014; Hogan & Agnello, 2004; Pawlowski et al., 2013). Thus, it is possible that validity evidence based on consequences of testing for the RSES will not be present.

**Research and Validation Synthesis**

Research synthesis is a process in which the activities that evaluate the quality of research are analyzed and reviewed to produce generalizations. It is also referred to as systematic review and meta-analysis. Although there is no clear consensus of difference among these terms (Chalmers et al., 2002; Cooper et al., 2009), one accepted definition of a systematic review is a study that uses original research as subjects in order to summarize findings on the selected topic, and meta-analysis is the statistical method that combines the results of these individual studies (Cook, Mulrow & Haynes, 1997; Egger, Smith & Altman, 2008). Evans and Kowanko (2000) distinguish a meta-analysis as an earlier term and intended to be a study that
summarizes research data, while the term “systematic review” is a development from the early 1990s that emphasizes the extensiveness and depth of the review topic.

Research synthesis was originally used in medical research to summarize findings of specific health care topics (Chalmers et al., 2002; Evans & Kowanko, 2000), such as effects of a vaccine against typhoid. In recent years, works of research synthesis have expanded to various areas such as neuropsychology and mental health (e.g., Meyer et al., 2014; Pawlowsky et al., 2013), education (e.g., Whittington, 1998), econometrics (e.g., Rhodes, 2012), psychology (e.g., Zumbo & Chan, 2014) and management (e.g., Rothstein, 1992). It has been used as a tool to evaluate the methodological quality of the design and analyses of research, reduce the impact of type II error by comparing the results against similar studies, and support studies with smaller sample size to evaluate meaningful significance (Altman, 2001; Egger et al., 2008).

One of the most valuable assets in a research synthesis is the inspection of different sources of evidence to better the components of research and its use of the cumulative reports of various studies to identify strengths, weaknesses, and discrepancies found in the results (Banks & McDaniel, 2012; Cooper, Hedges & Valentine, 2009; Evans & Kowanko, 2000).

A research synthesis includes relevant research reports that meet quality research standards in a systematic unbiased manner, and summarizes the findings (Chalmers et al., 2002; Rhodes, 2012). It serves as a valuable tool to reduce biases
and produces an integrative, processed summary of findings (Chalmers et al., 2002).

Cooper et al. (2009) report that research synthesis techniques in the social sciences began to be adapted around the 1970s. In 1971, Feldman proposed that a systematic review of literature could be a type of research on its own by sampling topics and studies, coding and summarizing the material, integrating the data and presenting a summary. In the same year, Light and Smith (1971) presented what they called a cluster approach that summarized data from various studies, and proposed that variations among outcomes could be a useful source of information. Cooper et al. (2009) noted that studies applying a methodology of research synthesis began appearing around 1977, although previous works as far back as 1904 supported the development of this methodology by using average combined estimates across studies. The purpose of using research synthesis to summarize the gathered research has remained constant through the years (Banks & McDaniel, 2012; Chalmers, Hedges, & Cooper, 2002); however, some of its characteristics have changed as the method developed.

In recent years, the efficacy of research synthesis as a research tool has been examined, and is reported to be a useful aid to researchers, although there is little consensus on the application and guidelines to follow when doing this kind of study (Banks & McDaniel, 2012; Chalmers et al., 2002; Evans & Kowanko, 2000; Leroy, 2012; Wolf, 1986). Some of the issues that might cause concern about research syntheses are the comparison of studies that might be too different, the interpretation of a synthesis that includes poorly designed studies, bias towards the
availability of published articles, selection bias on the sampled articles, the use of multiple results from the same study, and ensuring a fair coding of the papers (Leroy, 2012; Wolf, 1986).

Another development in research synthesizes is the use of journal databases; while earlier works had been done utilizing printed copies of journals such as Cook et al. (1997) or Hogan and Agnello (2004), more recent synthesizes have shifted to the use of online databases like PsycINFO, MEDLINE, or the Cochrane Database of Systematic Reviews (Chalmers, Hedges, & Cooper, 2002; Chinni & Hubley, 2014; Meyer-Moock, Feng, Maeurer, Dippel, & Kohlmann, 2014). Researchers doing synthesizes work may also choose to include or exclude previous works of meta-analysis (e.g., Whittington, 1998).

Research synthesizes have been used to compare the strengths and weaknesses of measures, as well as compare them against other measures in areas such as their ease of use, evidence of reliability and validity, scope, and target population (Meyer et al., 2014; Pawlowsky et al., 2013). Research synthesis focused on validity and reliability evidence tend to now be referred to more specifically as reliability synthesis and validity synthesis. Validity and reliability synthesizes can report prevailing or popular trends in current research, such as what type of analyses are being used most commonly or what is missing or is being done poorly in studies (e.g., Chinni & Hubley, 2014; Pawlowsky et al., 2013), or show the areas that lack proper evidence or need further study (Cook et al., 1997; Egger et al., 2008). In addition, they can provide the basis to suggest guidelines for future
planning and research for validation work on a specific measure (Chinni & Hubley, 2014; Meier & Davis, 1990; Pawlowsky et al., 2013; Whittington, 1998).

Reliability synthesis. Research syntheses that focus solely on reliability practices are known as reliability synthesis. For example, a research synthesis by Willson (1980) reported the percentage of studies that examined reliability, and the method used to obtain a reliability coefficient. The Standards (AERA et al., 2014) consider reliability as both the correlation of scores between two equivalent forms of a measure as well as the precision or consistency of scores across replications of a procedure. It recognizes three broad categories for reliability coefficients: derived from alternate forms, obtained by test-retest, and based on the interactions and relationships among scores from single items or subsets of items as an internal consistency coefficient.

The Standards (AERA et al., 2014) state that the interpretation of reliability of test scores depends on the population being tested and, as such, reliability should always be estimated for all relevant groups and subgroups to the extent feasible with large enough sample sizes for each subgroup. Additionally, when documenting reliability, potential sources of error should be reported as well as estimates that allow for a clear understanding of the data, such as standard error of measurement. Finally, when using raters, decision consistency in observed classifications of examinees should be maintained, as well as decision accuracy – which is the observed classifications agreeing with the true classifications.

Raykov and Marcoulides (2015) consider that reliability evidence for scores of a test can be compared across different studies to produce observations relevant to
the scale as a whole. Chan (2014) observes that reliability is part of the sources of measurement evidence in more than one standard or guideline, such as *The Standards* (AERA et al., 2014), the Consensus based Standards for the selection of health Measurement Instruments (COSMIN; Mokkink et al., 2010) and the Evaluating the Measurement of Patient-Reported Outcomes (EMPRO; Valderas et al., 2008).

Reliability syntheses are useful in documenting the practices of reporting reliability. Hogan, Benjamin and Brezinski (2000), for example, reported that the most commonly used estimate for reliability is Cronbach’s alpha, used in at least 66% of their sampled studies, followed by test-retest reliability estimates in 20% of studies. A study by Thompson and Snyder (1998) reported that 52% of the measures they examined only cited reliability, 4% obtained a reliability coefficient for their sample, and 32% did both, while 12% did not consider any reliability coefficient. Vacha-Haase, Ness, Nilsson, and Reetz (1999) found 35.6% of the sample studies reported reliability coefficients based on their own data, 22.9% of the studies only presented reliability coefficients from previous studies, 3.8% of studies cited studies in which a reliability coefficient was presented but did not report the coefficient, and 36.4% did not mention reliability coefficients. Furthermore, Slaney et al. (2009, 2010) reported that most studies do not calculate reliability coefficients that reflect a multidimensional factor structure in their model.

*The Standards* (AERA et al., 2014) present reliability as a characteristic of the test scores that has implications for validity as well as a component of validity evidence based on internal structure. Some studies have presented research
syntheses that study reliability practices, validity practices, and their interaction, such as the need to consider both validity and reliability of scores in order to interpret them adequately (Barry et al., 2014; Qualls & Moss, 1996), the relationship between reliability and internal structure validity evidence (Slaney, Tkatchouk, Gabriel & Maraun, 2009), and common errors in reporting reliability and validity (Whittington, 1998).

**Validation synthesis.** Just as validation is the process by which validity is collected and evaluated in order to make a judgement about it (Zumbo & Chan, 2014), validation synthesis collects information about the sources of validation to produce a synopsis, assess their characteristics, and produce judgement about them (Evans & Kowanko, 2000). Validation synthesis is a critical evaluation of the available sources of validity evidence.

Validation synthesis, like reliability synthesis, is considered a relatively recent methodology. Its goal is generally accepted to be about presenting generalizable statements about the validation sources used with a measure or measures, and presenting areas of opportunity for further validity research. The recently published book “Validity and Validation in Social, Behavioral, and Health Sciences” (Zumbo & Chan, 2014) is one, if not the only, publication currently that focuses on validation (and to some extent reliability) synthesis and reflects on the validation practices based on these analyses.

Earlier works report the use of research techniques, reliability and validity, and comment on the lack of standard practices on reporting reliability and validity evidence (Meier & Davis, 1990; Willson, 1980). More recently, validation syntheses
have used *The Standards* (AERA et al., 2014) as a reference and compared validity practices to these standards (Jonson & Plake, 1998; Qualls & Moss, 1996; Slaney, Tkatchouk, Gabriel, Ferguson, Knudsen & Legere, 2010; Vacha-Haase et al., 1999).

A review of validation sources has shown that validity evidence is commonly reported as content, construct, and criterion validity (Cizek et al., 2010). Using the classification based on *The Standards* (AERA et al., 2014), the most prevalent sources of validity evidence are internal structure and relations with other variables, while there are few studies that present validity evidence based on test content and response processes (Chan, Munro et al., 2014; Chan, Zumbo, Chen et al., 2014; Chinni & Hubley, 2014; Cizek et al., 2008; Cox & Owen, 2014; Gunnell, Wilson et al., 2014; Hogan & Agnello, 2004). Validity based on consequences of testing, however, has still not seen much prevalence in validation works; while there had been some studies that present this source of evidence (Ark, Ark & Zumbo, 2014; Chan, Zumbo, Darmawanti & Mulyana, 2014; Cizek et al., 2008; McBride, Wiens, McDonald, Cox & Chan, 2014; Sandilands & Zumbo, 2014), its utility and relevance to validity evidence has been debated (Chinni & Hubley, 2014; Cizek et al., 2010; Messick, 1995b; Zumbo & Chan, 2014). This might be explained as it has only recently begun to be adapted and has been slow to be accepted by the scientific community (Cizek et al., 2008; Hubley & Zumbo, 2011), although Shear and Zumbo (2014) argue that this could be due to the difficulty in establishing a study that examines validity evidence based on consequences due to the complexity of such a study.
Meier (1990) reported that authors have shown a trend of reporting reliability and validity related to a scale as “good”, but only “provide a citation that presumably contains quantitative and qualitative information about the scales” (p. 113). Hogan and Agnello (2004) pointed out that only 55% of studies report validity, whereas about 94% of studies report reliability (Hogan et al., 2000); other authors report similar findings (e.g., Barry et al., 2014; Slaney et al., 2009, 2010; Whittington, 1998). Some studies also report reliability and validity improperly or missing specific descriptions such as their method of assessment or the development process of adaptations of a measure. For example, Barry et al. (2014) reported that some adapted scales treat reliability and validity evidence from the original scale as evidence for the adaptation itself. Whittington (1998) reported that a majority of published articles failed to consider the characteristics of their sample when reporting reliability (64%) and validity (82%). Research has shown that reliability evidence was reported more often than validity evidence in studies, that several studies cited reliability and validity evidence from previous studies but did not compute these coefficients for their sample, and that a significant number of them did not report or reported incorrectly reliability, validity, or both (Barry et al., 2014; Hogan, et al., 2000; Hogan & Agnello, 2004; Qualls & Moss, 1996; Slaney et al., 2009; Thompson & Snyder, 1998; Vacha-Haase et al., 1999; Whittington, 1998). Furthermore, Green et al. (2011) reported that research rarely takes into consideration the sample’s characteristics and its implications for reliability and validity. However, the reporting practices of validity and reliability have been improving, and more authors are adhering to what The Standards (AERA et al.,
2014) suggest for correct methodology (Cizek et al. 2008; Meier & Davis, 1990; Green, Chen, Helms & Henze, 2011). Even so, guidelines to validation practices still need to be completely adapted, as there is discrepancy between contemporary validation theories and how validation practices should reflect these theories (Chan, Munro et al., 2014; Chan, Zumbo, Zhang et al., 2014; Sandilands & Zumbo, 2014).

