

THE GENERALIZABILITY AND RELIABILITY OF SCORES ON THE ROSENBERG  
SELF-ESTEEM SCALE OVER FORTY-EIGHT YEARS

by

Cristin Phibbs

---

Copyright © Cristin Phibbs 2020

A Thesis Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF ARTS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2020

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by: Cristin Phibbs  
titled: The Generalizability and Reliability of Scores on the Rosenberg Self-Esteem Scale Over Forty-eight Years

and recommend that it be accepted as fulfilling the thesis requirement for the Master's Degree.

*Monica K Erbacher*

\_\_\_\_\_  
Monica K Erbacher

Date: Jan 24, 2020

*Adriana Cimetta*

\_\_\_\_\_  
Adriana D Cimetta


Date: Jan 24, 2020

*Heidi Legg Burross*

\_\_\_\_\_  
Heidi Legg Burross

Date: Jan 24, 2020

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement. 

*Monica K Erbacher*

\_\_\_\_\_  
Monica K Erbacher  
Thesis Committee Chair  
Educational Psychology Department

Date: Jan 24, 2020



### Acknowledgments

I would like to express my appreciation to the department of Educational Psychology in the College of Education at the University of Arizona for the opportunity to earn my masters.

Thank you to my thesis advisor Monica K. Erbacher for her supervision, inspiration and comprehensive advice. Thank my advisor Adriana D. Cimetta, for her generous guidance, encouragement and astute feedback. Heidi Burross, thank you for agreeing to be on my committee and providing insightful comments and thought-provoking questions. This thesis would have no have been possible with out the support of these mentors.

I would also like to thank Niamh Wallace, Assistant Librarian in the Research and Learning department at the University of Arizona. Not only did she encourage me to apply to the Educational Psychology Master of Arts program, but she was also instrumental in finding many of my item-level data sources for this thesis.

I would like to thank all my amazing and supportive friends and family. This process would have been impossible without your support.

Thank you to my parents, the late Dr. Ray Phibbs and Barbara Phibbs. No words could ever express how grateful I am to be your child.

**TABLE OF CONTENTS**

LIST OF TABLES.....	5
ABSTRACT.....	6
1. INTRODUCTION.....	7
1.1 Global Self-Esteem versus Related Constructs.....	8
1.2 Common Measure of Self-Esteem.....	10
1.3 Why is Self-Esteem Important?.....	12
1.4 Why is it Important to Measure Self-Esteem Well?.....	14
1.5 Why the Reliability of Scores on This Measure May Have Changed.....	16
1.6 History of the RSES.....	18
1.7 Study Purpose.....	19
2. METHODS.....	21
2.1 Participants and Measures.....	21
2.2 Design Analysis.....	23
2.21 Classical Test Theory (CTT) .....	24
2.22 Generalizability Theory (G-theory) .....	24
3. RESULTS.....	28
3.1 Classical Test Theory Coefficients.....	28
3.2 Generalizability Theory Coefficients.....	20
4. DISCUSSION.....	36
4.1 Implications.....	38
4.2 Conclusions.....	40
5. REFERENCES.....	41
5.1 Literature.....	41
5.2. Data.....	47
6. APPENDIX.....	48

## LIST OF TABLES AND FIGURES

### Tables

Table 1. Data Sources.....	23
----------------------------	----

### Figures

Figure 1. Visualization of the G-study Design.....	27
Figure 2. Cronbach's $\alpha$ scores from 1971 to 2019.....	29
Figure 3. Guttman's $\lambda_2$ from 1971 to 2019.....	29
Figure 4. Percentage of G-study scores variance from 1971 to 2019.....	31
Figure 5. G-study Coefficients: 1-item.....	32
Figure 6. Variance Percentages Across 8-item D-studies from 1971 to 2019.....	33
Figure 7. $\rho^2$ and $\phi$ scores from the 8-item D-study.....	34
Figure 8. Variance Percentages Across 10-item D-studies from 1971 to 2019.....	35
Figure 9. $\rho^2$ and $\phi$ scores from the 10-item D-study.....	36

### Equations

Equation 1. Cronbach's $\alpha$ .....	23
Equation 2. Guttman's $\lambda_2$ .....	24
Equation 3. $\rho$ Coefficient.....	24
Equation 4. $\phi$ Coefficient .....	26
Equation 5. D-study item variance .....	26

### Abstract

The objective of this study is to test the reliability of scores on the Rosenberg Self-Esteem scale (RSES; Rosenberg, 1965) from 1971 to 2019. Self-esteem is how highly one thinks of themselves and how much worth they feel they possess. The RSES is not the only measure of global self-esteem, but it is the most widely used (Whiteside-Mansell & Corwyn, 2003). In the late 1980's and 1990's, the self-esteem movement was enacted in the United States as an effort to improve the lives of adults and children, which may have changed the way self-esteem is interpreted (Humphrey, 2004). For example, items on the RSES may now be measuring narcissism or self-efficacy more so than self-esteem. Thirteen existing item-level datasets that used the RSES were obtained. Sample sizes varied, but all samples contained young adults in the United States between the ages of 15 and 26. Two Classical Test Theory (CTT; Meyer, 2010) coefficients were used, Cronbach's  $\alpha$  and Guttman's  $\lambda_2$ , to test the reliability of scores on the RSES by finding the ratio of true score variance to observed score variance. Generalizability Theory (G-Theory; Shavelson & Webb, 1991) components, specifically G-studies and D-studies, were then used to identify the sources of variance present in scores on the RSES (i.e., variance from persons, variance from items, and remaining unexplained and error variance). The CTT coefficients and G-theory methods indicated scores on the RSES were reliable and, if anything, have increased slightly in reliability over the last 48 years. Despite the reliability of the scores on the RSES, the validity of the scores are still in question. It is important to periodically test the reliability of scores on widely used measures like the RSES to determine the extent to which they can be used for various forms of decision-making (Meyer, 2010) and for various research aims.

The Generalizability and Reliability of Scores on the Rosenberg Self-Esteem Scale  
Over Forty-eight Years

Rosenberg (1979) defined someone with high self-esteem as an individual who has "self-respect, and considers himself a person of worth", conversely an individual with low self-esteem lacks self-respect and considers themselves lacking as a person (Rosenberg, 1979, p. 54).

Individuals with high self-esteem are better equipped to mitigate stress, have greater happiness, and have better relationships than those with low self-esteem (Baumeister, Campbell, Krueger, & Vohs, 2003). High self-esteem is also associated with negative attributes, such as strong in-group association, which can lead to exclusionary behavior, more risky behavior, and perceiving oneself to be more attractive and more popular than they are (Baumeister, et. al., 2003). Thus, self-esteem is an important construct in mental and behavioral health.

Morris Rosenberg created the Rosenberg Self Esteem Scale (RSES) in 1965 to measure self-esteem. Initial analysis of the scores from the RSES indicated these scores were highly reliable (Rosenberg, 1965). In his book "Society and the adolescent self-image", he also discussed the languages the scale has been published in and the contexts in which the scale has been used (Rosenberg, 1965). In another book of Rosenberg's "Conceiving the Self" (Rosenberg, 1979), he provided detailed instructions for Guttman scale scoring, including which items should be considered a single item and how to score these combined items. According to Google Scholar, the first book (1965) has been cited 191 times in the last 5 years, 66 of which were in 2019. The second publication has been cited 2,530 times in the last 5 years, 455 of which were in 2019 (Google Scholar, 2019). The RSES was written 54 years ago but is still widely used. It is important to periodically revisit the reliability of RSES scores to ensure the data collected with this measure are still usable in current research. If scores are no longer reliable

enough, then scores will be inconsistent across occasions and contexts. Thus, the exact time and context in which the scale is administered to a specific sample could have an unreasonably large effect on the resulting scores. This is very problematic for research focused on detecting replicable effects involving self-esteem.

In his thesis, I explore the reliability of RSES scores over four decades, from 1971 to 2019. In the following sections, I first discuss the construct of self-esteem compared to other similar constructs. Second, I review other common measures of self-esteem. Next, I explain why self-esteem is important and specifically why it is important to measure self-esteem well. Then, I discuss one reason why the reliability of scores on the RSES may have changed over the past few decades. Last, I summarize the history of the RSES.

My interest is in the reliability of scores on the RSES over time. I obtained 13 item-level datasets collected between 1971 and 2019 and used Classical Test Theory (CTT) and Generalizability Theory (G-theory) to analyze various forms of reliability (i.e. generalizability and dependability) of scores on the measure. CTT separates observed scores into true scores and error scores,  $X = T + E$  (Meyer, 2010, p. 14), modeling a single source of measurement error. G-theory separates observed scores into true score and multiple other sources, modeling multiple sources of error simultaneously. G-theory also allows the researcher to test different scenarios to find the situation that would produce the most reliable scores for various types of decisions, such as relative decisions (i.e., rank ordering respondents' scores) and absolute decisions (i.e., comparing respondents' scores to a cut-off; Shavelson & Webb, 1991). For example, G-theory allows the researcher to explore how the number of items in a measure affects the reliability of scores.

### **Global Self-Esteem versus Related Constructs**



Self-esteem is most commonly measured globally, and the RSES is the most commonly used measure of global self-esteem (Whiteside-Mansell & Corwyn, 2003). Global self-esteem is usually measured with subjective self-report items, rather than observed behaviors, which prompts individuals to indicate their own relative level of overall positive versus negative feelings about themselves (Robins, Hendin, & Trzesniewski, 2001). Self-esteem is often related to narcissism and self-efficacy as these attitudes are also strongly connected with an individual's feelings of self-interest and sense of self (Baumeister et al., 2003).

Narcissism is an exaggerated positive sense of self accompanied by extroversion with little interest in forming relationships (Twenge, Konrath, Foster, Campbell, & Bushman, 2008). The Narcissistic Personality Inventory (NPI; Raskin & Terry, 1988) is the principal global measure of grandiose narcissism, described as high self-esteem paired with negative interpersonal functioning (Foster, 2015). The original NPI (Raskin & Terry, 1988) is 40-item measure. However, a 16-item measure, the NPI-16, was created to be used in place of the NPI-40 when a shorter measure is more appropriate. Scores on this measure demonstrate internal, discriminant, and predictive validity (Ames, Rose, & Anderson, 2006). The NPI-13 is an even shorter measure that is arguably superior to the NPI-16. Scores on the NPI-13 have good validity and reliability but also allow for the analysis of three subscales (Gentile, Miller, Hoffman, Reidy, Zeichner, & Campbell, 2013). It was once thought that high self-esteem can be good or bad, with bad high self-esteem often described as self-deception or narcissism (Baumeister et al., 2003). Researchers now know that self-esteem and narcissism are their own constructs, the main distinctions between the two are authentic pride and hubristic pride. Authentic pride is associated with healthy self-esteem, a realistic assessment of self. Hubristic pride is a maladaptive component of narcissism, specifically the tendency to overestimate the

degree to which one's personal beliefs line up with the rest of society (Tracy, Cheng, Robins, & Trzesniewski, 2009)

Self-efficacy is an individual's perception of their own ability to successfully perform to produce desired effects. A strong sense of self-efficacy positively influences an individual's likelihood of accomplishing difficult goals (Bandura, 2010). The Generalized Self-Efficacy scale (Sherer, Maddux, Mercandante, Prentice-Dunn, Jacobs, & Rogers, 1982) measures global self-efficacy. Rotter's Internal-External Locus of Control Scale (Rotter, 2011) is used to measure task-specific self-efficacy (Wang & Richarde, 1988). While self-esteem and self-efficacy are distinct constructs, they are related. Individuals who view themselves with high levels of worth generally view themselves as someone who can complete tasks. The important distinction between these two constructs is that self-esteem is a self-perception of one's value and self-efficacy is an individual's beliefs about their own capabilities (Gardner, & Pierce, 1998).