Another common observation that has been raised from validation syntheses is the misuse or misconception of the available research methodology (Cizek et al., 2010; Hogan et al., 2000; Slaney et al., 2010; Slaney et al., 2009; Willson, 1980), resulting in the use of inappropriate terms and conclusions that are unsupported by the presented data, such as using the term “criterion” to refer to a scale when using it as a convergent measure, or assessing reliability before internal structure validity.

Slaney et al. (2010) noted that the reporting trends of measurement-oriented journals were different from non-measurement-oriented ones, as the former would more likely identify sources of internal and external validity and include references to the psychometric characteristics of their tests, while the latter would identify more explicitly the theoretical structure.

Some previous research syntheses have used single journals and multiple journal comparisons for their sampling, usually examining multiple measures. These studies tend to use broader samples either by using a single database that has extensive entries or sampling more than one journal, as their aim tends to be an understanding of validation and reliability practices in general. The focus has been mainly quantifying the sources of validity and reliability presented, either cited or obtained for the sample. Cronbach and Meehl (1955), Hubley and Zumbo (2013),
and Slaney et al. (2009, 2010) noted the importance of the order in obtaining validity and reliability evidence, but most studies do not consider the particular details of the studies such as, if the measure had multiple factors, was reliability evidence based on each individual factor (subscales) or as a whole (total score), or if internal structure validity evidence was obtained before reliability and validity evidence based on relations with other variables.

Chan (2014) proposed that validation practices have veered towards collecting sources of validity evidence without emphasizing their synthesis to support the validity of inferences from test scores or their contribution to the overall validity. He also mentioned that the methodological approaches used to obtain validity should be examined, as not all standards or guidelines to obtain validity evidence might be suitable for every situation. Chan (2014) encouraged researchers conducting validation work to develop a “validation plan” (p. 21) situated within a view of validity appropriate to the inferences that are intended to be made. Zumbo et al. (2014) consider that a validation plan helps to give the ability to judge the strength of validity evidence, clarifying the purpose of the validation and pointing out what was not covered by the validation study.

As we refine our knowledge of validation and reliability practices as a whole, some research has begun focusing on the particulars of individual scales, such as Chinni and Hubley (2014) in their validation synthesis of the Satisfaction with Life Scale (SWLS). Scales such as the SWLS, and the measure proposed for this validation synthesis, have been in use for many years and several studies have examined the reliability of its scores and the validity of inferences made from those
scores. While validation syntheses that examine a variety of measures from single or multiple journals explain validation practices as a whole, a validation synthesis of a single scale can show how these practices have applied to a specific measure and over the life of that measure. A research synthesis—is a source that summarizes in a single document the sources of evidence sought to support the interpretation of RSES scores as well as the sources of evidence that still need to be explored. This type of validation synthesis is useful beyond the RSES because it shows the sources of evidence sought over the life of a measure rather than as just a snapshot of practices for many measures from a limited source (e.g., one journal or set of journals) or at a particular point in time.

**Study Purpose**

The overall purpose of this study is to document the sources of evidence used to support the interpretation of scores for a single measure, the RSES, over the life of that measure. Specifically, there are three main goals for this study. The first aim was to provide a description of the sources used to provide reliability and validity evidence related to the RSES, categorized by the three reliability sources and five validity sources described by *The Standards* (AERA et al., 2014). The second aim of this study was to determine if researchers follow a logical sequence when presenting reliability, validity evidence based on internal structure, and other validity sources. The third aim of this study was to highlight the sources of validity evidence that have not been examined for the RSES in the literature and provide some guidance for future validation research with the RSES and for reliability and validation synthesis in general.
Chapter 3 – Article

This chapter describes the methodology followed by the study, the results found and the discussion of these results.

Abstract

The Rosenberg Self-Esteem scale (RSES) is a commonly used measure, cited over 3,000 times in the past five years. The aim of this study was to produce a synthesis of the available sources of reliability and validity evidence for the RSES as classified by the Standards for Educational and Psychological Testing\(^1\) (AERA, APA & NCME, 2014). Despite the popularity of the RSES, only 27 articles have examined reliability and validity evidence for the scale. This study showed that the most prevalent source of reliability is based on internal consistency and the most prevalent validity evidence is based on internal structure, followed by relations to other variables. The latter source of evidence primarily consisted of convergent validity evidence. Evidence based on response processes is seldom examined and no studies examined validity evidence based on content or consequences of testing. When examining reliability, internal structure, and relations to other variables, studies tended to overlook the implications of the order in which these factors are studied. There is also a need for researchers to clearly state assumptions and criteria to interpret findings as well as more clarity in the reporting of results. The implications of these findings for researchers interested in the use of the RSES and for measurement experts will be discussed.

---

\(^1\) Henceforth referred to as “The Standards”
Keywords: psychometrics, reliability, research synthesis, Rosenberg Self-Esteem Scale, self-esteem, validity, validation

Background

This study examines the reporting practices of sources of reliability and validity evidence for the Rosenberg Self-Esteem Scale.

Self-esteem and the Rosenberg Self-Esteem Scale. Self-esteem is a popular construct in psychology with over 35,000 publications on the subject (Zeigler-Hill, 2013) and appears as a keyword in PsycINFO in 40,300 articles, and as a major heading in almost 17,000 articles; it has been viewed as an interpersonal signal of status in a social environment and as a protective resource against rejection and failure (Zeigler-Hill, 2013). The Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) is a well-known measure of self-esteem. A search of the PsycINFO database showed that the RSES has been cited 3,016 times from January 2010 to January 2015, showing that it is still a measure of great interest and common use.

The RSES is a 10-item measure typically scored using a 4-point Likert-type response format (Rosenberg, 1989). It was first published in 1962 as a modified Guttman scale; it was revised in 1989 using a Likert-type scale (Rosenberg, 1989). Given the widespread use of this scale, it is important to have strong evidence of reliability and validity, as well as a clear understanding of the sources that support the reliability of scores and the validity of inferences drawn from RSES scores.
**Sources of reliability and validity evidence.** Reliability is the precision or consistency of scores across replications of a procedure or the correlation of scores between two equivalent forms of a measure (AERA et al., 2014). *The Standards* (AERA et al., 2014) place reliability coefficients in three categories based on their source: derived from alternate forms, obtained by test-retest, and based on the interactions and relationships among scores from single items or subsets of items as an internal consistency coefficient.

Validity is the degree to which evidence and theory support the interpretation of the scores from a measure (AERA et al., 2014) whereas validation is the process of accumulating such evidence from relevant sources. *The Standards* (AERA et al., 2014) classifies validity evidence into five categories based on its source: test content, response processes, internal structure, relations to other variables, and consequences of testing. Validity evidence based on test content is focused on “the themes, wording, and format of the items, tasks, or questions on a test” (AERA et al., 2014; p. 11) as well as the administration and scoring methods and instructions. Content validity is threatened by construct underrepresentation and construct irrelevance variance (Furr & Bacharach, 2014) and can be examined using subject matter or experiential experts to rate the appropriateness and representativeness of these elements (Grant & Davis, 1997; Haynes, Richard & Kubany, 1995).

Validity evidence based on response processes typically relies on an examination of the behaviour or thought process of the person responding to the scale, or the examiner scoring the measure, to determine if these processes match what would be theoretically expected (AERA et al., 2014). Cognitive interviewing,
think aloud protocols, and focus groups are often used to gather such evidence (Padilla & Benítez, 2014).

Validity evidence based on internal structure shows the degree to which the test items and test components conform to the structure of the proposed construct (AERA et al., 2014). This evidence can be found by examining the scale’s dimensionality and measurement invariance (Rios & Wells, 2014). A measure’s dimensionality is the model of the factor structure for scale and usually examined with factor analysis, while item invariance means that scores are not affected by membership in a group (such as gender or race).

Validity evidence based on relationships with other variables examines how test scores relate to other scales, constructs, or known groups. These relationships can be categorized as test-criterion, generalization, convergent and discriminant (AERA et al., 2014). Test-criterion relationships examine how closely test scores relate or predict a criterion such as a medical diagnosis or an observable behaviour. Validity generalization is the evidence that suggests that inferences from test scores drawn from one group or test taker are supported in a different group. Convergent and discriminant validity evidence is examined in how test scores correlate to scores on other similar or dissimilar measures based on the theory of the construct. In simple terms, scales that should correlate relatively highly are considered convergent, and scales that should correlate relatively lowly or not at all are considered discriminant.

Finally, validity evidence based on consequences of testing evaluates the intended and unintended consequences of interpreting test scores when the test is
used appropriately for its intended purpose. This can be observed in the values implicit in the test development, and the intended and unintended social and personal consequences and side effects that occur by using the test (Hubley & Zumbo, 2011). It is important to note that validity evidence based on consequences does not include the misuse of a test or issues of social politics or ethics (Cizek, Bowen & Church, 2010; Hubley & Zumbo, 2011; Messick, 1995b).

**Reliability and validation synthesis.** A research synthesis, or systematic review, is a study that uses previous studies as the research subjects in order to summarize the findings of research on a specific topic to produce generalizations and find gaps in the literature (Cook et al., 1997; Egger, Smith, & Altman, 2008; Evans & Kowanko, 2000). Reliability and validation syntheses are specific kinds of research syntheses that review studies providing evidence of reliability or validity and report on the practices used to collect that evidence (Zumbo & Chan, 2014).

Previous research syntheses examining reliability and validity (e.g. Barry, Chaney, Piazza-Gardner & Chavarria, 2013; Chinni & Hubley, 2014; Cizek, Rosenberg & Koons, 2008; Hogan & Agnello, 2004; Zumbo & Chan, 2014) have classified sources of validity evidence according to *The Standards* (AERA et al., 2014); one of the outcomes that these previous research syntheses have found is that most studies report internal consistency reliability evidence, validity evidence based on internal structure and validity evidence based on relations with other variables, while only a small fraction of studies report validity evidence based on content, response process, and test consequences. Slaney, Tkatchouk, Gabriel, and Maraun (2009) argued that reliability and validity should be examined in a logical
sequence, considering that a scale should examine the internal structure validity of its scores before analyzing reliability, and reliability should be assessed before relations with other variables, although only 32% of the studies they examined followed this logical sequence.

Vacha-Haase, Ness, Nilsson, and Reetz (1999) and Hogan, Benjamin, and Brezinski (2000) reported that reliability is present in most studies. Vacha-Haase et al. (1999) showed that only about a third of studies calculated reliability coefficients based on their sample, and about 25% cited reliability from previous studies. Hogan et al. (2000) reported that about 75% of studies reported only one reliability coefficient, most commonly an internal consistency alpha or, to a lesser degree, a test-retest reliability estimate. Barry, Chaney, Piazza-Gardner, and Chavarria (2013) found similar results and reported that reliability is more commonly reported than validity evidence.

Qualls and Moss (1996) reported that a large number of authors do not follow the guidelines set by The Standards (AERA et al, 2014) when assessing reliability and validity. Jonson and Plake (1998) compared the reporting practices of studies in five time periods from 1937 to 1995, finding that authors were slowly including The Standards in their reporting practices and, while some used outdated standards, more updated versions of The Standards were increasingly present in recent works.

Whittington (1998) reported a lack of information on developing and piloting of tests, and suggested that, in order to draw adequate interpretations from test scores, the characteristics of the sample from which scores were drawn should be considered. Zumbo and Chan (Chan, 2014, Zumbo & Chan, 2014) expanded on this
idea, proposing that studies of reliability and validity should also consider a plan for how their practices affect the overall use of the scale and contribute to validity and reliability practices.

A majority of studies of validity and reliability synthesis examine a sample of articles from one or more journals (e.g. Hogan et al., 2000; Hogan & Agnello 2004, Hubley, Zhu, Sasaki & Gadermann, 2014; Slaney 2009, 2010; Vacha-Haase et al., 1999); this method is intended to describe reliability and validation practices as a whole or within a time frame or discipline. A reliability and validation synthesis can be done for a single scale, such as the study by Chinni and Hubley (2014), which focuses on the Satisfaction with Life Scale (SWLS). They found that reporting practices for reliability and validity for the SWLS were similar to prior findings for general reporting practices. For studies that examined the relations with other variables evidence, they noted there was a lack of description or hypotheses for how scores were expected to relate. Furthermore, a majority of studies treated translated versions that were treated as equal to the original.

The benefit of analyzing reliability and validation practices for a single scale is that you obtain a summary of available sources of reliability and validity, an asset for those interested in using the scale. It also allows one to find gaps in the literature as well as provide recommendations for further research. For a scale as widely used as the RSES and with a long history in the literature, such an analysis would help demonstrate in the ways in which validity and reliability sources have been examined, assisting those interested in the use of the RSES, those studying validity
and reliability in general, and those interested in research syntheses with a focus on validity and reliability.

Study Purpose

There are three main goals for this study. The first goal is to provide a description of the sources used to provide reliability and validity evidence related to the RSES as categorized by The Standards (AERA et al., 2014). The second goal is to determine if a logical sequence is followed when presenting reliability and validity evidence. The third goal is to highlight the sources of validity evidence that have not been examined in the literature and provide an area of opportunity for future validation research.

Method

The following are the factors considered in order to examine the reporting practices of reliability and validity evidence.

Database search. PsycINFO was selected as the optimal database to research reliability and validity studies related to the RSES because of the database’s focus in the area of psychology. PsycINFO covers over 3.6 million records on psychology dating from 1597 to present, is updated weekly, and contains publications from more than 50 countries in 29 languages (American Psychological Association, 2014). The search for publications in this study was limited to the span of January, 19891 to January 2015.

1 1989 was selected because in that year Rosenberg formally revised the scale to use a Likert-type response format.
A search for literature was made using the terms “analysis”, “consequence”, “content”, "convergent", "criterion", "discriminant", “factor analysis”, "generalization", “item response theory”, “internal structure”, "psychometric", “reliab**”, “response process”, “Rosenberg”, “self-esteem”, “talk aloud”, “think aloud”, and “valid**” The terms “Rosenberg” and “self-esteem” were searched separately, and combined with each other keyword (i.e., Rosenberg OR self-esteem AND *keyword*). Only articles published in English were included, although the sample includes articles with translated or adapted versions of the RSES. Articles that used the Guttman scale were excluded from this study as this is a different version of the RSES that has long since been revised. Modified versions of the RSES (e.g., revised item wording or response format) were also excluded, as evidence of validity found using the standard version may not apply to any modified versions and vice-versa. References for validity evidence for the RSES were also searched for in the citations presented in the relevant articles. The original search yielded 1,475 articles (See Fig. 1). By reading the title and abstract, relevant articles were selected, resulting in 32 articles. After a thorough read of these articles, inclusion and exclusion criteria were applied and five articles were removed\(^1\) from the sample. This resulted in 27 research articles (30 studies) presenting reliability or validity evidence that met the inclusion and exclusion criteria. These articles are listed in Appendix A.

\(^1\) One article was excluded because used the 1965 version of the RSES on a Guttman scale, one study was a meta-analysis, one study examined a different scale and used the RSES only as a reference, and two studies were modified versions of the RSES.
Reliability and validation synthesis coding. Coding was done for each of the 30 studies, with the following characteristics recorded: publication year, sample size, the country where the data were collected, if the research was done with the original revision or a translated or adapted version, the language in which the RSES was presented, if the translated instrument was piloted or pre-tested, and if there was a citation provided for translation method or guidelines used.

Each study was coded according to the sources of reliability and validity evidence provided in The Standards (AERA et al., 2014). This coding was based on the description by The Standards, regardless of how the authors described this method. For reliability, it was recorded what sources or estimates were used to provide evidence (i.e. internal consistency, item-total correlations, test-retest, alternate forms). For internal consistency, it was coded what type of estimate was used, if the criterion for said estimate was noted, the description of values obtained,
as well as if inter-item correlations were reported, and if reliability estimates were provided assuming the RSES is unidimensional or if the author(s) used or ignored internal structure results reported within the same study. For test-retest reliability, it was coded if the sample used for this purpose was the same at test and retest, was based on a subsample, the interval between testing sessions, and the rationale for this interval length. As the RSES is not known to have any parallel or alternate forms, this method of reliability estimation was not applicable.

For coding of validity evidence, this was classified by test content, response processes, internal structure, relations with other variables, and consequences of testing. Studies that presented validity evidence that matched more than one source were counted as a case for each of them. It was also noted if studies cited references to modern validity theory (e.g. The Standards, Messick).

**Evidence based on content.** Studies were coded for the following: use of subject matter, experiential, or practical experts; methodology used; and use of the terms “construct”, “representativeness”, “relevance” or “appropriateness” in the context of validation.

**Evidence based on response processes.** Coding captured if the following characteristics were present: use of measures of behaviour (e.g., eye movement or response time), measures of response strategies (e.g., using think aloud protocol, cognitive interviews), and theoretical or empirical analyses of the fit between the response processes of the sample with self-esteem construct and theory.
Evidence based on internal structure. Studies were coded for the use of factor analysis (FA) and, if so, what type (exploratory, confirmatory, or both). If the study conducted an exploratory factor analysis (EFA), further coding noted if the criteria to determine the number of factors was stated a priori, the criteria used to determine the number of factors (i.e., eigenvalues, scree plot, parallel analysis, percentage of variance explained), if factor loadings were provided, the criterion for loadings that was used, if more than one factor was identified and, if relevant, the rotation method reported. If the study conducted a confirmatory factor analysis (CFA), coding noted if the study stated the number of factors expected, which items loaded on which factors if more than one factor was expected, the fit indices used as well as their criteria, and the rationale for selecting those fit indices. If the study conducted both EFA and CFA, it was coded if the study used the same sample or different samples for each analysis. Additional coding noted if the studies examined measurement invariance and/or differential item functioning (DIF), and if the results of DIF were then related to internal structure or dimensionality.

Evidence based on relations with other variables. Studies were coded for the number of scales (or subscales) used for comparison, how the researchers referred to this process of validation (e.g., relations with other constructs/variables, convergent, discriminant, divergent, concurrent, construct validity), if the study aimed to provide convergent or discriminant validity evidence or both, if the researchers stated clearly what they expected for this source of validity evidence, if there was an explanation for the rationale for using specific constructs or measures, how the researchers showed if the evidence did or did not support validity, and if
they used other methods such a multitrait-multimethod (MTMM) matrix. While test-criterion relationships might not be relevant to the RSES, it was coded if the authors intended to use this evidence or misused the term “criterion” to refer to convergent validity evidence. Also coded was the intention to use known-groups evidence or between-groups differences as part of validation evidence, if this was described as something else, if the assumed differences were referenced in the literature, or if groups were only compared without providing validity evidence.

Because this source of evidence has shown to be poorly presented (Chan, Zumbo, Darmawanti & Mulyana, 2014; Chinni & Hubley, 2014; Gunnell, Schellenberg et al., 2014; Hogan & Agnello, 2004), coding for this section was inclusive in order to capture the most possible studies relevant to validity evidence based on relations to other variables.

**Evidence based on consequences.** Coding here focused on whether the studies used the terms “consequences”, “consequential”, “impact”, “implications”, “effect”, or “values” related to validity and if the study author(s) showed signs of misunderstanding evidence based on consequences of testing (e.g., by referring to test misuse).

**Coding check for open-ended items.** The sections that code test content, response processes, relations to other variables, and test consequences included open-ended items that search for any description of methodology and terminology corresponding to validation processes. Unlike the sections that code for translations, reliability, and validity evidence based on internal structure, these items describe the process that researchers used and the descriptions they provided. The reason for
this open-ended style was because of the difficulty in finding a clear framework for presenting these sources of evidence, as well as evident confusion with terminology (Chinni & Hubley, 2014; Cizek et al., 2008; Hogan & Agnello, 2004; Hubley et al., 2014).

The section coding for validity evidence based on other variables included open-ended items that could be misinterpreted, which could cause an incorrect description of the characteristics of the reporting practices of validity evidence. In order to reduce the impact of subjectivity, another examiner was asked to review this section. The second examiner was a doctoral student in measurement. She was presented with the articles that contained validity evidence based on relations to other variables and my completed coding sheets, and was asked to review them. She identified themes, information that might have been missed or was confusing and was asked if she considered the coding scheme to be clear, consistent and impartial.

**Logical sequence of reliability and validity evidence.** The number of sources of reliability and validity evidence were recorded within studies as well as the order in which this evidence was presented. It was also examined if this order followed a logic sequence (e.g., was internal structure measured before reliability?, was reliability measured before relations with other variables?, was internal structure measured before known groups evidence?), as well as if there was a methodology guide or validity “plan” as described by Chan (2014).
Results

The literature search resulted in 27 articles that fit the inclusion criteria. For one of these articles, the author conducted multiple studies using different samples and methods; each of these studies was treated independently, bringing the total of studies examined for this study to N=30. As seen in Table 1, 15 (50%) of the studies used a translated version of the RSES. Reliability evidence as a whole was reported in 25 (83.3%) of the studies; of the studies that reported reliability, internal consistency reliability was reported in all of them and test-retest reliability was reported in 5 (20%). Validity evidence, as classified in The Standards (AERA et al, 2014), was present as follows: validity evidence based on internal structure was present in 27 (90%) of studies, validity evidence based on relations with other variables was found in 13 (43.3%) of studies, and validity evidence based on response processes was found in 1 (3.3%) study. No studies examined validity evidence based on content or consequences of testing. Of the 30 studies, 12 (40%) identified their study clearly as a validation study, 9 (30%) noted how their study contributed to validity, and 9 (30%) did not identify their study as contributing to validity.
<table>
<thead>
<tr>
<th>Author</th>
<th>Translated language</th>
<th>Reliability</th>
<th>Alpha</th>
<th>Split-half</th>
<th>Composite</th>
<th>Rasch Person statistic</th>
<th>Sum of variance(^1)</th>
<th>Test-retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alessandri, 2005a</td>
<td>Italian</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alessandri, 2005b</td>
<td>Italian</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alessandri, 2005c</td>
<td>Italian</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alessandri, 2005d</td>
<td>Italian</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aluja, 2007</td>
<td>French</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baños, 2000</td>
<td>Spanish</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boduszek, 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen, 2007</td>
<td></td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demo, 1985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Franck, 2008</td>
<td>Dutch</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gana, 2005</td>
<td>French</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goldsmith, 1986</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gray-Little, 1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halama, 2008</td>
<td>Slovak</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hatcher, 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mannarini, 2010</td>
<td>Italian</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marsh, 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martin, 2006</td>
<td>Chinese</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martin-Albo, 2007</td>
<td>Spanish</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McMullen, 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rizwan, 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shapurian, 1987</td>
<td>Persian</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shevlin, 1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sinclair, 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supple, 2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supple, 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tomás, 1999</td>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td>Translated language</td>
<td>Reliability</td>
<td>Alpha</td>
<td>Split-half</td>
<td>Composite</td>
<td>Rasch Person statistic</td>
<td>Sum of variance&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Test-retest</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>-------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------------------</td>
<td>-----------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>V.-Raposo, 2012</td>
<td>Portuguese</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vermillion, 2007</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W.-Mansell, 2003</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>15</td>
<td>25</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