### **Common Measures of Self-Esteem**

According to PsychINFO the RSES has been used in 9,269 studies since 1967 (PsychINFO, 2019). The RSES may be the most widely used measure of global self-esteem, but it is not the only measure (Boyle, 2014). The Single-Item Self-Esteem Scale was developed in 2001 by R.W. Robins as a less time-consuming substitute to the RSES. The single item is "I have high self-esteem" and is measured on a Likert scale from 1 (*not very true of me*) to 5 (*very true of me*; Robins, Hendin, & Trzesniewski, 2001, pg. 153). Scores on this measure had strong convergent validity with scores on the RSES and had similar predictive validity as scores on the RSES, but reliability was not considered since as it is a one-item measure (Robins, et al., 2001, pg. 152).

Another measure of self-esteem is the Self-Linking/Self-Competence Scale – Revisited (SLSC-R), developed in 1995 by Tatarodi and Swann. This measure separates global self-esteem into two components: self-liking or a personal sense of worth and self-competence or a feeling of being capable. This 16-item self-report measure, developed with a 5-point Likert response scale, is comprised of two subscales, one for each component (i.e., self-competence and self-liking). Self-competence is the belief that individuals are responsible for events that happen in their lives and the results of those events. Self-competence is more closely related to self-efficacy than self-esteem. The difference between self-competence and self-esteem is similar to the difference between self-efficacy and self-esteem mentioned above. Self-competence is expectancy of behavior the future, while self-esteem is value of self (Tatarodi, & Swann, 2001). In the context of the SLSC-R, self-liking is how much an individual conducts themselves in accordance with their personal values. While the measure presents self-competence and self-liking as two halves of self-esteem, Tatarodi and Swan define self-liking as synonymous with self-esteem and self-competence as a source of self-esteem (Tatarodi, & Swann, 2001). Tatarodi and Swann (2001) conducted a study using the SLSC-R scale with 1,325 college students from the University of Toronto. Sum scores on these two subscales were positively correlated with correlation coefficient values ranging from .47 to .59. Cronbach's  $\alpha$  values ranged from .70 to .98 for scores on the self-liking subscale and from .56 to .92 for scores on the self-competence subscale (Tatarodi & Swann, 2001). The SLSC-R scale is similar to the RSES in structure as they are both self-assessments, both administered with Likert-type scales, and half of the items on each are reverse scored (Rosenberg, 1965; Tatarodi & Swann, 2001). The RSES items map on more closely to the self-liking items than the self-competence items as the self-liking items ask about attitudes and the self-competence items ask about actions. For example,

the second item on the RSES, “At times I think I am no good at all” (Rosenberg, 1975, p. 291; Appendix A), is very similar to the first item (reverse scored) on the self-liking subscale of the SLSC-R, “I tend to devalue myself” (Tafarodi & Swann, 2001, p. 670). The only RSES item that maps onto the SLSC-R self-competence item is the fourth item, “I am able to do things as well as most other people” (Rosenberg, 1975, p. 291; Appendix A), as it is close to the second (“I am highly effective at the things I do”) and twelfth (“I perform well at many things”) self-competence items on the SLSC-R (Tafarodi & Swann, 2001, p. 670).

While other measures have been used to collect scores on self-esteem, the RSES is by far the most popular. The Single-Item Self-Esteem Scale has been cited 2,523 since 2018 (Google Scholar, 2019). SLSC has been cited 618 times since 2010 (Google Scholar, 2019). Whereas the RSES has been cited 622 times in the past 5 years (Google Scholar, 2019), but used in 9,269 studies since 1967 (PsychINFO, 2019).

### **Why is Self-Esteem Important?**

The item-level data found for this study focused on the age group of 15- to 26-years of age. This is a very important time of development, as individuals transform from teenagers into young adults. The amount of value these individuals feel about themselves at this age could affect the trajectory of their lives in substantial ways.

Self-esteem is first shaped in childhood by social influences (Humphrey, 2004). Parent behavior towards a child is arguably the most important influencer of early self-esteem development (Humphrey, 2004). Teachers and peers are also very important in a child’s self-esteem development, as these interactions inform a child’s a sense of where they fit into the social order (Humphrey, 2004). Around the age of eight, children begin to combine their self-evaluations to later create their global self-esteem (Orth, 2019). Self-esteem begins to rise

around age 15 and continues to increase into adulthood. This increase is attributed to the fact that adults find themselves in social roles where they are expected to develop mature personality traits (Orth, 2019).

The benefits of high self-esteem are improved ingenuity and more agreeable moods (Baumeister et al., 2003). High self-esteem did not curb risky behavior in young adults, but did encourage experimentation (Baumeister et al., 2003). Global self-esteem was not strongly related to academic achievement (Humphrey, 2004). One of the earliest studies examining the relationship of self-esteem and academic achievement was a longitudinal study, started in 1966 by Bachman and O'Malley. It was a nation-wide study of 1,600 male tenth graders who were followed for eight years. The participants were asked to take a version of the RSES at several points in time, over eight years, 1966 to 1974. The researchers found self-esteem and academic performance had a correlation of .10 and a correlation between self-esteem and academic ability of .12 (Bachman, & O'Malley, 1977). Although Bachman and O'Malley did not find a correlation between self-esteem and academic performance or academic ability, others have found that some individuals with higher self-esteem may be more successful in school, because they set loftier goals for themselves and are more likely to persist when faced with failure (Baumeister et al., 2003). Low self-esteem has played a role in the development of depression in young adults (Brunet, Pila, Solomon-Krakus, Sabiston, & O'Loughlin, 2019). Young adults who have low self-esteem are less equipped to deal with negative stress, which can also lead to or deepen depression. Higher rates of body shame and guilt associated with poor body image have also been associated low self-esteem (Brunet, et al., 2019). Thus, self-esteem is an important construct to measure in research exploring the mental health, well-being, and persistence in education of young adults and adolescents.

**Why is it Important to Measure Self-Esteem Well?**

In the social sciences, we have the task of measuring unobservable characteristics, such as self-esteem or narcissism. These characteristics are often referred to as latent constructs or, depending on the method of analysis used, just constructs. First, these constructs need to be operationally defined by associated observable behaviors or self-report items. These behaviors are then reported and scored, and often a sum score across behaviors is calculated, called an observed score (Meyer, 2010). In the case of the RSES, the scores are self-reported based on the respondents' attitudes about themselves.

It is important to make sure that the scores reported to describe a certain construct are consistent (i.e., reliable) and representative. Science, particularly social science, is dependent on well-defined constructs and reliable measurements for clarity and study replication (Meyer, 2010). If the RSES scores are unreliable, then results using these scores may be inaccurate or inconsistent. Unreliable scores on the RSES can cause problems in relative decisions (i.e., rank ordering participants relative to one another by their RSES scores) about a participant's self-esteem. For example, if these scores were unreliable and were used in a correlation between self-esteem and motivation, this correlation may not accurately reflect the true relationship between self-esteem and motivation, because RSES scores represent measurement error too much and true levels of self-esteem too little. This misleading correlational study could limit future research or lead researchers in the wrong direction of research examining methods of increasing motivation. Unreliable scores could also lead to more serious problems. For example, imagine an effective intervention was developed to increase motivation in individuals with self-esteem below a certain cut-score (i.e., an absolute decision – comparing an individual's RSES score to an absolute cut score). Unreliable scores on the RSES would result in incorrect

decisions when comparing students' RSES scores to the cut-score, again because RSES scores contain too much measurement error and not enough true self-esteem. Consequentially, individuals who would most benefit from the intervention may not receive the intervention, as their RSES scores may be too high because of measurement error. One benefit of using G-theory in this thesis is that G-theory provides two reliability estimates: one for relative decisions and another for absolute decisions. Thus, results will indicate which types of research can be conducted with scores on the RSES in populations similar to those studied here.

In addition to relative versus absolute decisions in research, research can be characterized as low stakes or high stakes. An example of a low stake scenario using the RSES scale, which also happens to include relative decisions about scores, is a correlational study from 1993 looking at the correlation between self-esteem in high school students and global self-worth (Hagborg, 1993). An example of a high stakes scenario using the RSES, which also happens to include absolute decision about scores, is an intervention that used the RSES to assess which participants were eligible for a self-esteem intervention for positive symptomatology of mentally disordered offenders (Laithwaite, Gumley, Benn, Scott, Downey, Black, & McEwen, 2007). The participants in this 2007 pilot study had to have been perilously diagnosed with a primary diagnosis of schizophrenia, schizo-affective disorder or bi-polar disorder (Laithwaite, et al., 2007). In addition, these participants needed to earn a high score on the RSES, indicating low self-esteem, in order to be eligible for the intervention (Laithwaite, et al., 2007).

The two CTT reliability coefficients used in this thesis, Cronbach's  $\alpha$  and Guttman's  $\lambda_2$ , have various cut-offs for low stakes research. While self-esteem research is important, much of the self-esteem research in the literature is considered low stakes, because results do not have any major consequences on participants. Self-esteem research is not often in a life-or-death

context and does not often result in students being held back a grade in school, both of which are examples of high stakes situations. Researchers typically use .70 as a cut-off for Cronbach's  $\alpha$  values for scores to be considered reliable in low stakes situations (Taber, 2018). Researchers typically use a .80 as a cut-off for Guttman's  $\lambda_2$  values for scores to be considered reliable in low stakes situations (Guttman's Lambda-2, 2019). If Cronbach's  $\alpha$  and Guttman's  $\lambda_2$  values meet or exceed their designated cut-offs, scores are consistent enough to use in low stakes research.

### **Why the Reliability of RSES Scores May Have Changed**

In the 1980's the state of California created a task force to increase the self-esteem of its residents (Baumeister et al., 2003). The logic was that if Californians had higher-self-esteem, they would produce more revenue and in turn reduce many of the state's social problems, saving the tax payers money. Some of the expensive problems that the government of California thought they could solve with increasing the self-esteem of California citizens included unwanted pregnancy, school failure, crime, and drug abuse (Baumeister et al., 2003). This started what is now known as the self-esteem movement in America. The self-esteem movement became part of the education system in the last two decades of the 20th century. Millions of dollars were spent to develop programs to externally boost the self-esteem of America's children in hopes of improving academic achievement (Humphrey, 2004).