<sup>1</sup> As described by Alessandri, Vecchione, Eisenberg & Laguna, 2005

Table 2. Sources of validity examined

<table>
<thead>
<tr>
<th>Author</th>
<th>Validity</th>
<th>Response processes</th>
<th>Internal structure (EFA)</th>
<th>Internal structure (CFA)</th>
<th>Internal structure (other)</th>
<th>Relations with other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alessandri, 2005a</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Alessandri, 2005b</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Alessandri, 2005c</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Alessandri, 2005d</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Aluja, 2007</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Baños, 2000</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Boduszek, 2013</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Classen, 2007</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Demo, 1985</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Franck, 2008</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Gana, 2005</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Goldsmith, 1986</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Gray-Little, 1997</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Halama, 2008</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Hatcher, 2009</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Author</td>
<td>Validity</td>
<td>Response processes</td>
<td>Internal structure (EFA)</td>
<td>Internal structure (CFA)</td>
<td>Internal structure (other)</td>
<td>Relations with other variables</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------</td>
<td>--------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Mannarini, 2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Marsh, 2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Martín, 2006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Martín-Albo, 2007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>McMullen, 2013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rizwan, 2012</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Shapurian, 1987</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Shevlin, 1995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sinclair, 2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Supple, 2011</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Supple, 2013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tomás, 1999</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V.-Raposo, 2012</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vermillion, 2007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W.-Mansell, 2003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>30</td>
<td>1</td>
<td>8</td>
<td>17</td>
<td>6</td>
<td>14</td>
</tr>
</tbody>
</table>
Translation and adaptation. As mentioned earlier, 15 (50%) of studies used a translated version of the RSES for their studies. The RSES was translated into Chinese, Dutch, French, Italian, Persian, Portuguese, Slovak, and Spanish. Of those 15 studies, three (20%) of the translations were created for their use in the study, 10 (66.6%) used a previously translated version, and two (13.3%) were not clear on the source of their translated version. Of the studies translated, 3/15 (20%) specified the use of experts for their translation. One study described their experts as bilingual and native speakers of the target language who translated the RSES into Dutch and one bilingual native English expert who back-translated to correct inconsistencies. The second study described this process as parallel back translation by two bilingual individuals and two psychology professors. The third study described the experts as 12 bilingual Iranian experts in psychology and measurement. In addition, of the translations of the RSES done for the specific study, 1 (6.6%) study specified that the scale translation was adapted to the population by using psychology and measurement experts to judge if the content of each item served as a meaningful measure of self-esteem in the target culture. Of the remaining 12 studies, none provided any details on the process of translation or the characteristics of that version. No study reported use of pilot testing.

Reliability. Reliability estimates were reported for 25/30 (83.3%) of the studies. Out of those, internal consistency reliability was reported in all of the studies and test-retest reliability was reported in 5 (20%) studies. As there are no known alternative forms for the RSES, no studies reported alternate or parallel form reliability estimates.
Of the 25 studies that reported an internal consistency estimate, an alpha reliability estimate was reported in 18/25 (72%) studies, split-half reliability was reported once (4%), composite reliability once (4%), and Person reliability index once (4%); furthermore, in 4 (16%) of the studies, which were part of the same multi-study article, presented a reliability estimate calculated by the sum of the variances of the latent factors of the scale divided by the total scale variance. Of the 25 studies that reported an internal consistency reliability estimate, 11 (44%) stated what they considered an acceptable reliability coefficient, which varied on a range from .60 to .80 ($M= 0.74$, $SD= 0.06$).

Inter-item correlations were reported in 7 (out of 25, 28%) of the studies, 3 (42.9%) of which also reported average inter-item correlations. All 3 average inter-item correlations were calculated according the proposed or expected factor structure. Three out of the 25 internal consistency studies (12%) studies reported the standard error of measurement.

Out of the 5 (20%) studies reporting test-retest reliability coefficients, 4 (80%) reported their retest interval (i.e., 15 days, 3 weeks, 4 weeks and 6 months) and one (20%) was unclear about it. This latter study examined the test across four school year intervals, but was unclear on the exact interval period. One study used school years as the retest interval, the other four studies did not provide a rationale for the chosen time interval. For the retest sample, 3 (60%) of the studies used their full sample, and 2 (40%) used a sub-sample.
**Response processes.** Of the 30 studies, 1 (3.3%) examined validity evidence based on response process (Table 2). This study examined the RSES scores using IRT, describing the discriminating power of the items, and discussing the relevance and effect of wording on the responses. It also showed the response patterns and discussed the possibility of modifying or removing items.

**Internal structure.** Of the total 30 studies, 27 (90%) examined validity evidence based on internal structure (Table 2). Out of those 27 studies, 7 (25.9%) conducted an exploratory factor analysis (EFA), 16 (53.3%) conducted a confirmatory factor analysis (CFA), 1 (3.7%) conducted both analyses, and 6 (20%) studies examined validity evidence based on internal structure with another method.

**EFA.** Of the 8 studies that conducted an EFA (i.e., 7 studies using EFA and 1 study using both EFA and CFA), 7 (87.5%) used a Principal Components Analysis (PCA), and 1 (12.5%) used Pearson Product Moment Coefficients of Correlation. The criteria used to determine the number of factors in EFA was clearly stated in 5 (62.5%) of cases, and unclear in 2 (25%) of cases. In 2 (out of 5, 40%) of cases this criteria was presented a priori. Of the five studies that reported criteria to determine the number of factors, 3 (60%) used a single criterion, 1 (20%) used 2 criteria and 1 (20%) used 3 criteria. Eigenvalues greater than 1 was used as a criterion in 4 of the studies (50%), scree plots were used in 2 (25%), parallel analysis was used in 1 (12.5%), percentage of variance explained was used in 3 (37.5%), and an item response analysis was used in 2 (25%) studies. Three (37.5%) studies mentioned percentage of variance explained but it was not clear if this was used as a criterion.
Out of 8 studies, 3 (37.5%) reported factor loadings for their EFA, and 1 (12.5%) study only reported the range of factor loadings. Of the 8 studies, only 1 (12.5%) reported more than 1 factor found; this study used an oblique method of rotation, but did not specify which type. No study reports a criterion to interpret the factor loadings.

CFA. A total of 17 (out of 30, 56.6%) studies conducted a CFA (i.e., 16 studies used CFA and one study used both EFA and CFA). Of those 17, 16 (94.1%) specified the number of factors expected, 14 (82.3%) presented at least one multidimensional model, and 14 (82.3%) examined more than one model in the same study by comparing one and two-factor models. Of these 14 studies, 9 (64.2%) reported their models describing which items loaded on each factor. All of the studies that performed a CFA specified the fit indices they used, 12 (out of 17, 70.5%) specified the cut off value for the fit indices, and 4 (23.5%) specified the rationale used to select those fit indices. Of the 17 studies that conducted a CFA, 1 (5.88%) study used only one fit index, 1 (5.88%) study used 3 fit indices, 3 (17.64%) studies used 4 fit indices, 7 (41.17%) used 5 fit indices, 4 (23.52%) used 6 fit indices and 1 (5.88%) used 7 fit indices.

Chi Square ($\chi^2$) was used as a fit index in all studies that performed a CFA, 13/17 (76.4%) used Root Mean Square Error of Approximation (RMSEA), 12 (70.5%) used Comparative Fit Index (CFI), 10 (58.8%) used Tucker-Lewis fit index (TLI), 8 (47%) used Goodness of fit index (GFI), 7 (41.1%) used Standardized Root Mean Square Residual (SRMR), 5 (29.4%) used Akaike information criterion (AIC), 4 (23.5%) used Normed Fit Index (NFI), 2 (11.7%) used Adjusted goodness of fit
index (AGFI), 1 (5.8%) used INFIT/OUTFIT, and 1 (5.8%) used Incremental fit index (IFI).

**CFA and EFA.** The study that examined both EFA and CFA used the same sample to perform these studies. Of the 25 studies that performed either type of factor analysis, 8 (32%) found more than one factor; of those 8, 3 (37.5%) continued to use a single score in their analyses, and 1 (12.5%) was unclear. Of the 24 studies that performed a factor analysis, 14 (58.3%) did so using the original English version.

**Other validation evidence for internal structure.** Five (out of 30, 16.6%) studies examined measurement invariance, examining invariance among English, Italian, Polish, and Serbian groups; male and female groups; adolescents and adults; school grades; and social phobic vs non-social phobic groups. Differential Item Functioning was examined in 1/30 (3.3%) studies using Item Response Theory as their method, examining the reliability, factor structure and differential item functioning.

**Relations to other variables.** Fourteen out of 30 studies (46.6%) in this validation synthesis examined relationships with other variables evidence (Table 2). *The Standards* (AERA, et al., 2014) specify that, when presenting analyses of test items with data on other variables, the rationale for selecting said variables should be reported. Additionally, evidence concerning the constructs for additional variables should be presented or cited.
Of the 14 studies that examined relationships with other variables, 11 (78.5%) presented a rationale for selecting constructs or measures to compare with the RSES scores, and 1 (7.1%) was unclear about the rationale. For the studies that did not provide an explicit rationale, it was implied that the constructs chosen were related to self-esteem. As to why specific measures were chosen (as opposed to any other measures), there was no clear rationale provided.

**Coding check for coding relations to other variables.** The second examiner did not encounter coding discrepancies, and she noted similar themes to those presented in the current study. She noted there was lack of clarity in the description used in studies, a lack of reported values in their entirety -such as reporting all correlation values with p values or loadings when using factor analyses to determine if convergent and discriminant validity evidence load on different factors, the absence of criteria to support claims in the studies, and a need for more specificity in descriptions and interpretations.

**Convergent and discriminant evidence.** Of the 14 studies that presented validity evidence based on relationships with other variables, 12 (85.7%) used convergent measures and 2 (14.2%) used both convergent and discriminant measures. The number of measures used per study to be compared against the RSES scores ranged from 1 to 7 (\(M=3.5, \ SD=1.8\)). Overall, reported evidence was unclear and inconsistent in the description of the process of convergent and discriminant validation. A total of 7 different ways of referring to this process were found. Of the 14 studies, 4 (28.5%) described their method as correlational analyses, 3 (21.4%) described this process as convergent or discriminant evidence,
3 (21.4%) used the term 'construct validity', 2 (14.2%) described it as 'relations with other constructs', 2 (14.2%) described it as 'establishing predictors', 1 (7.1%) described it as 'robustness across methods', and 1 (7.1%) described it as 'concurrent validity'. In one study, the authors described their study as providing criterion-related evidence; however, this is a misidentification of convergent evidence as criterion-related, as the study examined the correlation between the RSES and four scales (Taylor Manifest Anxiety Scale, the Eysenck Personality Questionnaire, the UCLA Loneliness Scale, and the Beck Depression Scale).

Overall, studies were clear on establishing how the scores of other measures should correlate to the RSES scores. Out of the 14 studies, 12 (85.7%) stated clearly their expectation of correlates, and 2 (14.2%) were unclear. From the 12 studies that only examined convergent validity evidence, 9 (75%) explained that measures were expected to correlate, 6 (50%) described if the correlation should be positive or negative, and 4 (33.3%) expected the correlations to be statistically significant. From the two studies that examined both convergent and discriminant measures, one did not describe their expectations and the other explained that the RSES scores should correlate higher to mental health than physical health.

With respect to interpreting results, most studies examined the magnitude of the correlation (12/14, 85.7%), 7 (50%) interpreted the sign of the correlation, and 9 (64.2%) examined the statistical significance of the correlation.

Other validation approaches for relations to other variables. Aside from correlational analyses, 4 studies (out of 14, 28.5%) used other methods to examine the relationships of RSES scores with other variables. One study used a MTMM
matrix; in another study, a stepwise multiple regression analysis using the RSES scores as the outcome variable and the proposed correlates as predictors was conducted; and, in two studies, factor analysis was conducted to determine factor loadings of the correlated predictors. The description of these methods and their purpose was not clear; for example, the study that uses the MTMM describes it as “This multitrait–multimethod design (Campbell & Fiske, 1959) is likely to increase the validity of results” (Alessandri, Vecchione, Eisenberg & Laguna, 2015), but does not explain how validity evidence is demonstrated.

Table 3. Sources of reliability and validity by language version

<table>
<thead>
<tr>
<th>Language</th>
<th>N (%)</th>
<th>Reliability</th>
<th>Response Processes</th>
<th>Internal Structure (EFA)</th>
<th>Internal Structure (CFA)</th>
<th>Internal Structure (other)</th>
<th>Relations with other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>1 (3.3)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Dutch</td>
<td>1 (3.3)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>English</td>
<td>15 (50)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>French</td>
<td>2 (6.6)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Italian</td>
<td>5 (16.6)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Persian</td>
<td>1 (3.3)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1 (3.3)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Slovak</td>
<td>1 (3.3)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Spanish</td>
<td>3 (10)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
</tbody>
</table>

Table 3 shows the sources of reliability and validity evidence available in each language. The Chinese sample consisted of patients with a diagnosis of Myocardial infarction and unstable angina. The Dutch sample consisted of
volunteers who were tested in the context of two other studies investigating hurt feelings. The English sample included adolescents in grade 9 through 12, undergraduate psychology students, African-American single mothers, female nurse assistants, adults living in USA adults over 60 years of age, wheelchair basketball undergraduate students, students from a Hispanic background, ex-prisoners, and Pakistani undergraduate students. The French samples included university students and retired volunteers. The Italian samples included a cross-sectional nationwide survey, and undergraduate students. The Persian sample consisted of undergraduate Iranian students living in USA. The Portuguese sample consisted of adolescents between 15 and 20 years of age. The Slovak sample consisted of high school and university students. The Spanish samples included a random sample of general population, high school students, university students, and a group with a diagnosis of social phobia.

Order in which reliability and validity evidence is presented. Studies examined between 0 and 2 sources of reliability evidence ($M=1.03$, $SD=0.61$). Studies examined from 1 to 3 sources of validity evidence ($M=1.63$, $SD=0.76$). As seen in Table 4, most studies report reliability first, as seen in 19/30 (63.3%) of studies. Of the 30 studies examined in this synthesis, 4 (13.3%) studied a single source of validity evidence. Of the 26 studies that examined multiple sources of validity and reliability evidence, only 7 (26.9%) examined their data in an order that accounted for a logical sequence as described by Slaney et al. (2009), and 7 (26.9%) followed a logical sequence only partially. Internal structure is reported
before relations with other variables (6 out of 26 studies, 23%) more often that
relations with other variables before internal structure (4 out of 26 studies, 15.3%).

Table 4. Order in which evidence was examined¹

<table>
<thead>
<tr>
<th>Order of examination for the source of evidence</th>
<th>n</th>
<th>(%)</th>
<th>Logic²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability → internal structure</td>
<td>11</td>
<td>(36.6%)</td>
<td>No</td>
</tr>
<tr>
<td>Reliability → relations with other variables → internal structure</td>
<td>3</td>
<td>(10%)</td>
<td>Partially</td>
</tr>
<tr>
<td>Reliability → internal structure → relations with other variables</td>
<td>3</td>
<td>(10%)</td>
<td>Partially</td>
</tr>
<tr>
<td>Reliability → relations with other variables</td>
<td>2</td>
<td>(6.6%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Internal structure → reliability</td>
<td>2</td>
<td>(6.6%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Internal structure → reliability → relations with other variables</td>
<td>2</td>
<td>(6.6%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Internal structure → relations with other variables</td>
<td>1</td>
<td>(3.3%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Internal structure → reliability → internal structure³</td>
<td>1</td>
<td>(3.3%)</td>
<td>Partially</td>
</tr>
<tr>
<td>Relations with other variables → reliability → internal structure</td>
<td>1</td>
<td>(3.3%)</td>
<td>No</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

¹Only studies that examined more than one source of evidence are included in this table
²Logical order as described by Slaney et al. (2009)
³Study examined Many-Facet Rasch Model analysis, Cronbach’s alpha, and then internal principal component analysis.

Regarding a validation plan as suggested by Chan (2014) -which suggest studies should decide on a view of validity that is appropriate to the inferences that are expected from their study-, 15 (50%) of the 30 studies did not report any theoretical or conceptual orientation or a guideline such as \textit{The Standards} (AERA et al, 2014). From the studies that reported any such a plan, 11 (36.6%) based their guidelines from previous findings and prior research methodology, 2 (6.6%) were replicating a previous study, and 2 (6.6%) presented a clear theoretical orientation
plan (e.g., presenting theory on measurement invariance as a contribution to validity based on internal structure, or explaining how IRT contributes to the examination of response processes).

**Discussion**

The main goal of this thesis was to describe the sources, methods, and procedures of reporting practices for reliability and validity evidence for the RSES. To assess this, three objectives were described: (1) to describe the sources used to provide reliability and validity evidence for the RSES scores as categorized by *The Standards* (AERA et al., 2014), (2) to determine if a logical sequence was being followed when presenting these sources and if researchers appeared to have considered a theoretical orientation plan, and (3) to highlight the gaps in the literature on sources of validity evidence that have not been examined.

The findings of this study produce a detailed summary of the reporting practices for reliability and validity evidence over the life of the RSES. Similarly to previous works of validity synthesis (e.g. Chinni & Hubley, 2014; Cizek et al., 2010; Hogan & Agnello, 2004), this study uses *The Standards* (AERA et al., 2014) as a guide for validity sources and validation methods and procedures consistent with current validity theory. It also builds on the methodology for validation syntheses described in the book “Validity and Validation in Social, Behavioral, and Health Sciences” (Zumbo & Chan, 2014). This study is notable in that no other study has performed a reliability or validity synthesis on the RSES, and only a few studies...
have focused on a single scale (e.g. Chinni & Hubley, 2014; McBride, Wiens, McDonald, Cox & Chan, 2014).

Findings from this study are applicable to future validation works in general, by providing insight on current trends and underlining gaps, such as the need for homogeneity and clarity in the language used to report validity evidence based on relations to other variables. This study focuses solely on sources of reliability and validity evidence for the RSES and, because of that, every relevant study was included. Furthermore, by focusing on a single scale, this study is able to provide thorough summary of reliability and validation procedures regarding this scale.

Half of the studies examined were based on translations of the RSES, a third of which consisted of newly translated versions. These translation processes are unclear; this vagueness in the translation procedure has also been observed by Chinni and Hubley (2014), and Hubley et al. (2014). This lack of information on the adaptation process for the scale might be troublesome, as it implies a lack of consideration for how reliability and validity may be affected. In essence, a translated or modified version of a scale is a new measure and, as such, it is important to document clearly the process that led to this version of the test as this may provide insight into possible sources of construct irrelevant variance or content underrepresentation within a different language or cultural group. Furthermore, reliability and validity sources examined for the original RSES will not necessarily hold true for these translations, which is not acknowledged in any of these studies.

Overall, information provided on the translation process for the RSES was brief or poorly described. In 12 (80%) of the studies using a translated version of the
RSES, these translation works are cited but no description of the process was given. Only 1/15 (6.6%) translations described their procedural terms, and no study reported use of pilot testing. No study describes these translated versions of the RSES as independent from the original for the purposes of providing reliability or validity evidence.

Reliability evidence. Reliability evidence for the RSES scores was found in 83.3% of studies and stated clearly as such or simply reported as alphas. Internal consistency coefficients are the most used, reported in all studies that examined reliability. Cronbach’s alpha is the dominant estimate, being reported in 76% of the studies that reported reliability. This finding is consistent with previous studies such as Chan, Munro et al. (2014), Chinni and Hubley (2014) and Hogan et al. (2000) and shows that classical test theory is still a dominant approach for reliability estimation for RSES scores. Studies stated a criterion to acknowledge adequate or satisfactory reliability in 44% of cases. Little to no explanation was provided about the implications of reliability for subsequent validation or other statistical analyses. One possible interpretation for the manner in which reliability is reported is that it is an expected standard in research and generally well understood.

Test-retest reliability evidence was the presented source of reliability evidence in 20% of studies, and was presented along with an internal consistency coefficient in all cases. The intervals presented were described clearly in most cases, but there was no clear explanation on why such interval was chosen, although it is possible that in some cases the time period was selected for convenience.
**Validity evidence.** Of the five sources of validity evidence described by *The Standards* (AERA et al, 2014), only two sources of validity - internal structure and relations to other variables - were found in this study to have been sought with the RSES. No studies examined evidence based on content or consequences of testing.

*Test content.* There were no studies that examined validity evidence based on test content. While validity evidence based on content is one of the first identified sources of validity by *The Standards* (AERA et al, 2014), other validation syntheses (e.g. Chinni & Hubley, 2014; Cizek et al., 2008; Hogan & Agnello, 2004) have shown it is not commonly relied upon. One possible explanation is the concern for publication, as content validation might be considered by researchers or reviewers to be a process built into the development of a test and not worthy of a publication on its own. Another possibility might be that researchers consider other sources of evidence more relevant and prioritize them. One more explanation could be that *The Standards* lack practical guidance and so it is difficult to adapt to practical research (Shear & Zumbo, 2014). However, the lack of content validity evidence is a concern, as it is important to consider how the scale measures up to current self-esteem theories, and the content relevance it has in a modern setting.

*Response processes.* This source of evidence was only present in one study (3.3%) and used IRT methodology to present a detailed examination of each individual item and the association between given responses and the underlying trait measured (Gray-Little, Williams & Hancock, 1997). The use of IRT to examine the RSES provided a detailed description of the scale items and factor structure, but there was little discussion of the theoretical fit between the response processes of
the sample with the self-esteem construct and theory. The lack of studies that examine response process has been noted in many studies (e.g. Chan, Zumbo, Darmawanti & Mulyana, 2014; Cox & Owen, 2014; Chinni and Hubley, 2014, Cizek et al., 2008; McBride et al., 2014, Sandilands & Zumbo, 2014).

While *The Standards* (AERA et al, 2014) do not contemplate IRT as an example of a method to demonstrate validity evidence based on response process, it describes possible evaluation methods that match the inferences intended by the researchers in this particular case (e.g., analyses of the relationships among parts of scales, examining individual responses and the trends in said responses). Overall, the literature does not appear to report IRT as a typical approach, but neither is it excluded from being a validation procedure implicitly or explicitly.

Because there is only one study examining response process, this sample cannot be said to be necessarily representative or generalizable. Having few studies that examine validity evidence based on response process can highlight the gaps in literature, but it would be difficult to claim that all observations from the single study are representative of the studies that do examine this source of evidence. More studies that examine the validity evidence based on response processes of the RSES are necessary; Zumbo et al. (2014) note that this source of evidence is a substantive feature of validity evidence that requires further examination, Collie and Zumbo (2014) also remark that this source of validity evidence helps understand individual and cultural differences in the process of answering tests, and can assess whether the items are adequately measuring the intended construct or some other interpretation is taking place.
*Internal structure.* Validity evidence based on internal structure is very commonly reported and was present in 90% of the studies; this is comparable to results found in previous validity syntheses (Chinni & Hubley, 2014; Cizek et al., 2008; Hogan & Agnello, 2014). The majority of these studies used CFA (62.9%) to examine the internal structure of the RSES, which has been observed in previous studies such as Cox and Owen (2014) or Chan, Zumbo, Chen et al. (2014). Most studies (82.3%) presented two-factor models, and described their reasoning for the proposed models. A majority (90%) of studies were clear about the expected factor structure and the loading of items for more than one factor, as well as the fit indices chosen to assess fit. One issue was the lack of an explanation or rationale for choosing fit indices; this is a problem because different fit indices vary due to different factors, for example, NFI is sensitive to sample size and model complexity (De Wulf, 1999). Cutoffs for fit indices were also not specified in 29.5% of studies, which can make it difficult to evaluate some results.

EFA was only present in 29.6% of studies that examined internal structure, which perhaps makes sense for a measure that has been in the literature for such a long time. It could be argued that there is a need to use EFA to examine the factor structure of translated and adapted versions of the RSES, but this explanation was not found in studies. In studies using EFA, only two studies (25%) found a bi-factor structure; neither of these studies specified what rotation method was used. This is a discrepancy from the CFA proposed models which tended to toward two-factor models, whereas EFAs tended to support a single factor.
Another issue in the reporting of EFA is that 87.5% of studies performed a Principal Component Analysis (PCA). While this method is related to EFA, it is not an estimation method of common factor analysis, and is more appropriate as a data reduction technique because it does not differentiate common and unique variance (Brown, 2015); other methods such as common factor analysis or principal axis factoring might be more appropriate.

The criteria used to determine the number of factors in EFA was clearly stated in 62.5% of cases, but was unclear in 25% of cases. Only 40% of cases appeared to establish the criteria a priori. Factor loadings were only reported in 50% of studies; no study reported a criterion to interpret the factor loadings. These findings underline the need for clarity in the reporting of factor analysis in research for the RSES in future studies.