This push to bolster self-esteem externally may have changed the nature of how people view their own self-esteem. Baumeister, Campbell, Krueger, and Vohs (2003) believed haphazard acclaim could promote narcissism instead of self-esteem and recommended only using praise as a reward when earned. This potential shift in the interpretation of self-esteem may have contributed to a decrease in reliability of RSES scores. Self-esteem is often discussed with narcissism and self-efficacy, and the RSES may partially be measuring these attitudes. For



example, item four on the RSES is “I am able to do things as well as most people” (Rosenberg, 1975 p. 291; Appendix A). This item may not be interpreted in the same way now as when it was written in 1965. The RSES was initially written for high school students, and thus, these students would have been comparing themselves to their fellow students, family members and popular personalities in the media (i.e., on television and radio, in movies). When a young adult reads this question today, who are they comparing themselves to? Media has changed so much since 1965. This change has exposed young adults to unrealistic beauty standards. For example, the size of women presented in media has decreased steadily since the 1960’s (Park, 2005). Socialization has changed since the advent of the internet and subsequently social media; young adults now have to compare themselves to the unattainable portrayals of other people’s lives on-line. What “things” do they think they are doing as well as others at? Someone who rated themselves as *Strongly Agree* on this item may have a positive exaggerated sense of self, or some narcissistic tendencies (Twenge et. al., 2008). On the other hand, this person could also be exhibiting high self-efficacy if they feel they usually perform well at most tasks and expect to perform as well as others (Bandura, 2010). In our current society, young adults are constantly comparing themselves to highly curated representations of success that they see on-line, this may lead them to choose *Disagree* or *Strongly Disagree* no matter how competent they are. This comparison to social media influencers can also lead to envy, as more exposure to unattainable standards causes more social comparison and in turn leads to increased dissatisfaction and envy of those who have an unfair advantage (Chae, 2018). If this item is measuring narcissism, self-efficacy, or envy, that could mean this item no longer correlates with scores on the other RSES items. To be reliable, item responses on this item would need to be consistent with responses on

other items. Inconsistency would lead to decreased reliability of scores on the RSES. In either case, this measure may no longer be interpreted by young adult respondents as initially intended.

The process of researching the reliability of the RSES scores has led me to question the validity of RSES scores as well. In future research, I would like to conduct a qualitative study to investigate if items on the RSES represent current ideas about the phenomena of self-esteem (Creswell & Miller, 2000). Is it still correct to infer attitudes of self-esteem from scores on a measure written in 1965, or have interpretations of self-esteem changed? If interpretations of this attitude have changed, a possible source of this change is the self-esteem movement. If RSES scores are no longer measuring the attitude of self-esteem, what are they measuring? In order for scores on a measure to be valid, they must be reliable first (Meyer, 2010, p. 6). Thus, an extensive investigation of the reliability of RSES scores across decades is critical, before these questions can be answered.

### **History of the RSES**

The RSES was originally created in 1965 as a tool to measure self-esteem in teenagers (Rosenberg, 1979). The RSES is now used widely across social sciences to measure self-esteem in a variety of individuals from many countries. This 10-item measure was developed with a four-point Likert scale ranging from 1 = *Strongly Agree* to 4 = *Strongly Disagree*. In 1965, Rosenberg demonstrated scores on the RSES, when administered to high school aged students, had high reliability, with internal consistency (Cronbach's  $\alpha$ ) of .77 and Guttman scale Coefficient of Reproducibility of .92. Silber and Tippet (1965) also found scores on the RSES were reliable with a test-retest reliability over a two-week interval of .85 when administered to 37 college students from four different colleges.

In 1965, Rosenberg initially tested reliability of the scores from the RSES with the Guttman scale coefficient of reproducibility. While we now consider the RSES to be a Likert-type scale, at the time of its development, Rosenberg treated it as a Guttman Scale, an ordinal scale in which the items can be in ranked order such that respondents always agree to an easier item before agreeing with a more difficult item (Clayton, 2019). Initially, some of the 10 items were combined into multi-part items, producing a 6-item Guttman scale. Items 1, 8, and 10 (see Appendix A) were scored as single items. Items 3, 7, and 9 were grouped together, and at least two out of three answers needed to be on the agree side of the response scale in order for these three combined items to be considered a positive answer. Similarly, if either item 4 or 5 were answered positively, the combined scores were considered a positive item. Items 2 and 6 were also combined as one item, such that if one of them was answered on the agree side of the scale, the entire response for the two-part item was considered positive (Rosenberg, 1979). In the early 1970's researchers began to regularly use the RSES as a 10-item Likert-type scale, ignoring the potential Guttman scale structure (Dobson, Goudy, Keith, & Powers, 1979)

### **Study Purpose**

I became familiar with the RSES in my first semester of the Educational Psychology Department's Master of Arts program at the University of Arizona. I was enrolled in a measurements course, a requirement of which was to complete a final project. I had no data of my own and thus analyzed data collected by Dr. Erbacher (University of Arizona). Dr. Erbacher coordinated a multi-section survey to collect data about a variety of student attitudes. The survey was completed online by 229 undergraduates in the fall of 2016. In my final project, I used Generalizability Theory (G-theory) to determine whether the 10-item RSES, part of the online survey, could be shortened and/or if additional items were needed for scores to have adequate

reliability. The reliability coefficients in the generalizability-study (G-study) and decision-study (D-study) were calculated. G-study coefficient represent how generalizable an individual's observed score, the mean of a set of scores (i.e., their item responses), is to that individual's universe score (i.e., their score in the larger universe of all possible items). Generalizability coefficients are used to evaluate the reliability of relative decisions, for example ordering or comparing individuals relative to one another by their scores. Dependability Coefficients are used to evaluate absolute decisions, for example finding the position of objects in the entire universe or comparing individuals' scores to a cut-score. In a G-study, we are considering the measure as if it has an item set of 1. D-study coefficients use the information provided by the G-study to inform the optimal number of items that should be included in the measurement procedure (Shavelson & Webb, 1991). With multiple D-studies, I was able to create different scenarios to explore what would happen to score reliability if various numbers of items were administered. The results prompted me to explore the reliability of RSES scores at a larger scale, across several decades rather than in a single data set.

While the social movement of boosting self-esteem in communities and classrooms was very well intentioned, it may have had unintended results. Indiscriminate praise present during the self-esteem movement could have encouraged narcissism instead of self-esteem (Humphrey, 2004), and could change how individuals evaluate and view their own self-esteem.

The purpose of this study is to explore the reliability of scores on the RSES over 48 years to see if the reliability of these scores has changed. I expect to see the reliability of the RSES scores decrease over time, starting around the implementation of the self-esteem movement (i.e., 1990s). If the reliability of RSES scores has decreased over time, then this will bring into question the continued use of the RSES. On the other hand, if the reliability of RSES scores has

remained stable or increased over time, then the validity of RSES scores must be explored in future work. High reliability is necessary for quality measurement, but high reliability of scores does not guarantee validity of scores (Meyer, 2010, p. 6). If the construct being measured has changed, that also needs to be known. However, first, we must determine whether RSES scores from today's young adults are reliable.

### **Method**

Past research using the RSES has focused on comparing cohorts longitudinally to measure self-esteem over time, but has neglected to compare similar age groups at different points in time to confirm the continued reliability of the scores on the scale with new cohorts. I hypothesize that since its inception in 1965, the scores from the RSES have become less reliable. I predict that this change happened immediately after the beginning of the self-esteem movement of the 1980s and 90s.

### **Participants and Measures**

I collected 13 existing item-level datasets (see Table 1) using the RSES between 1971 and 2019. The samples are of varying sizes, containing young adults between the ages of 15 and 26.

The earliest data set is from the "Longitudinal Study of Generations, 1971, 1985, 1988, 1991, 1994, 1997, 2000, 2005" (Bengtson, 2008). This intergenerational study included 300 multi-generational families in California, including grandparents, parents, grandchildren and great-grandchildren (starting in 1991). Unfortunately, this study only administered the RSES to the age group of interest, 15-26, in 1971. This study also omitted two items from the RSES, items 3 and 6 (Bengtson, 2008). Since the RSES was initially considered a Guttman scale, Rosenberg considered items 2 and 6 as a single item and 2 out of 3 correct answers to items 3, 7,

and 9 were scored as a single item (Rosenberg, 1979). Therefore, I decided to keep the data from this 8-item version of the RSES. This set of item Responses from 1971 included 583 participants from ages 15 to 26. Items were re-arranged to reflect the initial order intended by Rosenberg (Rosenberg, 1975).

The second earliest datasets came from the "National Longitudinal Survey of Youth" (NLSY, 1979). This longitudinal study included data from seven cohorts between 1994 and 2006, with participants aged 15 to 26 in each dataset. This project followed the lives of a sample of Americans starting in 1979. The first year the survey included the RSES was 1994. The items in this data set also had to be rearranged to keep the data consistent across all years. The sets of item-level data from this source began in 1994 and continued every other year until 2006, with ages ranging from 17 to 26 years of age in 1994, 20 to 26 years of age in 1998, and 15 to 26 years of age in each of 2000, 2002, 2004 and 2006. Samples sizes ranged from 964 in 2006 to 4524 in 2004 (NLSY, 1979).

The third source of data came from the Lambda4 package (Hunt, 2013) in the free software environment for statistical computing, R (R Core Team, 2019). This data set, named "Rosenberg", is from 2010 and contains RSES item responses from 837 high school and college aged students.

The fourth source of data is the Attitudes and Behavior in Learning and Education (ABLE) lab at the University of Arizona, led by Monica Erbacher, Ph.D. The ABLE lab included the RSES in four surveys administered to students at a large, public university in the southwestern US in spring 2017, fall 2017, fall 2018, and spring 2019. These students were part of an educational research participant pool. The sample sizes in the four datasets ranged from 79 (spring 2017) to 226 (fall 2017). Data sets from the same school year but different semesters

were considered different cohorts, particularly because attitudes tended to differ between fall and spring freshmen (M. Erbacher, personal communication, November 28, 2019)

Table 1

*Data Sources*

<b>Index</b>	<b>Study</b>	<b>Source</b>	<b>Year</b>	<b>Age Range</b>	<b>N</b>
1	Bengston*	Interuniversity Consortium for Political and Social Research.	1971	15-26	583
2	National Longitudinal Survey of Youth	NLSY 79 survey	1994	17-26	970
3	NSLY	NLSY 79 survey	1996	15-26	1656
4	NSLY	NLSY 79 survey	1998	20-26	2127
5	NSLY	NLSY 79 survey	2000	15-26	1636
6	NSLY	NLSY 79 survey	2002	15-26	1411
7	NSLY	NLSY 79 survey	2004	15-26	4524
8	NSLY	NLSY 79 survey	2006	15-26	964
9	Rosenberg	R, Lambda4	2010	High school & College	837
10	Erbacher	ABLE Lab, University of Arizona	2017 (Spring)	College	79
11	Erbacher	ABLE Lab, University of Arizona	2017 (Fall)	College	226
12	Erbacher	ABLE Lab, University of Arizona	2018 (Fall)	College	140
13	Erbacher	ABLE Lab, University of Arizona	2019 (Spring)	College	206

*Note.* \*= 8 item scale

**Data Analysis**

I have treated all the data collected as Likert-type items. To analyze various forms of reliability for scores on these items in each of the data sets described above, I used two internal consistency coefficients from CTT and two coefficients, one relative and one absolute, from G-

theory. Each of these measurement frameworks, along with the coefficients used, are explained briefly below.

**Classical Test Theory (CTT).** CTT separates observed scores into true scores and error scores with the equation  $X = T + E$  (Meyer, 2010, p. 14). Reliability is the ratio of true score variance to observed score variance (Meyer, 2010, p. 20). In other words, reliability is the proportion of variance in observed scores on a measure that is due to true differences between individuals on the construct of interest. Only one source of error can be modeled in CTT. To estimate internal consistency reliability, I used Cronbach's  $\alpha$  (see equation 1) and Guttman's  $\lambda_2$  (see equation 2).

$$\alpha = \frac{k}{k-1} \left( \frac{\sigma_x^2 - \sum_j^k \sigma_{y_j}^2}{\sigma_x^2} \right) \quad (1)$$

$$\lambda_2 = \left( \frac{\sqrt{\frac{k}{1 < -1}} (\sigma_x^2 - \sum_j^k \sigma_{y_j}^2)}{\sigma_x^2} \right) + \left( \frac{\sigma_x^2 - \sum_j^k \sigma_{y_j}^2}{\sigma_x^2} \right) \quad (2)$$

In the Cronbach's  $\alpha$  equation,  $k$  is the number of items used in a weight at the beginning of the equations,  $\sigma_x^2$  is the total variance (i.e., variance of sum scores) across items and  $\sum_j^k$  sums the item variances ( $\sigma_{y_j}^2$ ). In other words, in the numerator in equation 1, item variances are subtracted from sum score variance, which leaves only item covariances or overlap among items in the numerator. The Guttman's  $\lambda_2$  equation is almost identical to the Cronbach's  $\alpha$  equation, the only difference is the weight at the beginning. In equations above, the variance of total scores contains variance of scores on each item as well as covariance between scores on pairs of items. On the left-hand side of both equations is the weight. On the right hand side, the variance of scores for each item is separately being subtracted from the variance in the sum scores of the entire measure. Thus, all that remains is the overlap covariance among item responses. This



overlap represents the part of the scores that is consistent from item to the next. In other words, this overlap should represent self-esteem. Then, this aggregate of covariances among item scores is multiplied by a weight.

Cronbach's  $\alpha$  and Guttman's  $\lambda_2$  determine the internal consistency reliability of scores on a measure, to answer the question: Will the items elicit the same response if administered again to the same sample, if it were possible to wipe participants' memories (Santos, 1999; Callender & Osburn, 1979)? Both of these reliability estimates are meant for essentially tau-equivalent measures (Meyer, 2010). To use  $\alpha$ , three assumptions must be met; true scores must be equal across items, true score variance must be equal across items, and covariances between scores on pairs of items must be equal (Meyer, 2010). When the last assumption about item covariances is unreasonable, Guttman's  $\lambda_2$  is a more accurate estimate of internal consistency, as  $\alpha$  will be artificially low (Meyer, 2010).

**Generalizability Theory (G-theory).** G-theory informs the reliability of scores in terms of how generalizable relative and absolute decisions using those scores are to a larger domain or to the Universe of Admissible Observations. The Universe of Admissible Observations is made up of the observations similar to those in the data collected (e.g. similar respondents, similar items; Shavelson & Webb, 1991, pg. 3). This G-theory study includes the common four components of G-theory; a G-study, one or more D-studies, the Generalizability Coefficient ( $\rho^2$ ; see equation 3) and the Dependability Coefficient ( $\Phi$ ; see equation 4) calculated for each G and D-study. When calculating  $\rho^2$  it has two components,  $\sigma_p^2$  is variance attributed to respondents or persons and  $\sigma^2_{relative.error}$  is made up of error variance, which is confounded with variance attributed to the interaction between persons and items (see equation 3). Note, item variance is not included in relative error variance. All variance components that influence the rankings of an

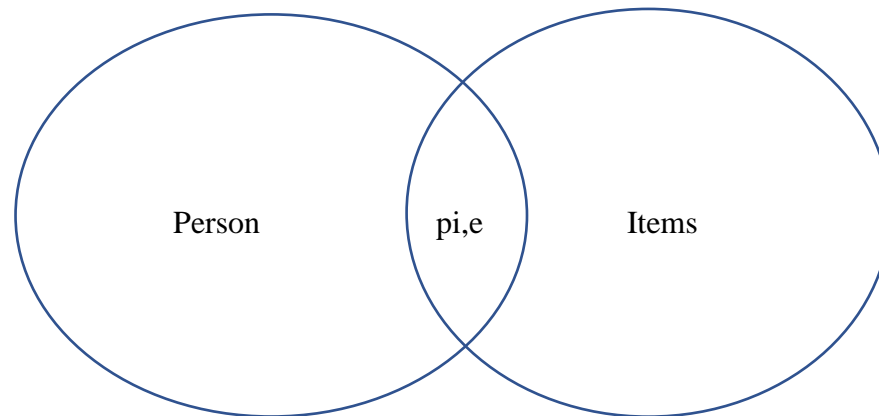
individual contribute to relative error, these components interact with the object of measure (Shavelson & Webb, 1991, pg. 84). The equation for  $\Phi$  is almost identical, with the exception that absolute error is in the denominator rather than relative error (see equation 4). Absolute error is all variance components except the object of measurement, in this instance persons (Shavelson & Webb, 1991, pg. 84). Thus, item variance and error variance both contribute to absolute error variance.

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{relative.error}^2} \quad (3)$$

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{absolute.error}^2} \quad (4)$$

$$\sigma_x^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{p,i,e}^2 \quad (5)$$

This is a crossed, random, 1-facet design. The objects of measurement, the cases we are trying to rank order or make absolute decisions about, are the persons ( $p$ ), and the facet is the items ( $i$ ) on the self-esteem scale. This study is crossed because every student answered every item on the scale. The items on the RSES are considered random because all items are interchangeable and because we want to generalize these scores to a larger universe of self-esteem items, beyond the 10 RSES items (Shavelson & Webb, 1991). The strength of G-theory is that it lets us break the variance of scores on a measure into multiple sources. In this 1-facet design, total score variance ( $\sigma_x^2$ ) is broken down into variance between persons ( $\sigma_p^2$ ), variance explained by items ( $\sigma_i^2$ ), and remaining error variance which also contains variance explained by specific person-item combinations ( $\sigma_{p,i,e}^2$ ), see equation 5 and Figure 1).



*Figure 1.* Visualization of the G-study Design

G-studies break variance in test scores down into multiple components, based on the sources of variance included in the G-study design (here, a 1-facet, fully crossed, random design). In this design, I found how much of the variance can be attributed to Persons, Items, or the Person and Item interaction, which is confounded with error (Shavelson and Webb, 1991). To perform the G-study, I used the *gtheory* package (Moore, 2016) in the statistical software R (R Core Team, 2019). Half the items were reverse scored, items 2, 5, 6, 8 and 9, in each of the obtained data sets. Each dataset was formatted in wide format for the CTT reliability analyses. Wide format was structured with 10 columns (one for each item) and a row for each person. For G- and D-study analyses, data was transformed into long format with 3 columns (person ID, item ID, and scores) and 10 rows for each person, one row for each item within each person. Total observed variance ( $\sigma_x^2$ ) was broken down into three components: item variance ( $\sigma_i^2$ ), person variance ( $\sigma_p^2$ ), and variance for the item-by-person interaction and error ( $\sigma_{pi+e}^2$ ; Shavelson and Webb, 1991). Importantly, the variance component for the item facet (or all facets other than persons and error), represents variance in observed scores explained by differences between

single items (i.e., as though single items were administered to participants rather than 10- or 8-item sets).

The second step of G-theory is conducting one or more D-studies. D-studies allow you to test theoretical measurement scenarios. I conducted two D-studies for each data set: one in which 8 items were hypothetically administered and the other in which 10 items were hypothetically administered, to mimic the two actual measurement scenarios found in the obtained datasets.

I calculated new variance amounts and proportions for the 8-item 10-item scenarios. To find the new theoretical variance estimates, I used the G-study variance estimates and the theoretical number of items. The persons variance does not change from G-studies to D-studies.

The CTT coefficients, Cronbach's  $\alpha$  and Guttman's  $\lambda_2$ , estimated the reliability of scores on the RSES measure in each dataset, assuming one source of measurement error. The strength of G-theory coefficients is that the reliability coefficients (generalizability and dependability coefficients) allow for multiple sources of variance.

## Results

### Classical Test Theory Coefficients

Cronbach's  $\alpha$  was explored, along with the upper and lower bounds of a 95% confidence interval around each alpha value. The low Cronbach's  $\alpha$  bound had a range of .79 to .88, and the high Cronbach's  $\alpha$  bound had a range of .83 to .92. The median Cronbach's  $\alpha$  coefficient score across all years was .87, with a range of .80-.90 (Figure 2). The median Guttman's  $\lambda_2$  coefficient score across all years was .87 with a range of .81 to .91 (Figure 3). Both coefficients increased over time at a similar rate. The cut-offs for Cronbach's  $\alpha$  and Guttman's  $\lambda_2$  are  $\geq .70$  for group-level studies and  $\geq .80$  for low-stakes evaluations, for example grades. These alpha scores are

above the cutoffs for group level studies and low-stakes evaluations (Stephanie, 2018; Taber, 2018).

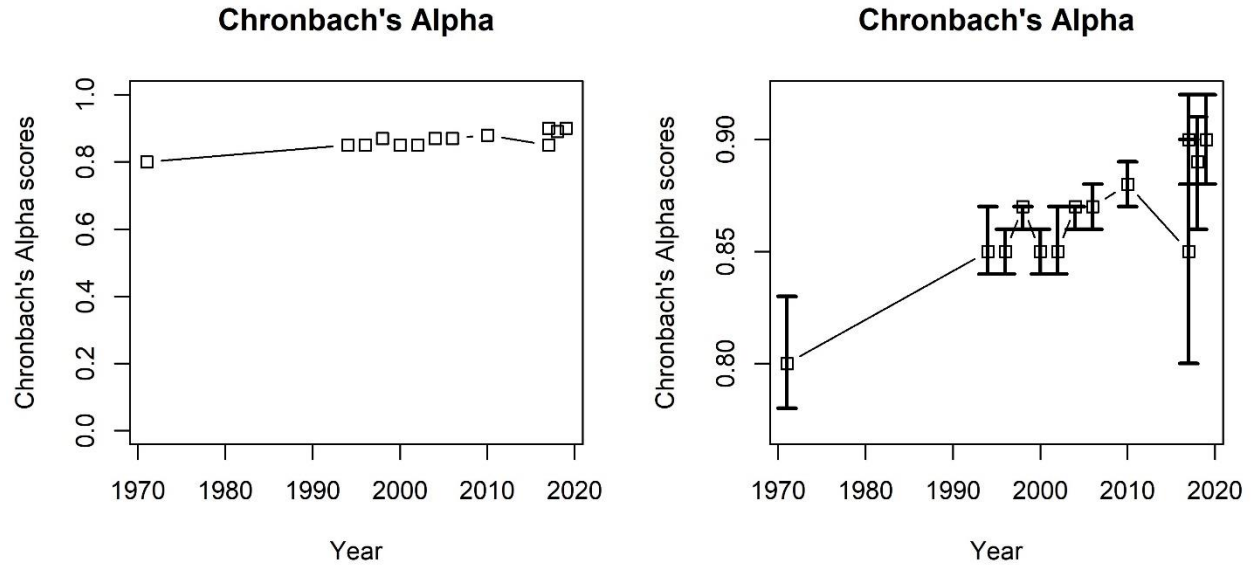


Figure 2. Cronbach’s  $\alpha$  scores from 1971 to 2019. The left plot displays results within the possible range of  $\alpha$ . The right plot displays results within the observed range of values.