There was only one instance of a study that examined both EFA and CFA. This study used the same sample for both factor analyses and, unsurprisingly, found that the structure suggested by their EFA was supported in the CFA. In order to obtain more significant data, an EFA and CFA should be performed on two different samples from the same population. Because there is only one study that performs both types of FA, it is difficult to generalize observations relevant to studies that perform both EFA and CFA.

Measurement invariance was examined in 20% of cases, and differential item functioning in only one study. Overall, descriptions of these methods were clear. Studies that described MI stated their comparisons of model invariance as structural level, weak, and strong invariance.
Overall, methods used to describe validity evidence based on internal structure were explained clearly. It stands to reason that adequate reporting practices are better known and more likely to be held because of how commonly the methods are examined (Chinni & Hubley, 2014; Cizek et al., 2008; Hogan & Agnello, 2004). However, there appears to be a lack of detail on the description of a rotation method, the selection of fit indices, and selection of cut-off values; a report for a factor analysis should include information that allows for a clear interpretation for the model, including the criteria to identify a model, the fit indices used, parameter estimates, and the rationale followed to make decisions (Brown, 2015). It is possible that these issues stem from a lack of clarity on how factor analyses can explain validity and are instead considered as validity itself, and that validity could be taken for granted (Gunnell, Wilson et al., 2014). It is also possible that internal structure is examined often not because of its relevance but rather because of it is opportunistic evidence as described by Sandilands and Zumbo (2014), lacking a theoretical rationale for interpretation.

Despite the large number of internal structure analyses, it must be remembered that almost half (41.6%) of the studies that examined the dimensionality of the RSES did so on translated versions of the test. Cox and Owen (2014) comment that it is important to consider the representativeness and adequacy of the samples examined to the validation of scales, as validity is dependent on context.

*Relations to other variables.* Relations to other variables was the second most commonly found source of validity evidence and present in 46.6% of studies.
Most (78.5%) of these studies provide a reason for why they chose the constructs they did to compare to the RSES, and 85.7% provided an expectation for how they thought measures should correlate. A majority (85.7%) of the studies that examined correlations only provided convergent coefficients, explaining that measures should correlate, some indicating expected sign of the correlation as well, and only a few explicitly expecting the correlations to be significant. Additionally, 14.2% of studies examined both convergent and discriminant evidence, but only one study described an expectation for how measures should correlate. Studies tended to be vague in their interpretations of obtained correlation values, such as “academic motivation [was] positively associated with both RSES factors” (Supple, Su, Plunkett, Peterson & Bush, 2013, pp. 760).

Previous studies have also found that while relations to other variables is commonly examined, it is mostly convergent evidence, whereas discriminant evidence is not examined as commonly (Cizek et al., 2008; Hubley et al., 2014; Collie & Zumbo, 2014, Sandilands & Zumbo, 2014).

The description of convergent and discriminant validity methodology was vague and heterogeneous. This is similar to findings from previous validation syntheses, such as Chan, Zumbo, Darmawanti and Mulyana (2014), Chinni and Hubley (2014), Gunnell, Schellenberg et al. (2014), and Hogan and Agnello (2004), which noted ambiguity in the language used to report correlations. This study identified seven different ways this evidence is referred to, the most common being correlational analysis, construct validity, and convergent/discriminant evidence.
Of the studies that examined relations to other variables, four (28.5%) used other methods in addition to correlations. Of those four studies, one was a multitrait-multimethod (MTMM) matrix, another was stepwise multiple regressions using the variable factors to confirm their utility as predictors, and two studies performed factor analyses to determine factor loading of the correlated predictors. While MTMM matrices are not a commonly used method, it has been found in previous research and is recommended for examining validity evidence based on relations to other variables on previous studies (Chan, 2014; Gunnell, Schellenberg et al., 2014; Sandilands & Zumbo, 2014). Using more than one method to assess validity of the RSES scores can reduce the bias from method effects (Campbell & Fiske 1959). However, it is important to be clear in the reporting of these methods, such as describing every step of the methodology followed, the criteria used to draw inferences, and the how the obtained results contribute to validity evidence for RSES scores.

One study described the RSES as a criterion measure, but performed a correlation between the RSES and four other scales. This highlights a possible misunderstanding of the use of term “criterion” when examining validity evidence based on relations to other variables. While this misuse of the term was not common in the studies examined, it has been found in validation research (Hubley, Zhu, Sasaki & Gadermann, 2014). This issue adds to the previous evidence that researchers need to be more careful with their chosen language, as well as clearer and more specific in their reporting practices.
Consequences of testing. There were no studies of the RSES that addressed consequences of testing. While the importance of validity evidence based on consequences has been challenged by some authors, such as Cizek et al. (2008), other authors have described the utility of this source of validity (Hubley & Zumbo 2011, 2013; Messick 1995b). The Standards (AERA et al., 2014) consider validity evidence based on consequences to be an important source of evidence to provide relevant interpretations of test scores, as it can trace sources of invalidity such as underrepresentation or construct-irrelevant components, as well as the evaluation of adequacy for the claims and interpretations based on the test scores. Because of the widespread use of the RSES, validity evidence based on consequences should be considered, as it can be a source of information for policymakers and healthcare providers, as well as educational settings such as programs to build self-esteem in students. It is to be noted that, while self-esteem is a popular construct examined in psychology and education, it is not known to us that the RSES has been used as an indicator for policy-making. This reduced concern on the use of the RSES added to the debate on the utility of validity evidence based on consequences of testing might explain why this source of evidence is absent.

General observations. The majority (83.3%) of studies examined in this synthesis reported reliability for their sample, and were clear on describing what source of reliability was calculated. Similarly, when describing the factor analysis performed, studies are overall clear describing the type (e.g. PCA, EFA, or CFA); the language utilized for these studies was generally homogeneous and unambiguous. Models examined for CFA were usually detailed and conclusions
about the factor structure of the RSES were discussed clearly. Multiple fit indices tended to be examined. For validity evidence based on relations to other variables, studies considered a mix of magnitude, statistical significance, and sign of correlations to support their inferences.

There are some concerns based on the observed studies. Overall, the results collected showed that, while most studies describe their contributions to validity, almost a third (30%) do not appear to consider how their studies contribute to the validity of inferences made from RSES scores. This could reflect that researchers fail to understand the implications that their studies have on validity or there is a disconnection between their studies and theory as a whole.

Clarity and description of some of the methodology utilized seems to be another concern. There are many studies that do not report all relevant findings or use inadequate or outdated terminology such as face validity (Ark, Ark & Zumbo, 2014). Establishing assumptions and interpretation guidelines a priori should be present in studies.

Logical sequences of reliability and validity evidence. The order in which studies report reliability and validity evidence suggests that there is little consideration given to the implications that internal structure has on reliability and validity. If a factor structure shows two different latent variables (i.e. a bi-factor model), reliability should be examined for each factor. In 63.3% of the studies in this synthesis reliability tends to be examined first, and factor structure second. Similar findings were reported by Slaney, Tkatchouk, Gabriel and Maraun (2009).
Studies also tend to present evidence that is more easily examined with computations, such as examining reliability with an internal consistency coefficient or model fit with fit indices. This can also be seen when a study reports convergent and discriminant validity evidence using correlation values without discussing the expected findings and how they fit with theory. Validity evidence based on content, response processes and consequences is harder to interpret and analyze, which could explain the lack of studies examining these sources. The absence of validity evidence based on consequences of testing also reflect the findings by previous research such as Cizek et al. (2008) and Chinni and Hubley (2014), which report a large absence in this source of evidence. While Cizek et al. (2010) suggest a lack of this source of evidence suggests it has no place in validity evidence sources, Shear and Zumbo (2013) argued that this absence is due to the difficulty of incorporating validity evidence based on consequences in a validation study. Furthermore, arguing that absence of a source of evidence means it is not an adequate source of validity seems flawed, as a similar argument could be made about validity evidence based on content or response processes. A more insightful response might take into account the similar characteristics that these sources of evidence present, such as a less standardized method to assess these sources of evidence or the need for an in-depth discussion of what findings might be expected, and how they fit with the available theory.

**Recommendations for future validation research.** One of the most common trends in research on the RSES is treating translated or adapted versions as equal to the original English version RSES in terms of using reliability and validity
evidence. Studies should consider that modifications to the scale result in what can be considered an entirely different measure and should be examined as such. Translations should also include detailed reporting of the process they followed as well as how any adaptations of the scale to the population beyond a simple work of direct language translation took place. Evidence of pilot work conducted is important. Furthermore, validation research should consider the context of the population it samples; as described by Cox and Owen (2014), population and subpopulation should be meaningful and chosen with a logical rationale aimed to the intended use of the test scores.

Studies examining validity should also provide the rationale for selected comparison measures, cutoffs, criteria, sample selection, and other relevant variables. Studies should also state their expectations and supporting theory both for their rationales and their results and interpretations. Methodology and results should be explained in detail and include all relevant information.

Overall, studies seemed to require more guidance in following a structured methodology, including in the use of The Standards (AERA et al, 2014) on their research. A lack of clarity in the methodology and language of some studies could be observed, particularly in those dealing with validity evidence based on relations to other variables. It would benefit reliability and validity research to have an in-depth description of possible methods of assessing reliability and validity evidence that expands on the suggestions set by The Standards.

**Recommendations for RSES research.** Reliability and validity evidence related to the RSES has been examined in various ways. The most common
methods are reliability evidence based on internal consistency, and validity evidence based on internal structure (both for the English version as much as for translated versions). Less common but still present is reliability evidence based on test-retest estimates, and validity evidence based on relations to other variables. There is one study that examined validity evidence based on response processes. Given this, some of the gaps in the literature can seem obvious; there is a lack of studies on validity evidence based on content, response processes, and testing consequences.

A less obvious gap is found in the examination of relations to other variables, as there seems to be little examination of discriminant sources of evidence to pair with convergent evidence. There is also some concern about the examination of factor structure, as 41.6% of this evidence was examined on translations, so it is worth considering if the evidence on the English version is sufficient. For studies using EFA, consideration needs to be given to using common factor analysis methods rather than PCA.

Studies should also consider how their findings contribute to validity theory as a whole, and to the support inferences made from RSES scores. A validity plan as described by Chan (2014) should be considered, examining the appropriateness of the approach and utility to overall validity. This validity plan should make the purpose of the validation study clear, give a sense of what the study has not presented, and allow the reader to judge the strength of validity evidence (Zumbo et al., 2014). Furthermore, as part of the planning of a study, when examining multiple sources of reliability and validity studies should plan a logical order to follow as
described by Slaney et al. (2009), i.e. factor analysis before reliability, and reliability before relations to other variables.

There are certainly gaps in the literature, but the RSES is a scale that continues to be scrutinized even to this date. Reliability and validity evidence is examined for the English version and for multiple translations, as well as the use of the scale in various population groups with different characteristics. While modifications and translations exist, the revised version from 1989 is still a commonly used scale, and, as such, the RSES continues to be relevant in research. As such, it is important to know the sources of reliability and validity evidence provided thus far and consider how well that evidence is reported and what gaps in the evidence currently exist.
Chapter 4 – Concluding Comments

Reliability and validity evidence are two essential parts of psychometric evidence. Reliability is a theoretical property of scale scores obtained from a given sample for a given purpose, while validity describes the quality of inferences made from scale scores supported by theory and empirical evidence (Chan, 2014). A test that lacks both reliability and validity would be unfit to use. Reliability ensures that observed scores will remain constant through repeated measures, and allows us to calculate a point estimate of true scores and confidence intervals (Furr & Bacharach, 2014). A measure with poor reliability might limit the possible magnitude of correlational analyses used for research, impact validity evidence (especially that based on internal structure and relations to other variables), and alter the variance and skewness of scores. Evidence of validity ensures that the score interpretations are meaningful, useful and appropriate (Messick, 1995a); because of this, having poor validity can lead to misinformed, wasteful, or even harmful decisions based on the inferences drawn from test scores (Furr & Bacharach, 2014; Messick, 1995b).