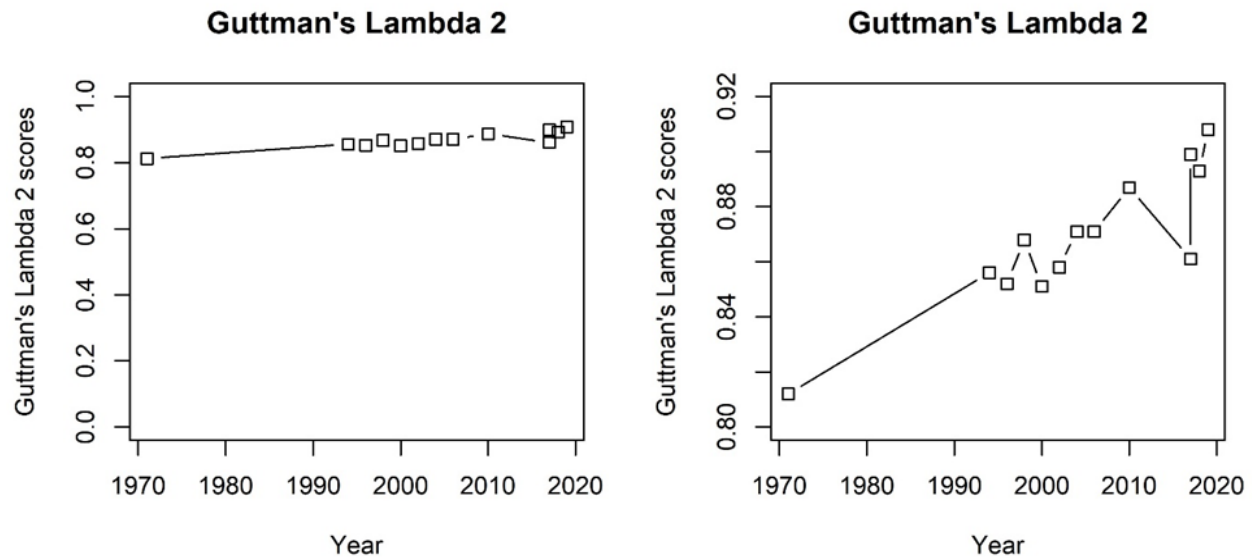


Figure 3. Guttman’s  $\lambda_2$  from 1971 to 2019. The left plot displays results within the possible range of  $\lambda_2$ . The right plot displays results within the observed range of values.

### Generalizability Theory Coefficients

Within the G-theory analyses, I first explored the variance components from the G-study conducted for each dataset as percentages. The median percent of variance explained by persons across all years was 35.53% with a range of 27.40% to 42.00%. We want person variance to be the highest source of variance. The data with the highest source of person variance came from the ABLE Lab. The median percent of item variance across all years was 10.34%, with a range of 5.10% to 19.50%. In other words, on average, items accounted for 10.34% of the variance in observed item responses (out of items, persons, and the confounded item-person interaction and error). The lowest percentage of item variance came from the NLSY data sets in the 90's and 2000. The median percent of residual variance across all years, was 54.21% with a range of 44.50% to 60.70%. The highest residual variance came from the NLSY data and the lowest came from the ABLE lab data (Figure 4).

Generalizability ( $\rho^2$ ) and dependability ( $\Phi$ ) coefficients were also calculated in each G-study. Recall that G-study results apply to 1-item sets. The median generalizability coefficient across all years was .40, with a range of .34 to .49. The lowest  $\rho^2$  values came from the earliest data set (Bergtson) and the highest came from the most recent data, (ABLE Lab), suggesting a slight upward trend. The median  $\Phi$  coefficient was .36 with a range of .28 to .42 (Figure 5). The pattern of the  $\Phi$  coefficient values were similar to those of  $\rho^2$  coefficient values. The lowest  $\Phi$  also came from the earliest data set (Bergtson) and the highest came from two of the most recent data sets (ABLE Lab), mirroring the upward trend of the  $\rho^2$  values. If this measure was administered as a 1-item measure, it would result in unreliable scores. These coefficients represent the proportion of variance explained by persons and they are too low.

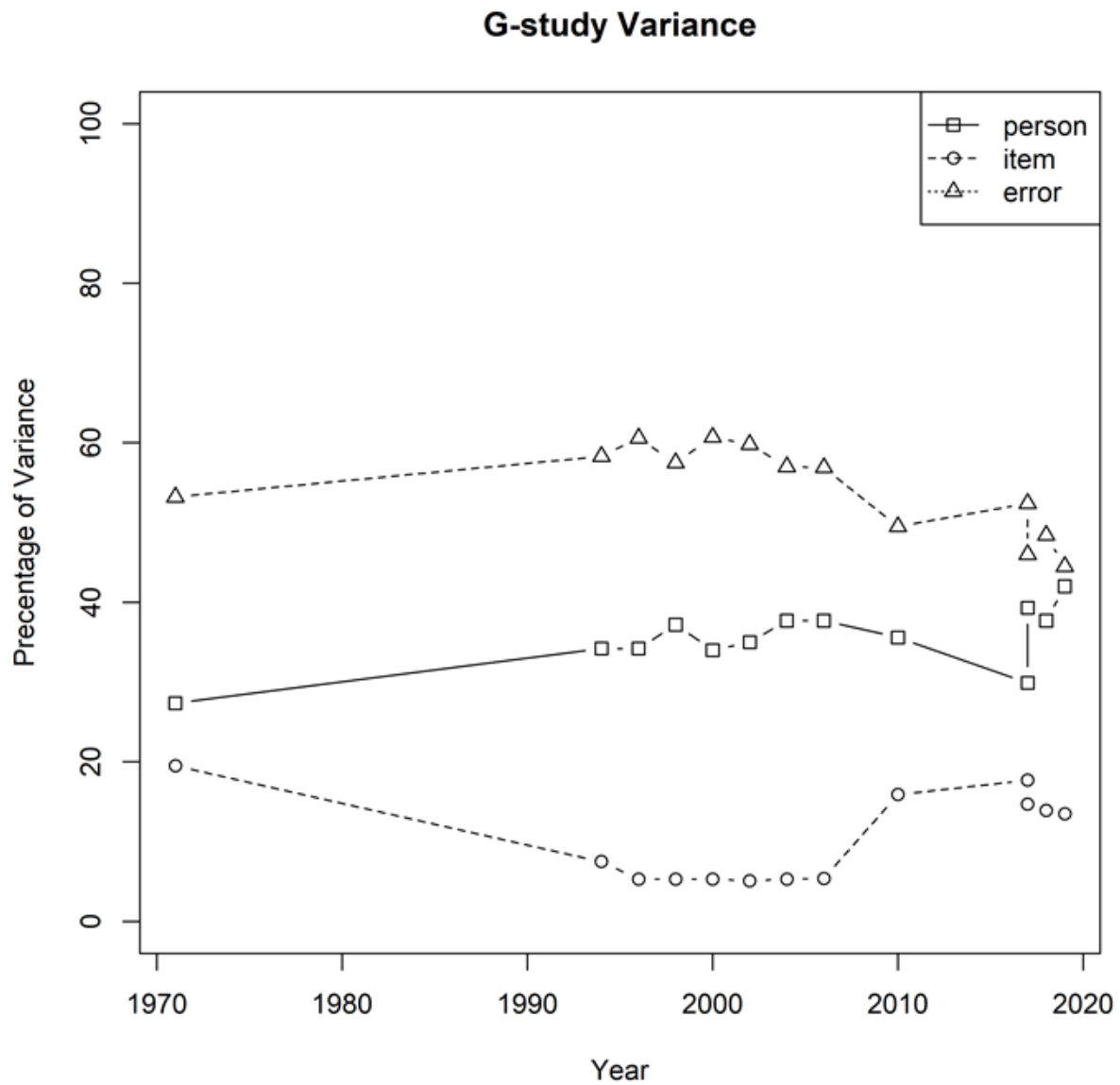


Figure 4. Percentage of G-study scores variance from 1971 to 2019

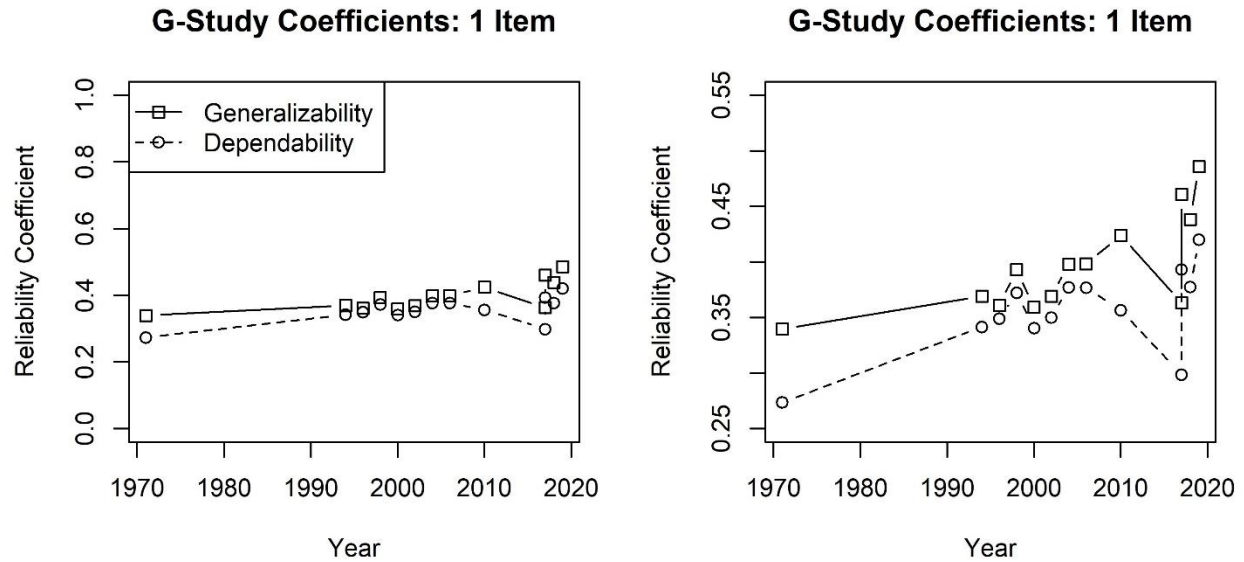


Figure 5. G-study Coefficients: 1 item. The left plot displays results within the possible range of the reliability coefficients. The right plot displays results within the observed range.

The second set of analyses were 8-item D-studies. These D-studies estimated percentages of variance as though eight RSES items had been administered in each study. The median percent of person variance across all years, was 81.60% with a range of 75.06% to 85.31%. The median percent of item variance across all years was 2.23%, with a range of 1.46% to 6.70%. The lowest item variance percentage again came from the NLSY data sets in the 90's and 2000. The median percent of residual variance across all years was 15.94% with a range of 6.38% to 18.25%. The lowest residual variance came from the NLSY data in 2004 (Figure 6).

Generalizability and dependability coefficients were calculated for each D-study as well. D-study variance analysis revealed a median  $\rho^2$  value of .84 for the D-study, with a range of .74 to .88. The lowest Generalizability Coefficient came from the 2017 spring ABLE lab data set, and the highest came from the latest ABLE lab data set, Spring 2019. The median  $\Phi$  scores for the D-study variance was .82 with a range of .71 to .85. The pattern of the Dependability coefficient scores were similar to the Generalizability coefficient scores (Figure 7).



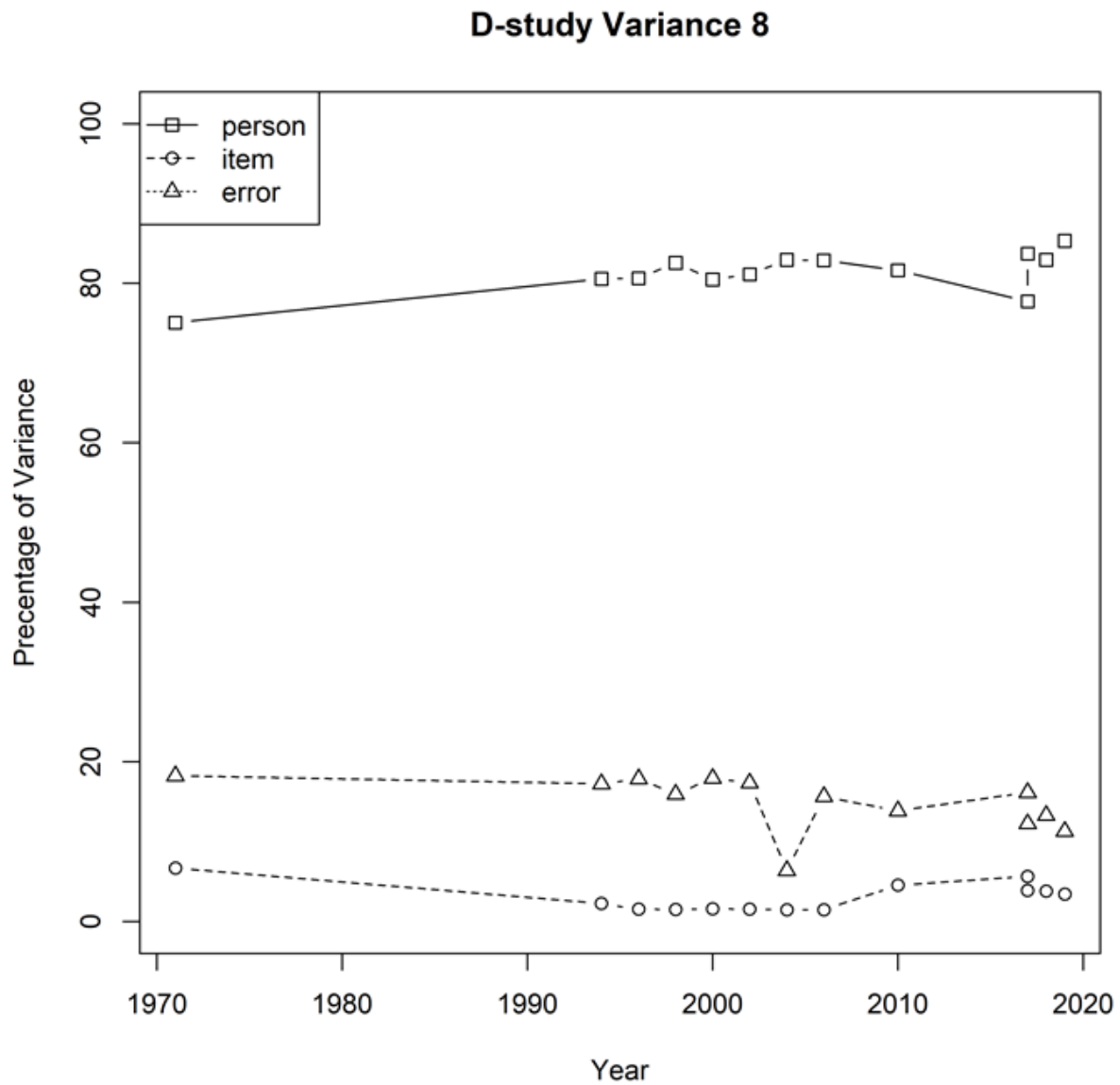


Figure 6. Variance Percentages Across 8-item D-studies from 1971 to 2019

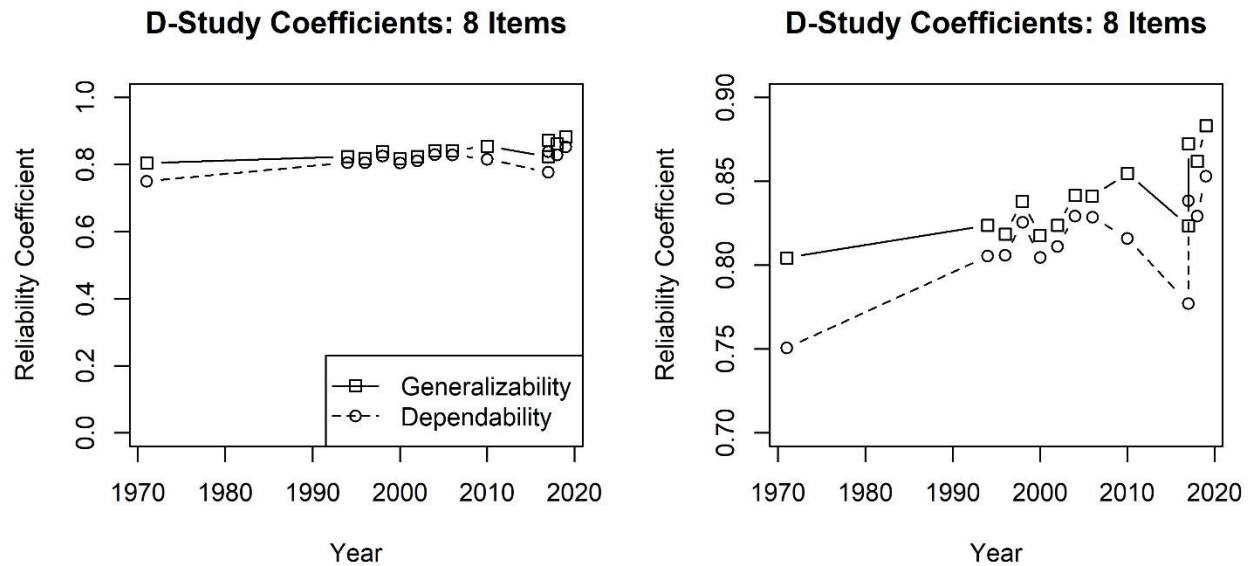


Figure 7.  $\rho^2$  and  $\Phi$  scores from the 8 - item D-study. The left plot displays results within the possible range of the reliability coefficients. The right plot displays results within the observed range.

The third analysis conducted for each data set was a 10-item D-study. The median percent of person variance across all years was 84.71% with a range of 79.01% to 87.87%. The median percent of item variance across all years was 1.83% with a range of 1.22% to 5.63%. The lowest item variance percentage again came from the NLSY data sets in the 90's and 2000. The median percent of residual variance across all years, was 13.23% with a range of 9.26% to 15.37%. The lowest residual variance came from the last three ABLE Lab data sets (Figure 8). The next step was finding the Generalizability and Dependability coefficients. D-study variance analysis revealed a median  $\rho^2$  value of .87 for the D-study, with a range of .84 to .90. The median  $\Phi$  scores for the D-study was .85 with a range of .79 to .88. (Figure 9.) The patterns of  $\Phi$  and  $\rho^2$  squared were like those found in the 8-item test, indicating a slight increase over time.