Evidently, the importance of reliability and validity has not gone unnoticed, and reliability and validity theory has grown and developed greatly over the years with contributors such as Chan (2014), Hubley and Zumbo (2011; 2013), Messick (1989; 1995a, 1995b), and The Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014). Thus, the more evidence on these psychometrics become available, and because of this, it becomes increasingly difficult to keep track of what reliability and validity evidence has done, what gaps are there to cover, and what are relevant sources available for particular tests.
Reliability and validity syntheses provide an analysis of the sources for reliability and validity evidence. They can be a useful tool to observe what has been done to examine these psychometrics, point out the gaps in literature, report issues in the reporting practices on reliability and validity, and provide guidelines or suggestions for future studies.

Previous reliability syntheses have shown that internal consistency reliability using a Cronbach’s alpha was the most common method of assessing reliability (Hogan, Benjamin & Brezinski, 2000); this held true when examining reliability for the RSES. One common finding in reliability syntheses of journals was that reliability was reported for the study’s sample in about a third of studies (Thompson & Snyder, 1998; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). However, in a reliability synthesis on the Satisfaction with Life Scale (Chinni & Hubley, 2014), reliability was reported on 78.3% of cases. This is more comparable to the current study, in which reliability for the RSES was estimated in 83.3% of the studies.

Validation syntheses seem to agree that the most common source of validity evidence is that based on internal structure and relations to other variables, with significantly less evidence based on response processes and rarely based on content or consequences of testing (Chinni & Hubley, 2014; Cizek, Rosenberg, & Koons, 2008; Hogan & Agnello, 2004, Zumbo & Chan, 2014). The evidence found in this study seems to support the previous findings. Interestingly, the use of item response theory as a source of validity evidence based on response processes appeared once in this study; this approach does not seem to be contemplated by The Standards (AERA et al, 2014), but meets the criteria to support this source of
evidence. This was similarly documented before by Chinni and Hubley (2014), where an IRT study was presented in one occasion (3%) as a source of validity evidence based on response process.

The studies that were presented here are the entirety of those that met the inclusion/exclusion criteria rather than a random sample of such articles, and, as such, the analysis are descriptive statistics, not inferential. Similar to the analyses presented in Zumbo and Chan’s (2014) book, this study presents a systematic review of the included papers, focusing on analyzing frequencies of the characteristics of the efforts of validation and reliability work related to the RSES. This analysis described the language version of the RSES, a general description of the samples used, the sources of reliability and validity evidence reported, the methodology used to obtain this evidence (such as factor analysis, test-retest correlations), the order in which these analyses are performed, and whether there was a “validation plan” present in the studies.

In the current study, reliability was usually presented briefly, and only 44% of studies reported a criterion to establish reliability. For test-retest reliability, the reasoning for the chosen retest interval was not discussed. It seems a positive note that the majority of studies are reporting reliability, although this lack of detail gives the impression that reporting an alpha is taken as a mere formality. Validity evidence based on internal structure presented arguments about the dimensionality of the RSES, debating the possibility of one or two factors. All five studies that reported using an EFA were, in fact, using a PCA. All of these studies identified clearly their criteria to determine the number of factors, and two out of five studies examined
more than one criterion. Studies that examined the RSES with a CFA were, in general, clear and descriptive about the number of factors and models expected; most of these studies considered models with two factors, but only six retained models with two factors. Furthermore, CFA studies used several fit indices in their samples, the majority using between five and six indices. Overall, studies that performed a factor analysis were clear and descriptive in their methods, and followed a clear methodology, although these studies rarely explained their rationale for selecting criteria to identify the number of factors or to select fit indices. It could be noted that studies tended to favor the use of factor analyses, and this method seemed the most homogeneous and standardized, reducing methodological errors; however, studies seldom discussed how their findings are relevant to validity or discuss the implications of their analyses on validity or reliability evidence.

A concerning finding in validation syntheses is that some studies have been found to use vague or inadequate language when presenting their findings (Chinni & Hubley, 2014; Cizek, Bowen & Church, 2010; Hogan et al., 2000; Slaney et al., 2010; Slaney, Tkatchouk, Gabriel & Maraun, 2009; Willson, 1980), such as the use of the term “criterion” to refer to convergent validity evidence, or in some cases to all sources of evidence. This study found similar heterogeneity in the language used, predominantly with validity evidence based on relations to other variables.

Similarly troublesome is the lack of consideration to the implications that internal structure validity evidence has on reliability, and the implications of reliability on validity evidence based on relations to other variables. Slaney et al. (2009, 2010) noted that very few studies consider a logical sequence for their methodology. This
study observed that few studies followed such a logical sequence, although most studies report reliability first and briefly, followed by validity evidence.

Studies that examined sources of reliability and validity should consider following a more planned approach to their methodology, considering the implications of their methodology and their results. Zumbo and Chan (2014) suggest that many studies that report reliability or validity evidence are “haphazard or opportunistic” (pp. 323); haphazard because of a perceived lacked of planning or order, and opportunistic because some studies appear to focus on an independent area of research and use the available data to report reliability and validity as supporting a secondary goal. This can be observed in the language used in the studies examined in this research, as only 40% of studies reported their studies explicitly as a validation study, while 30% of studies only described how their results contribute to validity but were not strictly a validation study, and the other 30% of studies did not report their studies as supporting validity.

**Study Strengths and Limitations**

This study builds on previous works of research synthesis to consolidate the presented validation efforts, and focuses on the validation and reliability evidence of a single scale. This study examines only the RSES, which is a longstanding and commonly used scale, and as such synthesis of reliability and validity evidence of its scores in various populations is sure to be of interest to various researchers interested in the RSES, measurement experts and those interested in reliability and validity syntheses. Furthermore, focusing on a single measure allows us to see the
parallels between reliability and validity syntheses that focus on journals and syntheses that focus on single scales.

Because this study used only those articles relevant to the RSES, it has a smaller relevant population that can actually be examined in its entirety. Because of this, the conclusions obtained from this study are gathered from the population, as opposed to inferences based on a sample, and provide a summary of reliability and validation procedures used to date for the RSES.

Another strength of this study is the use of The Standards (AERA et al, 2014) as a methodological approach. The Standards is a longstanding set of standards that are well-recognized, and have been adopted by various researchers as an accepted view of validity (e.g. Barry, Chaney, Piazza-Gardner & Chavarria, 2013; Hogan & Agnello, 2004; Jonson & Plake, 1998; Pawlowski, Segabinazi, Wagner & Bandeira, 2013; Qualls & Moss, 1996; Slaney et al., 2009; Whittington, 1998).

Adopting The Standards in this study provides a view of validity theory that Chan (2014) advocates for the idea of a “validation plan”.

There were studies that did not necessarily fit the categories described by The Standards (AERA et al., 2014), but still provided reliability and validity evidence; for example, the use of IRT (Gray-Little, Williams & Hancock, 1997) or person’s reliability index value (Classen, Velozo & Mann, 2007). While these approaches are not contemplated in The Standards, they still provide reliability and validity evidence and were included here.
A limitation that comes from focusing on a single scale is that the studies examined in this research might not be representative enough of the general trends for reliability and validation, as the group of studies that examine the RSES is a very specific kind of sample. The results found in this study are compared to previous findings in other reliability and validity syntheses to provide a sense of reference, but it is advisable to remember results may differ when applied to other measures or on a broader scope.

Another possible limitation on the methodology is that, in general, reliability and validation syntheses that focus on a single scale might be limited in the amount of studies available that focus on that scale; in the case of this particular study, the sample of 30 studies was considered large enough to require such analysis, but future studies should consider this limitation when planning their methodology. Furthermore, the studies examined here were collected only from those presented in English and, as such, left out relevant literature that is only available in other languages.

Another potential weakness in the study is the possibility of subjectivity of the analysis; while coding was done with care and following planned guidelines, the fact that in some studies there is a lack of homogeneity when referring to validity evidence—in particular based on relations to other variables—can lead to possible misunderstandings or missed information. For that reason, a coding check was done to ensure a fair coding of the studies and to determine if the same observations could be drawn from an unbiased second source with a clear understanding of reliability and validity.
One more limitation is the publication bias of positive results over negative ones (Kicinski, 2013). This means that studies that fail to find evidence of reliability or validity evidence for the RSES would not be included because they might not have been accepted for publication.

**Contributions to Reliability and Validation Syntheses**

This study contributes to the field of reliability and validations syntheses by including the classification of reliability as classified by *The Standards* (AERA et al., 2014); previous studies such as Hogan, Benjamin and Brezinski (2000), Hubley, Zhu, Sasaki and Gadermann (2014) and Willson (1980) have examined reliability present in studies, but have not classified this information according to the categories presented in *The Standards*. This study also included an examination of the order in which particular reliability and validity evidence is presented in articles. The examination of evidence presented in a logical order is not often studied even though there are implications for subsequent validation and statistical analyses, similar to the findings by Slaney, Tkatchouk, Gabriel and Maraun (2009). *The Standards* (AERA et al., 2014) note that reliability should be reported for total scores, subscales or factors; thus, it would be inconsistent to obtain a reliability score before a factor analysis. Similarly, Slaney et al., (2009) argue that before being able to compare the scale against other external constructs, it must be adequately assessed for reliability and internal structure validity.

The findings in this thesis about the reporting practices of reliability and validity evidence were focused on the RSES. Similar to previous studies that focus
on a single scale (e.g. Cox & Owen, 2013; Gunnell, Wilson et al., 2014; McBride, Wiens, McDonald, Cox & Chan, 2014; Sandilands & Zumbo, 2014), this study examines the various sources of validity evidence described by *The Standards* (AERA et al., 2014), and examines reliability as classified by *The Standards*. Furthermore, it integrates the analysis of a “validation plan” (Chan, 2014) and the presence of a logical sequence when analyzing sources of reliability and validity evidence (Slaney et al., 2008).

Findings in this study seem consistent with findings from previous reliability and validation syntheses (Cizek et al., 2008; Chinni & Hubley, 2014; Hogan & Agnello, 2004; Slaney et al, 2009), and show how the findings from reliability and validation syntheses based on a sample of articles from journals within a given time frame can be compared to such findings from syntheses focused on individual scales. As such, it strengthens the generalizability of this study, and shows that trends in reporting reliability and validity for journals are analogous to those focused on scales.

The methodology of this study builds on previous works of reliability and validity syntheses, particularly those on the book by Zumbo and Chan (2014), but also expands on the method. The focus on a single scale is not typical for a reliability or validity synthesis, as they tend to focus on publications across one or more journals (e.g. Cizek et al., 2008; Hogan et al., 2000; Hubley, Zhu, Sasaki & Gadermann, 2014; Qualls & Moss, 1996), so this study contributes to the methodology for studies interested on focusing on a single scale. Furthermore, this study incorporates the assessment for a logical order when examining reliability and
validity evidence, as described by Slaney et al. (2009); by doing this, the study goes beyond analyzing what sources of evidence have been examined, and questions how it is examined.

**Contributions to Researchers Using the Rosenberg Self-Esteem Scale**

Researchers interested in the use of the RSES should consider that there is a relative lack of psychometric studies about the RSES compared to the number of articles that reference the RSES; only 27 psychometric articles focused on the RSES have been published since 1989. Only about half of those focus on an English version of the RSES; the rest provide evidence on RSES versions translated or adapted into eight other languages. Given this, there might not be enough evidence supporting the English version of the RSES.

Most reliability evidence is based on an internal consistency estimate (specifically Cronbach’s alpha) and most validity evidence is based on internal structure evidence, followed by relations to other variables. There are still gaps in the literature that require further examination. The lack of validity evidence based on content and response processes in the past 25 years of research is a large oversight that requires further attention. Studies should consider how the RSES holds on current testing and self-esteem theory, analyze the representativeness of the construct of self-esteem and examine the adequateness of the format in which this test is presented. Similarly, examining response processes might yield useful information, including the use of think aloud protocols or cognitive interview.
Studies examining validity evidence based on consequences require one to recognize implied values and uses of the RSES and the possible effects that it may have within a context in a given sample. Research that focuses on it might comment on the way it is applied, and if it is used in the creation of programs, healthcare, policy and decision making, as well as the adequacy of such use.

Regarding the factor structure of the RSES, research appears to focus on examining its dimensionality; however, many of these studies were performed on translated versions of the RSES, and there might be a need for the examination of internal structure validity evidence. Further studies interested in the internal structure of the RSES might want to examine how the factor structure supports subsequent analyses and its interaction with reliability.