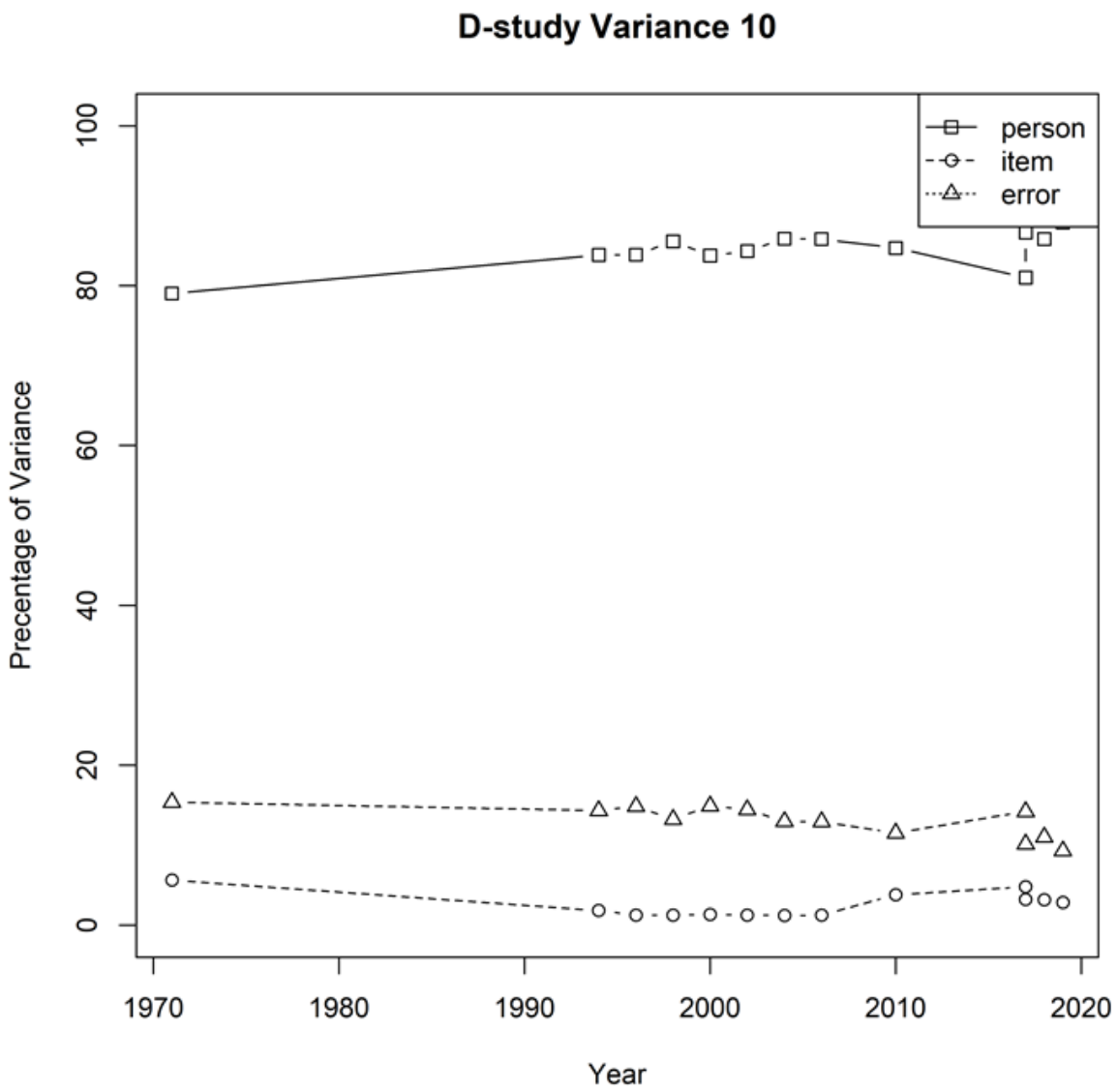


Figure 8. Variance Percentages Across 10-item D-studies from 1971 to 2019

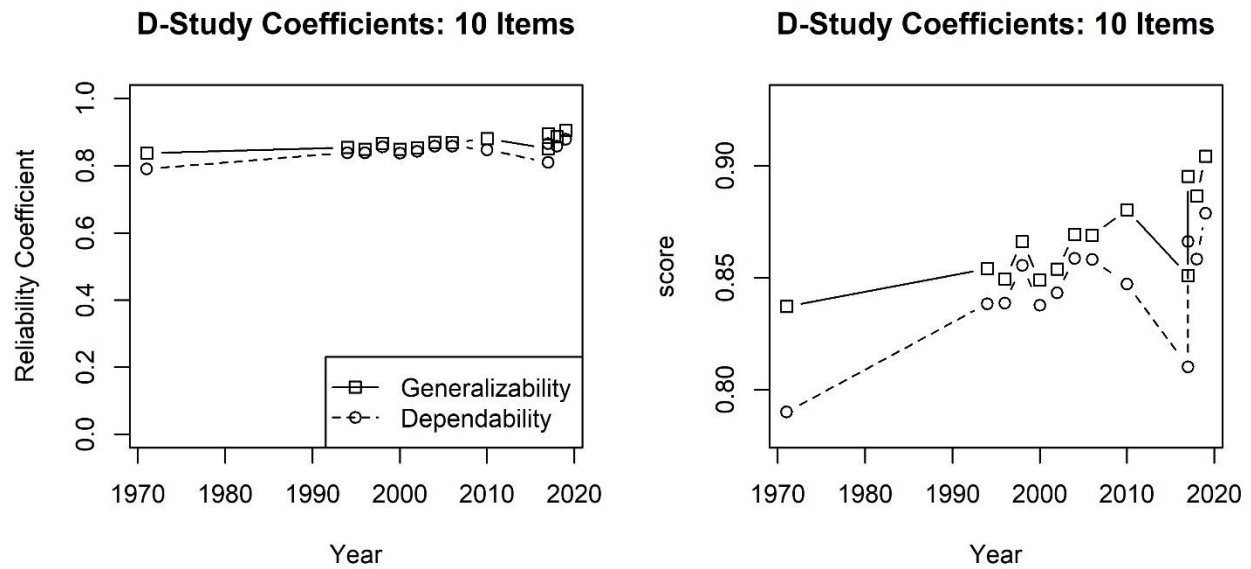


Figure 9.  $\rho^2$  and  $\Phi$  scores from the 10 - item D-study. The left plot displays results within the possible range of the reliability coefficients. The right plot displays results within the observed range.

### Discussion

If the scores on the RSES are no longer generalizable nor reliable, it is important to know when they stopped being reliable. Comparing data from different points in time will be a very effective way to evaluate whether the scores of the RSES are still generalizable and reliable, and if they are not, to pinpoint when the shift happened.

The CTT coefficients scores of Cronbach’s  $\alpha$  and Guttmann’s  $\lambda_2$  were above the low stakes cut off scores. Cronbach’s  $\alpha$  coefficient scores on the RSES ranged from .80 to .90, which are all above the low stakes  $\alpha$  cutoff of .70 often used in social science research (Pmean.com, 2019). The Guttman’s  $\lambda_2$  coefficient scores on the RSES ranged from .81 to .91., all of which exceeded the low stakes cut off of .80 (Stephanie, 2018). These values suggest scores on the

RSES have acceptable reliability for low-stakes research contexts in the populations explored here (Stephanie, 2018).

Through a G-study, I was able to simulate a one-item RSES. The  $\rho^2$  coefficient scores ranged from .34 to .49. The  $\Phi$  coefficient score ranged from .27 to .42. These coefficients are too low to rank order people or compare to a cut score. While these coefficients have increased over time, the scores on a one-item RSES cannot be considered reliable in the populations studied here. The reliability of scores on the Single-Item Self-Esteem Scale (SISE) were not considered because it is a one-item scale (Robins, Hendin & Trzesniewski, 2001, pg. 152). The generalizability and dependability coefficients from the G-study results suggest that other one-item self-esteem measures may also yield unreliable scores.

In the 8-item D-study, the reliability coefficients were higher than the G-study reliability coefficients. The  $\rho^2$  coefficient values were all greater than .80, which is high enough for a relative decision low-stake research (Boyle, Saklofske & Matthews, 2014). However, the  $\rho^2$  coefficients values are not high enough to make relative decision evaluations in high stakes research since none of the coefficients are greater than .90 (Boyle, Saklofske & Matthews, 2014). The  $\Phi$  coefficients scores on the RSES were all greater than the cut-off of .80, except the coefficient value from the oldest set of scores which was .75, which is high enough to make absolute decisions in low stakes research (Boyle, Saklofske & Matthews, 2014). However the  $\Phi$  coefficient values are not high enough to make absolute decisions in high stakes research because the coefficient scores do not meet the cut off of .90 (Boyle, Saklofske & Matthews, 2014). Thus, with these RSES scores, it is acceptable to rank order people to make relative decisions in low stakes situations and it is acceptable to use RSES scores in low stakes absolute decisions but unacceptable to use these scores for absolute decisions (i.e., compare RSES scores

to a cut-score). For example, a researcher could run analyses that rely on rank ordering individuals by their scores, such as a correlation between self-esteem and another variable, or group differences on self-esteem. The reliability coefficients were not high enough to make absolute decisions. For example, it would not be appropriate for researchers to compare these RSES scores to a cutoff score in order to identify participants with low self-esteem as selection criteria for an intervention study.

The 10-item D-study reliability coefficient scores were high. The  $\rho^2$  coefficient values for the RSES responses were all greater than .83, and the  $\phi$  coefficient values were all larger than .79. These coefficients are a bit higher than the 8-item reliability coefficient scores. However, the same pattern emerged:  $\rho^2$  and  $\Phi$  values were above the .80 cut off, except for one  $\phi$  value which was just below. These coefficients values are higher than the 8-item D-study coefficients, but still might not be high enough to make absolute decisions because the coefficient does not meet the .90 cut (Boyle, Saklofske & Matthews, 2014).

### **Implications**

Given the results, it is acceptable for researchers to use 8 or 10 RSES items to measure self-esteem for research in which participants are being rank ordered by their scores. It is not appropriate to use 8 or 10 RSES items to measure self-esteem research in which participants' scores are being compared to an absolute cutoff.

It is appropriate to use scores on the RSES in research is when the scores are used to make relative decisions in low stakes research. For example in 2004, Grzegorek, Slaney, Franze and Rice study use scores on the RSES to look at the relationship between self-esteem and perfectionism. This study had 273 participants, all were undergraduate students who were enrolled in Educational Psychology classes at Pennsylvania State. To measure perfectionism the

Almost Perfect Scale-Revised (APS-R) was used (Slaney, Mobley, Trippi, Ashby & Johnson, 1996). The APS-R is a 43-item measure with six subscales that measure perfectionism (Ashby & Kottman, 1996). These subscales include, personal standards, need for order, the discrepancy between performance and standards, interpersonal relationships, anxiety around performance, and procrastination (Ashby & Kottman, 1996). A cluster analysis was used to distinguish three groups, adaptive perfectionists (AP), maladaptive perfectionists (MP) and nonperfectionists (NP; Grzegorek, Slaney, Franze, & Rice, 2004). The subjects were given the instruments, RSES and APS-R, at two points in the semester and a demographic survey which included GPA (Grzegorek, et. al., 2004). Those identified as AP had significantly higher scores on the RSES than those identified as MP or NP. The NP and AP groups were not significantly different (Grzegorek, et. al., 2004). This is a low stakes study to look for relationships between attitudes, this is an appropriate use of the scores on the RSES. Relative decisions in high stakes research would not be appropriate given the  $\alpha$  and  $\lambda_2$  scores and the  $\Phi$  and  $\rho^2$  coefficients do not exceed high stakes cut scores.

It is appropriate to use the scores on the RSES to make absolute decisions in low stakes research but not high stakes research. An example of scores on the RSES used to make absolute decisions in high stakes research is an intervention that used scores on the RSES to assess which participants were eligible for an early intervention to prevent relapse of schizophrenia (Gumley, Karatzias, Power, Reilly, McNay, & O'Grady, 2006). The participants were 144 individuals who had met three requirements. The first requirement was a schizophrenia, or a related diagnosis as defined by the DSM-IV (Frances, First, & Pincus, 1995). The second requirement was that these participants were considered prone to relapse and were receiving antipsychotic medication. The criteria to be considered prone to relapse included, a history of relapse, living in a stress full

environment, social isolation, non-adherence to antipsychotic medication, or participating in a neuroleptic dosage-reduction program (Gumley, et. al., 2006). The third requirement to participate in this intervention was scores on the RSES from 10 to 40, which indicates lower self-esteem (Gumley, et. al., 2006). This is considered an absolute decision because cut-off scores are being used to determine which participants are included in an intervention which could be beneficial and improve the quality of life of the individual. Absolute decision making in low stakes individuals would be appropriate given the  $\alpha$  and  $\lambda^2$  scores and the  $\Phi$  and  $\rho^2$  coefficients exceed the low stakes cut scores.

### **Conclusions**

The reliability of scores on the RSES have not decreased over time. If anything, they have increased slightly. It is important to re-asses the measures we use to determine if scores on these measures are still sound. While the scores on the RSES have been shown to be increasingly reliable from 1971 to 2019, the question of the validity of these scores still needs to be evaluated.

The self-esteem movement has not affected the reliability of the on the RSES, one of the most commonly used measures of self-esteem. However, the scores may be less valid than when the measure was developed in 1965. The self-esteem movement focused on boosting self-esteem and self-worth. This movement may have changed the way people view self-esteem and the RSES could now be measuring other attitudes, for example narcissism. If self-esteem is now viewed as an internal assessment of self that can be enhanced externally then the validity of the scores on the RSES need to be re-evaluated in future research.



### References

- Ashby, J. S., & Kottman, T. (1996). Inferiority as a distinction between normal and neurotic perfectionism. *Individual Psychology, 52*(3), 237.
- Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality, 40*(4), 440-450.
- Bachman, J., & O'Malley, P. (1977). Self-esteem in young men: A longitudinal analysis of the impact of educational and occupational attainment. *Journal of Personality and Social Psychology, 35*(6), 365-380.
- Bandura, A. (2010). Self-efficacy. *The Corsini encyclopedia of psychology, 1-3*.
- Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles?. *Psychological Science in the Public Interest, 4*(1), 1-44.
- Boyle, G. J., Saklofske, D. H., & Matthews, G. (Eds.). (2014). Measures of personality and social psychological constructs. Retrieved from <https://ebookcentral.proquest.com>
- (brownj@hawaii.edu), J. (2019). Questions and answers about language testing statistics: Generalizability and Decision Studies. [online] Hosted.jalt.org. Available at: [http://hosted.jalt.org/test/bro\\_21.htm](http://hosted.jalt.org/test/bro_21.htm) [Accessed 3 Dec. 2019].
- Brunet, J., Pila, E., Solomon-Krakus, S., Sabiston, C. M., & O'Loughlin, J. (2019). Self-esteem moderates the associations between body-related self-conscious emotions and depressive symptoms. *Journal of Health Psychology, 24*(6), 833-843.
- Callender, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's lambda-2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement, 16*(2), 89-99.

- Chae, J. (2018). Explaining females' envy toward social media influencers. *Media Psychology*, 21(2), 246-262.
- Clayton, R. R. (1973). Guttman scaling: An error paradigm. *Pacific Sociological Review*, 16(1), 5-26.
- Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice*, 39(3), 124-130.
- Dobson, C., Goudy, W. J., Keith, P. M., & Powers, E. (1979). Further analysis of Rosenberg's Self-esteem Scale. *Psychological Reports*, 44(2), 639-641.
- Foster, J. D., McCain, J. L., Hibberts, M. F., Brunell, A. B., & Johnson, R. B. (2015). The Grandiose Narcissism Scale: A global and facet-level measure of grandiose narcissism. *Personality and Individual Differences*, 73, 12-16.
- Frances, A., First, M. B., & Pincus, H. A. (1995). *DSM-IV guidebook*. American Psychiatric Association.
- Gardner, D. G., & Pierce, J. L. (1998). Self-esteem and self-efficacy within the organizational context: An empirical examination. *Group & Organization Management*, 23(1), 48-70.
- Gentile, B., Miller, J., Hoffman, B., Reidy, D., Zeichner, A., & Campbell, W. (2013). A Test of Two Brief Measures of Grandiose Narcissism: The Narcissistic Personality Inventory–13 and the Narcissistic Personality Inventory–16. *Psychological Assessment*, 25(4), 1120-1136.
- Google Scholar. (2019, November 23). Retrieved from [https://scholar.google.com/scholar?hl=en&as\\_sdt=0%2C3&q=rosenberg+self-esteem+scale&oq=](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C3&q=rosenberg+self-esteem+scale&oq=)
- Gumley, A., Karatzias, A., Power, K., Reilly, J., McNay, L., & O'Grady, M. (2006). Early

- intervention for relapse in schizophrenia: Impact of cognitive behavioural therapy on negative beliefs about psychosis and self-esteem. *British Journal of Clinical Psychology*, 45(2), 247-260.
- Grzegorek, J., Slaney, R., Franze, S., & Rice, K. (2004). Self-Criticism, Dependency, Self-Esteem, and Grade Point Average Satisfaction Among Clusters of Perfectionists and Nonperfectionists. *Journal of Counseling Psychology*, 51(2), 192-200.
- Hagborg, W. J. (1993). The Rosenberg Self-Esteem scale and Harter's Self-Perception profile for adolescents: a concurrent validity study. *Psychology in the Schools*, 30(2), 132-136.
- Humphrey, N. (2004). The Death of the Feel-Good Factor?: Self-Esteem in the Educational Context. *School Psychology International*, 25(3), 347–360.
- Laithwaite, H. M., Gumley, A., Benn, A., Scott, E., Downey, K., Black, K., & McEwen, S. (2007). Self-esteem and psychosis: a pilot study investigating the effectiveness of a self-esteem programme on the self-esteem and positive symptomatology of mentally disordered offenders. *Behavioural and Cognitive Psychotherapy*, 35(5), 569-577.
- Meyer, J. (2010). Reliability (Series in understanding measurement). Oxford ; New York: Oxford University PRESS.<https://doi.org/10.1177/0143034304046906>
- [Moore, C. T. \(2016\). \*gtheory: Apply Generalizability Theory with R. R package version 0.1.2.\* https://CRAN.R-project.org/package=gtheory](https://CRAN.R-project.org/package=gtheory)
- Orth, U., & Robins, R. W. (2019). Development of self-esteem across the lifespan. *Handbook of Personality Development*, 14(3), 328-344.
- Park, S. Y. (2005). The influence of presumed media influence on women's desire to be thin. *Communication Research*, 32(5), 594-614.
- Pmean.com. (2019). Stats: What's a good value for Cronbach's Alpha? (September 9, 2004).

[online] Available at: <http://www.pmean.com/04/CronbachAlpha.html> [Accessed 3 Dec. 2019].

PsycINFO. (2019, April 30). Retrieved from

[http://eds.a.ebscohost.com.ezproxy1.library.arizona.edu/ehost/RESultsadvanced?vid=2&sid=80819c0c-dbd2-46c2-8a0f-c3b0f0efe69d%40sdc-v-sessmgr05&bquery=TM+rosenberg+self-  
esteem+scale&bdata=JmRiPXBzeWgmdHlwZT0xJnNIYXJjaE1vZGU9U3RhbmRhcmQ  
mc2l0ZT1laG9zdC1saXZl](http://eds.a.ebscohost.com.ezproxy1.library.arizona.edu/ehost/RESultsadvanced?vid=2&sid=80819c0c-dbd2-46c2-8a0f-c3b0f0efe69d%40sdc-v-sessmgr05&bquery=TM+rosenberg+self-<br/>esteem+scale&bdata=JmRiPXBzeWgmdHlwZT0xJnNIYXJjaE1vZGU9U3RhbmRhcmQ<br/>mc2l0ZT1laG9zdC1saXZl)

[R Core Team \(2019\). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.](https://www.R-project.org/)

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, 54(5), 890.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 27(2), 151-161.

Rosenberg, M. (1965). Rosenberg self-esteem scale (SES). Society and the adolescent self-image. Princeton, NJ: Princeton.

Rosenberg, M. (1979). *Conceiving the self*. New York: Basic Books.

Rotter, J. (2011). Rotter internal-external locus of control scale. *28 Measures of Locus of Control*, 10.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension*, 37(2), 1-5.

- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Sage.
- Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports*, *51*(2), 663-671.
- Silber, E., & Tippett, J. S. (1965). *Self-esteem: Clinical assessment and measurement validation*.
- Slaney, R. B., Mobley, M., Trippi, J., Ashby, J. S., & Johnson, D. G. (1996). The almost perfect scale-revised. *Unpublished manuscript, The Pennsylvania State University*.
- Stephanie. (2018, August 14). Guttman's lambda-2: Definition, Examples. Retrieved December 2, 2019, from <https://www.statisticshowto.datasciencecentral.com/gutmans-lambda-2/>.
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273-1296.
- Tafarodi, R. W., & Swann Jr, W. B. (2001). Two-dimensional self-esteem: Theory and measurement. *Personality and Individual Differences*, *31*(5), 653-673. *Psychological reports*, *16*(3\_suppl), 1017-1071.
- Tracy, J., Cheng, J., Robins, R., & Trzesniewski, K. (2009). Authentic and Hubristic Pride: The Affective Core of Self-esteem and Narcissism. *Self and Identity*, *8*(2-3), 196-213.
- Twenge, J. M., Konrath, S., Foster, J. D., Keith Campbell, W., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the Narcissistic Personality Inventory. *Journal of Personality*, *76*(4), 875-902.
- Wang, A. Y., & Richarde, R. S. (1988). Global versus task-specific measures of self-efficacy. *Psychological Record*, *38*, 533-541. Retrieved from <http://ezproxy.library.arizona.edu/login?url=https://search-proquest-com.ezproxy3.library.arizona.edu/docview/57879190?accountid=8360>

Whiteside-Mansell, L., & Corwyn, R. F. (2003). Mean and covariance structures analyses: An examination of the Rosenberg Self-Esteem Scale among adolescents and adults.

*Educational and Psychological Measurement, 63*(1), 163-173.

## Data Sources

Bengtson, V. L., (2008): *Longitudinal Study of Generations, 1971, 1985, 1988, 1991, 1994, 1997, 2000, 2005* [California]. Archival Version. Version: v0. ICPSR - Interuniversity Consortium for Political and Social Research. Dataset.

<http://doi.org/10.3886/ICPSR22100>

Hunt, T. (2013). *Lambda4: collection of internal consistency reliability coefficients*. R package version 3.0. Retrieved July 1, 2019.

NLYS, 1979, "*National Longitudinal Survey of Youth*",

<https://www.nlsinfo.org/content/cohorts/nlsy79>

### Appendix

#### Rosenberg Self-Esteem Scale (RSES)

Please record the appropriate answer for each item, depending on whether you Strongly agree, agree, disagree, or strongly disagree with it.

- 1 = Strongly agree
- 2 = Agree
- 3 = Disagree
- 4 = Strongly disagree

- \_\_\_\_\_ 1. On the whole, I am satisfied with myself.
- \_\_\_\_\_ 2. At times I think I am no good at all.
- \_\_\_\_\_ 3. I feel that I have a number of good qualities.
- \_\_\_\_\_ 4. I am able to do things as well as most other people.
- \_\_\_\_\_ 5. I feel I do not have much to be proud of.
- \_\_\_\_\_ 6. I certainly feel useless at times.
- \_\_\_\_\_ 7. I feel that I'm a person of worth.
- \_\_\_\_\_ 8. I wish I could have more respect for myself.
- \_\_\_\_\_ 9. All in all, I am inclined to think that I am a failure.
- \_\_\_\_\_ 10. I take a positive attitude toward myself.

(Rosenberg, 1965)