As for relationships to other variables, the RSES has been related to various constructs, but discriminant evidence is seldom examined. This suggest an incomplete examination of the RSES, as examining how scales relate on a continuum can provide a more descriptive view of how scale scores relate to other constructs. For example, a study examining the relationships between self-esteem, depression, extraversion and achievement should not only show that these constructs are significantly correlated, but should rather consider the magnitude of these relations to self-esteem relative to each other and how they fit within self-esteem theory. Known-groups evidence and test-criterion might not be adequate evidence sources for this scale, as self-esteem does not have a known criterion that could be used for the purposes of such research.
Studies that further examine internal structure and relations to other variables need to be more descriptive of the information they present as well as include more information on their assumptions and expectations. More specifically, studies examining relations to other variables need to correctly identify their methodology and use standard terminology in order to communicate their findings more effectively.

**Contributions to Measurement**

There appears to be various issues in the way that researchers present reliability and validity evidence. A very commonly found issue is the lack of clarity and thoroughness in the assumptions and the criteria chosen to make decisions and interpretations. This lack of clarity also appears when reporting procedures followed and subsequent results. While there is a possibility that this could be caused by the need to limit the length of articles for publication, it is more likely that these issues stem from a lack of rigour in developing validation studies, misunderstood terminology, a lack of or poor training in measurement, or a lack of a guidelines for studies to follow, as some authors have underlined the need for clearer methodologies and unified guidelines (Cizek et al., 2010; Green, Chen, Helms & Henze, 2011; Hogan & Agnello, 2004; Meier & Davis, 1990).

While *The Standards* (AERA et al, 2014) is a widely accepted resource for measurement specialists, it is not always applied in practice. Previous studies noted the discrepancy between theory and practice when reporting reliability and validity (Qualls & Moss, 1996; Slaney et al., 2009). This might indicate lack of clarity in the
guidelines presented, but it could also be a sign that there needs to be more
emphasis in establishing general guidelines for measurement practice or a need for
more exposure to, and training in, measurement by researchers in general.
References


doi:10.1093/oxfordhb/9780199732579.013.0009

review of seven journals. *Health Education & Behavior, 41*(1), 12-18. doi: 10.1177/1090198113483139


doi:10.1177/0163278702025001003


Meyer-Moock, S., Feng, Y., Maeurer, M., Dippel, F., & Kohlmann, T. (2014). Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC)


Morgan, G. J. (1981). The Rosenberg Self-Esteem Scale as a predictor of the extracurricular activities of summer college freshmen. *(ERIC Document Reproduction Service No. ED 291751).*


doi:10.1177/0829573510375549

doi:10.1177/0734282909335781

*Psychological Assessment, 17*(4), 396-408. doi:10.1037/1040-3590.17.4.396

*Journal of Counseling & Development, 76*, 436-441.


Appendices

Appendix A: List of Articles Included in the Study


Baños, R., & Guillén, V. (2000). Psychometric characteristics in normal and social phobic samples for a Spanish version of the Rosenberg Self-esteem Scale. *Psychological Reports, 87*(1), 269-274. doi:10.2466/PR0.87.5.269-274


Appendix B: Rosenberg Self-Esteem Scale

BELOW IS A LIST OF STATEMENTS DEALING WITH YOUR GENERAL FEELINGS ABOUT YOURSELF. IF YOU STRONGLY AGREE, CIRCLE SA. IF YOU AGREE WITH THE STATEMENT, CIRCLE A. IF YOU DISAGREE, CIRCLE D. IF YOU STRONGLY DISAGREE, CIRCLE SD.

<table>
<thead>
<tr>
<th></th>
<th>1: STRONGLY AGREE</th>
<th>2: AGREE</th>
<th>3: DISAGREE</th>
<th>4: STRONGLY DISAGREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I feel that I'm a person of worth, at least on an equal plane with others.</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>I feel that I have a number of good qualities.</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>3</td>
<td>All in all, I am inclined to feel that I am a failure.**</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>I am able to do things as well as most other people</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>I feel I do not have much to be proud of.**</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>6</td>
<td>I take a positive attitude toward myself.</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>On the whole, I am satisfied with myself.</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>8</td>
<td>I wish I could have more respect for myself.**</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>9</td>
<td>I certainly feel useless at times.**</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>10</td>
<td>At times I think I am no good at all.**</td>
<td>SA</td>
<td>A</td>
<td>D</td>
</tr>
</tbody>
</table>
Appendix C: Coding Sheet

Keywords:
Sample Description:

Translation and Adaptation of Measures

<table>
<thead>
<tr>
<th>Translation and Adaptation of Measures</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translated version of measure used</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was it translated for this study? (vs previously translated)</td>
<td></td>
<td></td>
<td>Partially</td>
</tr>
<tr>
<td>Translation done by an expert?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Describe process:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the measure adapted to the population? (vs just translated)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other comments:

Reliability

Internal consistency:

<table>
<thead>
<tr>
<th>Internal consistency:</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported an internal consistency reliability estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, type of int. consist. estimate (e.g., Cronbach’s alpha, split half (odd/even), KR-20, ordinal alpha):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, criterion used for int. consist. is clearly referenced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, describe criterion (e.g., &gt;.80, &gt;.90; citation):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, list the obtained value(s):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reported inter-item correlations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reported average inter-item correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, was it done for single factors?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reported standard error of measurement (SEM)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test-retest reliability:

<table>
<thead>
<tr>
<th>Test-retest reliability:</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported a test-retest reliability estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, sample used</td>
<td>Full sample</td>
<td>Subsample</td>
<td>New sample</td>
</tr>
<tr>
<td>If yes, reported the retest interval used</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, describe retest interval (e.g., 1 week, 3-41 days):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If yes, provided rationale for choice of retest interval</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
If yes, list the obtained value(s):

### Alternate forms reliability:

<table>
<thead>
<tr>
<th>Reported an alternate (parallel) forms reliability estimate</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>• If yes, describe how this was examined (e.g., Pearson correl.):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• If yes, list the obtained value(s):</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Were there other sources of reliability? Describe:

### Validation

#### General Info:

<table>
<thead>
<tr>
<th>Authors describe study as a validation study (e.g., title, purpose) [strict]</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>• If no or unclear, authors describe evidence as contributing to validity [more lenient]</td>
<td>Y</td>
<td>N</td>
<td>Unclear</td>
</tr>
<tr>
<td>• Comments:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Content

Things to look for:

- Use of experts (subject matter, experiential or practical experts)
- Use of the keywords “construct”, “representativeness”, “relevance” and “appropriateness” in the context of validation

Describe:

### Internal Structure

#### General Info:

<table>
<thead>
<tr>
<th>Conducted a factor analysis</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>• If yes, type of factor analysis</td>
<td>EFA</td>
<td>CFA</td>
<td>Both</td>
</tr>
</tbody>
</table>

#### Exploratory Factor Analysis:

<table>
<thead>
<tr>
<th>Type of EFA</th>
<th>PCA</th>
<th>FA</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not stated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If FA, was the extraction method (see below) identified?</td>
<td>Y</td>
<td>N</td>
<td>Unclear</td>
</tr>
<tr>
<td></td>
<td>Not stated</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Describe extraction method (e.g., PAF, ML, ULS, GLS): 

| Criteria to determine the number of factors was stated a priori | Y | N | Unclear |
| Criteria to identify number of factors clearly stated | Y | N | Unclear |
| • Used criterion of eigenvalues > 1 to decide # of factors | Y | N | Unclear |
| • Used criterion of scree plot to decide # of factors | Y | N | Unclear |
| • Used parallel analysis (PA) | Y | N | Unclear |
| • Used % of variance explained to decide # of factors | Y | N | Unclear |
| • Used another criterion to decide # of factors; describe: | Y | N | Unclear |

• Factor loadings reported 

| Criterion for factor loadings reported | Y | N | Unclear |

• If yes, describe criterion used (e.g., >.30, >.35, >.40): 

| More than 1 factor found | Y | N | Unclear |
| • If yes, factors were rotated | Y | N | Unclear |
| • If yes, type of rotation method used | Oblique | Orthogonal | Both |
| • If orthogonal, specific method used | Varimax | Quartimax |
| • If oblique, specific method used | Promax | Oblimin |

Notes:

**Confirmatory Factor Analysis:**

| Specified the number of factors expected | Y | N | Partially |

• If >1 factor, specified which items load on which factors 

| Specified fit indices (e.g., NFI, RMSEA) used | Y | N | Unclear |
| List fit indices used: | Y | N | Partially |
| • Provided criteria / cut-offs for fit indices | Y | N | Partially |
| • List criteria used for each: | Y | N | Partially |
| • Provided rationale for fit indices selected | Y | N | Partially |

**Measurement Invariance:**

| Examined measurement invariance | Y | N | Unclear |

• Describe software used, estimation method used, fit indices and criteria used, terms used:
Differential Item Functioning (DIF):

<table>
<thead>
<tr>
<th>Examined DIF</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Describe:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Related DIF results to factor/internal structure or multidimensionality

<table>
<thead>
<tr>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
</table>

EFA & CFA Both Used:

<table>
<thead>
<tr>
<th>Did they use the same or different samples/subsamples?</th>
<th>Same</th>
<th>Different</th>
<th>Unclear</th>
</tr>
</thead>
</table>

Notes:

Factors

<table>
<thead>
<tr>
<th>If &gt;1 factor found, was a single score still used?</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
</table>

Notes:

Relations to Other Variables

Other Measures Used:

Total number of measures/subscales used for comparison:

Convergent / Discriminant Evidence:

- How did researcher describe this process (e.g., relations with other constructs/variables, use of terms convergent, discriminant, divergent, 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)?

Type of measures used

<table>
<thead>
<tr>
<th>Convergent</th>
<th>Discriminant</th>
<th>Both</th>
<th>Unclear</th>
</tr>
</thead>
</table>

- Did the researchers provide any rationale for why particular constructs or measures were selected?

- 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?

• Notes/comments:

<table>
<thead>
<tr>
<th>Criterion-Related Evidence:</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher stated intention to provide criterion-related evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mis-identified convergent evidence as criterion-related</td>
<td>Y</td>
<td>N</td>
<td>Partially Unclear</td>
</tr>
<tr>
<td>Type of criterion-related evidence*</td>
<td>Concurrent</td>
<td>Predictive</td>
<td></td>
</tr>
<tr>
<td>Both C&amp;Pr</td>
<td>All 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postdictive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclear</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Describe:

* Concurrent = criterion collected at (roughly) same time as measure of interest; Predictive = criterion collected at a later date than measure of interest; Postdictive = criterion collected prior to measure of interest

<table>
<thead>
<tr>
<th>Known-Groups Evidence:</th>
<th>Y</th>
<th>N</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher stated intention to conduct known-groups evidence or between-group differences as part of validation evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Researcher called known-groups evidence something else (e.g., discriminant validity; criterion validity)</td>
<td>Y</td>
<td>N</td>
<td>Partially Unclear</td>
</tr>
</tbody>
</table>

• Describe:

| Researcher provided evidence of known difference in literature | Y | N | Partially Unclear |
| Researcher just compared groups and did not actually provide validity evidence | Y | N | Partially Unclear |

Other comments:

**Response Processes**

Things to look for:

• Measure of the sample's behaviour (e.g. eye movement or response time)
- Measure of response strategies (e.g. use of think aloud process, cognitive interview of participants, record of responses to items or revisions to answers)
- Analyses of the fit between the response process of the sample with the construct
- Note citations related to response processes (e.g. AERA Standards, Messick, Schwarz)

Describe:

**Consequences**

Things to look for:

- Use of the terms “consequences”, “consequential”, “impact”, “implications”, “effect”, or “values” related to validity
- Consequences misunderstood as test misuse
- Note citations related to consequences (e.g., AERA Standards, Messick, Kane)

Describe:

**Other characteristics**

- What sources of reliability and validity evidence were examined? Describe:
- In what order are these sources examined? Describe:
- Does this order follow a logic sequence? (Internal structure > Reliability > Relations with other variables). Describe:
- Is there an explicit “validation plan”? (Guideline such as The Standards, conceptual or theoretic orientation) Describe:
- Other notes: