EXPLORATORY STRUCTURAL EQUATION MODELING AS A TOOL FOR VALIDATION OF SCALES USED TO MEASURE RESEARCH SELF-EFFICACY AND RESILIENCE IN RESEARCH TRAINING PROGRAMS FOR UNDERREPRESENTED STUDENTS

by

W. Haydon Ekstrom

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2025

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: *WILLIAM HAYDON EKSTROM*, titled: *EXPLORATORY STRUCTURAL EQUATION MODELING AS A TOOL FOR VALIDATION OF SCALES USED TO MEASURE RESEARCH SELF-EFFICACY AND RESILIENCE IN RESEARCH TRAINING PROGRAMS*

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Ronald W Marx
Ronald W Marx (Apr 28, 2025 09:24 PDT)
Ron Marx, Ph.D. Date: 04/28/2025

Jessica J Summers
Jessica Summer, Ph.D. Date: 04/28/2025

Lia D Falco
Lia Falco, Ph.D. Date:

Carol Brochin (Apr 28, 2025 08:55 PDT)
Carol Brochin, Ph.D. Date: 04/28/2025

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

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

Ronald W Marx
Ronald W Marx (Apr 28, 2025 09:24 PDT)
Ron Marx, Ph.D. Date: 04/28/2025
Dissertation Committee Chair
Educational Psychology

Acknowledgements

I am deeply grateful to my committee members, whose support and belief in me helped carry me through the most challenging experience of my academic career. Their guidance, encouragement, and patience made all the difference at times when I doubted myself.

I owe a profound debt of gratitude to my family. Their unwavering support convinced me that I could withstand the stress and uncertainty that accompanied this journey. To my sister, thank you for always reminding me that turmoil is temporary, and for helping me keep perspective when it felt like the challenges would never end. To my brothers, thank you for bringing humor and music into my life when I needed them most, often without even realizing how much I needed them. To my parents, your selflessness, steady encouragement, and faith in me have been the foundation beneath every step I have taken toward this achievement.

Above all, I want to thank my wife. Her unconditional love, patience, and daily encouragement were what made me believe that I could finish this doctorate, even when I could not always see it myself. This accomplishment would not have been possible without her.

Table of Contents

List of Tables/Figures

Abstract

Using multiple forms of structural equation modeling, this study demonstrates a progression of analyses from the most basic confirmatory factor analysis through more complex exploratory structural equation modeling to interrogate both the construct and discriminant validity of the RSES and CD-RISC-10 when used as an outcome assessment for research training programs that target underrepresented students. Not only does this study inform the understanding of the factor structures for both measures, but it also demonstrates the utility of using more complex modeling for localized instrument validation. The psychometric assessment of the CD-RISC-10 found adequate model fit using CFA of the original factor structure, but improved model fit utilizing a two-factor structure defined through ESEM. However, the modifications made to the original RSES before being applied to this population resulted in more model fit and factor loadings inconsistent with its original form. Potential adjustments to the RSES and suggestions for interpretation are discussed.

Keywords: structural equation modeling (SEM), exploratory structural equation modeling (ESEM), Research Self-Efficacy, Resilience

CHAPTER I

Introduction

In the design of any educational intervention, demonstrating the psychometric quality of the instruments used to evaluate that intervention is crucial to ensuring the scientific soundness of the research design and properly informing changes to that would enhance the quality of the intervention. Interventions targeting the improvement of less easily quantified attributes such as self-efficacy and resilience need to demonstrate that the instruments evaluating outcomes are valid in the context of their use. This importance is amplified when those measures are being applied outside of the context in which they were originally developed and validated or have been modified significantly from their original form. This dissertation examines two such measures, the Research Self-Efficacy Scale (RSES; Brancolini & Kennedy, 2017; Greeley et. al., 1989) and the Connor Davidson Resilience Scale (CD-RISC-10, Connor & Davidson, 2003) that are being used to evaluate the outcomes of research training programs for underrepresented undergraduate students. Each measure is being applied to a population outside of which they were previously validated, and one of them, the RSES, has been modified to fit the specific evaluation goals of the programs from which the data used in this study comes. Determining the validity of these measures requires not only new psychometric testing using rigorous forms of analysis but also a reconsideration of how the content is being interpreted and utilized.

Problem Statement

The need for psychometrically sound measures to evaluate research training programs for underrepresented students is evident, given the tenuous position of significant investment in such initiatives (National Center for Education Research, 2024). The context-dependent nature of validity in measurement, particularly in educational settings, is a crucial consideration when assessing these programs. For the research training programs, this would involve ensuring that the assessment tools adequately represent all facets of the intended outcomes. Modifying existing measures like the RSES and CD-RISC-10 to address program outcomes more directly necessitates reevaluation in this new context. To do this adequately requires assessments of both the construct validity of individual measures in context and potential issues with discriminant validity that arise from the tandem use of assessments for closely related attributes.

Localized validation of measures used for psychological assessment is also important from a social justice perspective. In an official apology published in 2021, the American Psychological Association asserted that the widespread misuse and overuse of instruments for psychological measurement perpetuated social and racial inequities (American Psychological Association, 2021). Wearing a mask of scientific objectivity, uncritically accepted psychological instruments were used to legitimize oppression. Many of the ideas these instruments used to support left a legacy of systemic issues. For example, measures claiming to test intelligence were used to make decisions about the kinds of jobs people could do or the kinds of scholastic opportunities they would be provided. For decades, the Stanford-Binet scale claimed to place students into special education programs based on intelligence. However, the measure was subject to such cultural bias and interpreter bias that, in effect, it only served to separate Black and immigrant children from opportunities taken advantage of by their white peers (Garcia & Hamilton, 2023; Terman & Merrill, 1937). Even more recently developed tests of intelligence are primarily normed on white, middle-class people, leading to innately lower scores for people of color. (Ford, 2004).

Measures with questionable psychometric properties are still used to make high-stakes decisions, notably in the justice system. Intelligence testing has been used to determine parental fitness in custody hearings, despite not accurately representing the full range of qualities necessary for effective parenting, such as emotional responsiveness and the ability to ensure a child's physical and emotional safety. (Brodzinsky 1993, Sanders & Katz, 2013). Tests intended to assess specific attributes in clinical populations have been used to determine whether someone who is suspected of a crime is lying when denying guilt (Wakefield & Underwager, 1993). Additionally, highly subjective measures such as Rorschach tests are still commonly utilized, and the conclusions drawn from them are seldom challenged on the grounds of validity (Neal, 2020; Viglione et al., 2022). Despite the questionable validity of their application in this way, judges and juries, who are not experts in psychological testing and cannot spot misuse, often value the outcomes of psychological tests over practical observations (Sanders & Katz, 2013). In a large-scale analysis of "junk science" in the courtroom, Neal et. al. (2020) found that psychological tests with poor or unknown psychometric properties are disproportionately used to convict BIPOC defendants. While accepting misapplied and flawed assessments in the courts has direct and detrimental acute effects, accepting conclusions of studies without examining their methods can result in harmful assumptions about psychological attributes.

Shoda, Mischel, and Peake (1990) claimed to be able to predict cognitive ability and self-regulatory competency via a test of delay of gratification and further connected the ability to delay gratification to numerous other positive gains later in life, such as lower rates of poverty. The test in question, the now-famous "marshmallow test," involved providing a child with a marshmallow and telling them that if they could refrain from eating it for a period of time, they would be given an additional marshmallow. The amount of time children could wait was compared with achievement in their teenage years. However, the actual size of the gains in achievement from delay of gratification was called into question by later researchers who noted that the original research largely examined a homogeneous group of children with regard to parental environments. Nearly three decades later, Watts, Duncan, and Quan (2018) sought to replicate the study, but primarily focused on children whose mothers had not completed college. In a larger and more diverse sample, while controlling for socio-economic status, family background, and home environment, the strength of the correlation between early delay of gratification and later achievement markedly diminished. The findings of Watts, Duncan, and Quan (2018), as well as a later examination of the marshmallow test by Falk, Kosse, and Pinger (2020), suggest that a child's home environment may be what is truly predictive of future achievement. Without a critical reexamination of the original findings of the marshmallow test, achievement may have continued to be attributed to an inherent characteristic, such as the ability to delay gratification, rather than circumstances outside the control of a child. It is crucial that both the measures used to evaluate psychological attributes and the conclusions drawn from the outcomes of those measures are critically examined.

Research objectives

This study aims to provide a psychometric assessment of the RSES and CD-RISC-10 in the context of the research training programs for undergraduate students. This study will assess construct validity from multiple angles, examining how well the data fit present theoretical models and exploring how the latent variables in those models may change in this new context. Bandura (2006) noted that resilience and self-efficacy are closely related attributes. Therefore, confirming that the two measures used to assess those attributes can distinguish between them is important. Finally, this study offers guidance for interpreting the data collected using these measures to ensure validity in its applications, not only in its psychometric properties.

Significance of the Study

Providing evidence for the validity of these measures when used in the context of research training programs will not only impact the interpretation of the outcomes of the programs but also inform the use of these tools when evaluating undergraduate research training programs in the future. Further, examining a progression of model fit utilizing more complex forms of structural equation modeling will allow future researchers and evaluators to more completely grasp the complexity of the constructs of research self-efficacy and resilience, aiding in the interpretation of data produced by the RSES and CD-RISC-10. Allowing for cross-loading of observed variables onto latent variables that are theoretically distinct from them, as in ESEM, informs the nature of conclusions that can be drawn from these data.

Research Questions

This dissertation answers the following questions and tests the following hypotheses:

Question 1: Does the established single-factor structure of the CD-RISC-10 demonstrate better model fit than a two-factor structure when applied to a population of participants in research training programs for underrepresented undergraduate students?

Hypothesis 1: A two-factor solution will demonstrate an improved model fit compared to the original unifactorial model.

Question 2: Does the factor structure of the RSES demonstrate adequate model fit when modified to include only 3 of 4 originally defined latent variables?

Hypothesis 2: The modifications made to the RSES will result in poor model fit. Changes to the factor structure will be necessary to identify a functional abbreviated version of the measure.

Question 3: Do the RSES and CD-RISC-10 discriminate from one another while simultaneously demonstrating a positive correlation, as theorized by Bandura (2007)?

Hypothesis 3: Significant cross-loading between items will indicate overlap between the two measures.

Question 4: Does exploratory structural equation modeling (ESEM) produce better model fit statistics and more useful factor structure for interpreting the complexity of both research self-efficacy and resilience as constructs, compared to traditional CFA?

Hypothesis 4: Utilizing ESEM will produce better model fit statistics and the cross-loadings identified in the factor structure can be used to better interpret the outcomes of the RSES and CD-RISC-10.

CHAPTER II

Theoretical Framework and Literature Review

This chapter details the multiple theoretical frameworks at play in this work. These frameworks encompass the basis for psychological measurement and validity testing, the foundational understanding of self-efficacy as a component of social cognitive theory, and the conceptualization and operationalization of resilience as both a state and a trait.

Measurement

To fully understand the methods and conventions of psychological measurement and test validity, it is important to understand the philosophical underpinnings that form the theoretical framework of measurement theory. The question of what can be measured is both practical and philosophical and has no definitive answer. For many early measurement theorists and physical scientists, it was a common assumption that only physical attributes of an object (i.e. mass, length) could be measured (Campbell, 1920; Stevens, 1946; Jones, 1971). Even those who were later instrumental in the development of psychological measurements were skeptical of its feasibility (Thorndike, 1918; Guilford, 1946).

The perspective that anything outside of the physical is inherently immeasurable continues with a number of academics and scientists arguing that psychological constructs lack the definitive clarity, numerical properties, and empirical basis for true measurement (Michell, 2003, 1997; Borsboom, 2004). However, this approach to measurement both underestimates the extent to which psychological attributes are normally distributed and overestimates the extent to which physical properties are stable (Thorndike, 1927; Jones, 1971; Cronbach, 1971 Heisenberg, 1927). The Central-Limit-Theorem holds that any attribute that is made up of a large number of relatively independently functioning factors tends to normally distribute (de Moivre, 1738). To that end, one can consider any psychological attribute the sum of a large number of influences (Jones, 1971; Thorndike, 1927). However, being the sum of many influences also creates an opportunity for more extensive errors in measurement.

Any measurement procedure is impacted by a component of random error. In psychology intra-person variability can result in significant errors. The errors produced by variability in an individual may not correlate to the errors produced by another individual on the same measure. When looking at something measured by a single indicator (i.e., a direct question) the impact of measurement error on normality can be diluted with more samples. For example, asking 10 people “How happy are you?” would likely produce less normally distributed data than asking 10,000 people “How happy are you?”. But often there is not the capability to get thousands more responses on a measure. Utilizing a latent variable model, in which multiple questions are asked to assess the same attribute, can have the same mitigating effect on normal distribution (de Ron et. al. 2022). For example, asking 10 people a series of questions designed by content experts to address the central construct of happiness, such as close non-familial relationships, self-confidence, and loneliness, can provide a more reliable assessment in a smaller sample than a single direct question (Cheng & Furnham, 2002).

One measurement issue that psychologists and education researchers must reconcile is how measurement error is treated. A general assumption necessary for many statistical tests is that the errors are uncorrelated (Raykov, Marcoulides & Patelis, 2015). When true random sampling is used to gather data, it is more likely that the errors caused by intra-person variability are not correlated. Human beings are complex, and the dramatically different life experiences each person has influence errors in measurement. However, the reality of studying human beings is that a truly random sample is often not feasible. Even when random sampling methods are used, the pool of participants being sampled often shares geographical and cultural proximity. Due to these connections, it is very likely that errors assumed to be uncorrelated are actually related in a way that cannot be accounted for in the measurement. New and more complex psychometric models are needed to account for these potentially correlated errors (Marsh et. al., 2011; Morrin, Arens, & Marsh, 2014; Marsh, Morin, Parker & Kaur, 2014; Summers & Falco, 2020).

Another theoretical hurdle for measuring psychological attributes is the absence of a meaningful *zero* point for psychological measures. *Zero* on a physical measure represents the absolute absence of something. However, it is much more difficult to meaningfully describe what it looks like to have *zero* self-efficacy or *zero* resilience. One might think of a particular set of behaviors and traits that could be interpreted as having none of those attributes, but that description would vary based on the observer. Stevens (1946) introduced levels of measurement that are commonly used to describe data today, and in that text stressed that the most precise measurements require a true zero point. Other researchers question whether anything without a true zero point can be accurately measured at all (Borsboom & van; Cliff, 1992; Hand 1996; Heerden, 2004; Lord & Novick, 1968; Michell, 1997, 1999). Their skepticism is primarily rooted in the idea accepted since the early 20th century that ratio scales are the most precise and *must* contain an absolute zero to be valid (Campbell, 1920; Stevens 1946).

While the presence of an absolute zero point on a scale aids in *precision*, it is not required for that scale to be *useful*. What changes primarily is how that measure can be interpreted and applied. For example, the summing of scores on a measure with no absolute zero value is meaningless, as the difference between one point and another is not determinable. However, the difference between one score and the *mean* of a set of scores is meaningful and does have an absolute zero point at which the score is identical to the mean. (Jones, 1971). By taking the mean of a set of scores and comparing any individual score to that mean, one creates a ratio scale that can be interpreted. The degree to which an individual score differs from the mean score conveys valuable information.

Just as with assessments of psychological attributes, assessments of physical properties are “limited by the smallest resolution magnitude that can be detected by the instrument used” in that measurement. (Jones, 1971, pg. 346). That measurement is then subject to the errors that occur when the observer is interpreting the scale, and human beings are notoriously not free from error in observation and interpretation (de Ron, et. al. 2022). Absolute measurements for physical or psychological attributes are therefore not practically possible. The question then is not whether or not something can be measured, but the degree to which interpretation by the observer is involved. How a measure is interpreted and applied contributes significantly to that measure's validity.

Validity

Many early works on test validation put the onus for validity on the measure itself. (Bingham, 1937; Cureton, 1951; Guilford, 1946; Gulliksen, 1950; Shaw & Crisp, 2011). Validity was simply an issue of whether a test measures what it purports to measure (Kelley, 1927). A valid test was a test used to measure “anything with which it correlates” (Guiliford, 1946, p. 429). However, early educational measurement researchers argued strongly for the application of context when considering the issue of validity (Thorndike, 1918).

Much in the way one does not measure an object, one measures an *attribute* of that object (i.e. I am not measuring a tree, I am measuring the *height* of a tree), statistical validation is not a validation of a test: it is the validation of “an interpretation of data arising from a specified procedure" (Cronbach, 1971, p. 447; Sierci, 2007). At its simplest level, determining validity is done by comparing the scores on a measure to another observation or measure that addresses the same attribute. However, oversimplifying how this comparison is made neglects the contextual nuances at play and represents only a small aspect of validity.

Types of validity are clearly distinguishable. Broadly, validity can be discussed in terms of either *discriminant validity* or *construct validity* (Shaw & Crisp, 2011). The former answers the question of whether something other than the desired attribute is being assessed (Campbell & Fiske, 1959). Do scores on one measure correlate strongly with another measure that is evaluating a supposedly independent attribute? If so, the discriminant validity of the measure comes into question. It cannot *discriminate* between those two attributes.

Construct validity addresses the more traditional conception of validity, asking "does the test measure the attribute it is said to measure?" (Cronbach, 1971, p. 446) Construct validity encompasses three subcategories: predictive, criterion, and content validity (Cronbach, 1988; Cronbach & Meehl, 1955; Meehl, 1954; Westen & Rosenthal, 2005). To have predictive validity, a test must be interpreted so that future outcomes can be accurately inferred from scores on that test. For example, scores on a valid driving safety exam should predict the driving practices of the individual being assessed. Criterion validity is the classical conception of validity: how well do scores on a measure correlate with another measure known to address the same attribute? For example, there is a positive correlation between SAT scores and first-year college GPA, because both measure academic ability (Korbin et al., 2008). Content validity examines the appropriateness of the items on the exam. Cronbach (1971) plainly states that "correlations have nothing to do with content validity." There may well be a negative correlation between the number of horse-drawn carriages in use and the rate of climate change. However, the connecting factor is that technology has changed with time, and the rate of climate change increases over time. Including a question about horse-drawn carriages would not have content validity in assessing climate change rates. Content validity affirms whether the test is relevant to the domain in which it is used and is a fundamental prerequisite for construct validity (American Educational Research Association et al., 2014; Messick, 1989). Determining content validity often relies on domain experts to verify that the questions asked accurately capture the complexity of the attribute they are intended to measure. (Cronbach, 1971).

Ultimately, a decision about validity depends on the utilization of the test. “Validation requires a clear statement of the proposed interpretations and uses” of the test (Kane, 2006, p. 23). There needs to be sufficient evidence that supports the use of a test not generally, but for its specific intended purpose (Sireci, 2007). It requires an understanding of how outcomes of the test are intended to be applied (Cronbach, 1971; Cronbach and Meehl, 1955; Messick, 1989; Moss 1992). The standards for validity on a test used to inform changes to an educational program are likely not the same standards that should be used to determine the validity of a test designed to determine the effectiveness of cancer treatments. Statistically valid models and theories do not provide rules for action but rather aid in interpreting results in a given context (Cronbach, 1971). To accurately make judgements on the validity of a measure requires an understanding of how that measurement is intended to be used, whether it is for assessment or instructional purposes, for example (Moss, 1992). Attention to the social consequences, particularly in performance-based assessments, is integral to appropriately determining validity (Moss, 1992).

Given enough data and enough iterations, a theoretical model of how a particular attribute functions can be developed (Jones, 1971). An approach to assessing general construct validity is the use of structural equation modeling (SEM) to compare a theoretical model structure of the attribute with the model structure from real data (Bacci & Crutchfield, 2020). Two commonly used approaches to using SEM for validity testing are exploratory factor analysis (EFA) followed by a confirmatory factor analysis (CFA) of the structure that the EFA suggests.

Self-Efficacy

Self-efficacy describes one's belief that they can perform a given task and is a key component within broader social cognitive theory. (Bandura 1977, 1978, 1981, 1982, 1986 1997). "This theory states that psychological procedures, whatever their form, alter the level and strength of *self-efficacy*" (Bandura, 1977, p. 191). Self-efficacy is derived from four primary sources. In order of strength of impact, those sources are previous accomplishments and mastery experiences, vicarious experiences, verbal support, and psychological state (Bandura 1977, 1978; Bandura & Adams, 1977; Weinberg, Gould & Jackson, 1979; Zimmerman, 2000).

The first of these influences, previous experiences, describes the extent to which the individual has accomplished the task in the past. If one has already accomplished a task or previously demonstrated mastery of that task, it is reasonable to expect that they feel good about their ability to accomplish it in the future. Vicarious experience describes the extent to which an individual has seen a task successfully completed in the past. If someone of a similar skill level accomplishes something an individual is attempting, that individual may feel better about their ability to accomplish it. Beyond personal and vicarious experience, the level of verbal support one receives can also increase self-efficacy when attempting new tasks. One may feel better attempting something new if another person offered encouragement. Or conversely, feel worse if someone has discouraged them. The final influence is one's physiological state. For example, a singer with a strep throat may have lower levels of efficacy around performing a song.

While outcomes do inform behavior, self-efficacy has been shown to influence behavior to a stronger degree than outcome experiences, the belief that a behavior will result in desired outcomes (Bandura, 1977; Shell, Murphy, & Bruning, 1989; Zimmerman, 2000; Zimmerman, Bandura, & Martinez-Pons, 1992). In fact, self-efficacy strongly predicts both outcome and persistent behavior (Schunk, 1981). Further, interventions focusing on fostering self-efficacy are likely to have a longer-lasting effect on behavior than outcome expectancies alone (Bandura & Adams, 1977). One method of improving and sustaining improvements in self-efficacy is reframing long-term distal goals into several more proximal ones. Multiple proximal goals provide an ongoing example of improved achievement towards a distal goal that the learner can recognize as concrete improvements in their ability (Bandura & Schunk, 1981). These incremental achievements are a sort of self-generated reinforcement, ensuring that changes in behavior are intrinsically motivated and more likely to be long-lasting (Deci et al., 1991; Ryan & Deci, 1985, 1987, 2020; Zimmerman, Bandura, & Martinez-Pons, 1992).

Self-efficacy influences behavior in various ways beyond the likelihood of behavior being repeated or the probability of achieving a given outcome. As far as completing tasks, self-efficacy beliefs determine whether someone attempts a task, how much effort they put into that task, and how persistent they will be at achieving that task, especially when that task is challenging (Bandura, 1977, 1978; Bandura & Adams, 1977). Research has shown that self-efficacy motivates behavior more than fear of an outcome or experience (Bandura, 1983). Interventions targeting self-efficacy through repeated exposure have reduced phobic reactions by changing how people perceive their ability to manage a given threat (Bandura, 1983; Bandura & Adams, 1977). “Persistence in activities that are subjectively threatening but in fact relatively safe procedures, through experiences of mastery, further enhancement of self-efficacy and corresponding reductions in defensive behavior" (Bandura, 1977, p. 191).

As previously discussed, self-efficacy is primarily informed by personal experiences with a task, vicarious experiences, verbal support (encouragement/discouragement), and physical state. (Bandura, 1977). Addressing one influence without the others is likely an ineffective way to foster growth in a skill. For example, allowing a student to attempt the task in an environment in which failure to complete the task has minimal consequences provides mastery experience that directly informs their self-efficacy beliefs around that task. However, the process of trial and error for a task may not provide the quickest route for learning a skill initially, and verbally encouraging the student to perform a task for which they have no context might provide an inaccurate perception of the task's demands. That is particularly true if the task has not been modeled for the learner. Without context or verbal support, the learner may attribute their failure to unchangeable qualities, leading them to no longer pursue the task. "People convinced vicariously of their inefficacy are inclined to behave in ineffectual ways that" confirm their low efficacy beliefs (Bandura, 1981, pg. 203). Repeated discouraging experiences create a risk for self-fulfilling prophecy. A combination of rehearsal, modeling, and constructive feedback typically accelerates skill acquisition through self-efficacy (Bandura, 1978).

Modeling a task is an effective way to begin self-efficacy intervention (Bandura, 1978). Compared to didactic instruction, modeled instruction resulted in more significant increases in efficacy and skill development (Schunk, 1981). Because misattributions about the feasibility of a task are often influenced by a lack of complete information about what the task involves, modeling the learning task can provide important contextual information needed for the learner to accurately identify the cause of the outcome. "People see the extraordinary feats of others, but not the unwavering commitment and countless hours of persevering effort that produced them" (Bandura, 1997, p. 119). Knowing what kind of efforts and regulatory processes are needed to perform a task successfully, and providing context for metacognitive reflection, allows for effective evaluation of outcomes such that efficacy can be increased (Bandura, 1997; Schunk, 1981, 2012).

Within the context of skill acquisition, the primary aspects of self-efficacy at play are the *level* and *strength* of self-efficacy. *Level* refers to the degree to which self-efficacy depends on the task's difficulty. *Strength* of self-efficacy refers to the degree of certainty one has around the ability to perform a task (Bandura, 1997). However, the extent to which that efficacy transfers across tasks merits exploration. Improvements in self-efficacy for one task can change behavior in others. For example, Bandura and Adams (1977) found that efficacy interventions that addressed one phobia also reduced avoidant behavior around unrelated tasks. In another study, participants with higher levels of self-efficacy demonstrated more positive self-talk and were less likely to attribute successes to luck in future tasks (Weinberg, Gould & Jackson, 1979). While efficacy in one task may be generalizing to efficacy in another, it is important to note that this does not indicate general self-efficacy. Self-efficacy is *task-specific*. This distinguishes it from constructs such as self-esteem: You can have generally low self-esteem about something (i.e., I am bad at math) and high self-efficacy around a specific, related task (i.e., I can add two numbers together). The reverse is also possible (Bandura, 1997; Zimmerman, 2000). This distinction is particularly important when conceptualizing tools for the measurement of self-efficacy.

"[M]easures of self-efficacy are not only conceptually distinct from closely associated constructs such as outcome expectancies, self-concept, and perceived control, they have discriminant validity in predicting a variety of academic outcomes" (Zimmerman, 2000, pp. 85-86). Bandura (2006) provides a guide for creating self-efficacy scales. First and foremost, scales for self-efficacy need to be domain-specific. They must address the issue of self-efficacy while situated within the task of interest, capturing a variety of difficulty levels to cover the varying demands involved in the task; the determination is never "efficacious or not". To adequately capture this range, Bandura (2006) recommends the use of items on a 0-100 scale, going from "cannot do at all" to "highly certain can do". The scales must also be temporally situated, evaluating what someone thinks they *can do now*, not what they think they *will do* in the future.

These measures should be administered privately, not only to minimize social pressure, but also to aid the scales' nonreactivity; the scale's administration should not alter behavior. To that end, efforts to minimize stereotype threat should be taken. For example, if examining self-efficacy in math, reminding the participant of stereotypes connecting race and gender to mathematical ability alters the outcome of the assessment such that it will not accurately represent the sample (Appel & Kronberger, 2012; Logel & Davies, 2016; Ryan & Ryan, 2005). This may mean taking steps to account for autonomic perception that occurs during the test-taking process. Feltz and Mungo (1983) found that self-efficacy mediates autonomic perception and performance.

Bandura (2006) also speaks to the necessity of strong psychometric interrogation of the measure, stating that it should show accurate reliability (measured through *Cronbach's alpha*) and provide predictive power over other variables such as motivation, perseverance, and emotional resilience within the respective domain. While many of these guidelines for measuring self-efficacy are modern best practices for executing any survey today, special attention should be paid to the criterion for validity laid out by Cronbach (1971) regarding how the data produced by the measure will be interpreted (Blair, Czaja & Blair, 2013). Any survey designed to examine self-efficacy in a specific domain can *only* be interpreted in the context of that domain.

Research Self-efficacy

A goal of this work is to evaluate a measure that examines self-efficacy in the context of performing research. Improving self-efficacy around research tasks is a primary goal for many research training programs. Incorporating interventions designed to enhance self-efficacy can synergize, improving the training program's effectiveness (Goulart et al., 2022). These programs aim to create a more equitable landscape for scientific inquiry (National Center for Education Research, 2024). Studies have shown that the synergistic effect is particularly salient for students from underrepresented groups (Butz et al., 2018). Improving research self-efficacy can help address disparities in research fields. Increased research self-efficacy also improves disposition toward research, mediating an increase in scholarly productivity (Hemmings & Kay, 2016). Studies on research training programs have consistently found that those with higher research self-efficacy tend to have increased publication output (Hollingsworth & Fassinger, 2002; Hemmings & Kay, 2010; Pasupathy & Siwatu, 2014; Hemmings & Kay, 2016; Livinţi et al. 2021).

There are several measures that are used to examine research self-efficacy. A modified version of the 53-item Research Self-Efficacy Scale (Bieschke et al., 1996; Greeley et al.,1989) that had been reduced to 38 items was used to evaluate the research training programs of interest to this study. Research suggests that because the 38-item RSES was validated on present doctoral students, it is more suitable for those trained in research practices (Bieschke et al., 1996; Brancolini & Kennedy, 2017). This abbreviated version was used because these programs are designed to equip undergraduate students, who often do not have experience with research skills, with the necessary knowledge and efficacy to do research at a professional level. Hierarchical regression of this measure identified three components of research self-efficacy: early tasks, conceptualization, and implementation (Brancolini & Kennedy, 2017). These three components were originally identified by Bieschke et al. (1996) as being the components from the 53-item version that accounted for unique variance in whether the participant was interested in pursuing further research.

Resilience

Resilience refers to one's ability to attain positive outcomes despite challenging circumstances, allowing one to effectively cope with stress or trauma (Zoloski & Bullock, 2012). Resilience is not necessarily attached to achievement outcomes, but rather to the absence of detrimental outcomes that are produced by adverse experiences. For example, many people will encounter some form of trauma within their lives, but only a small few will develop conditions such as PTSD or depression in response to those traumas (Wu et al., 2013). More resilient individuals often display more empathy and better problem-solving skills, and children who are more resilient tend to be more future-oriented (Cowen et al., 1992; Seginer, 2008). Resilience also plays a role in many major life decisions such as what topics to study or what jobs to pursue (Herman et. al. 2011)

Resilience and self-efficacy are undeniably linked. Bandura (1977, 1982) established that efficacy plays an important role in our decisions about whether to persevere in difficult situations by shaping how one interprets their abilities and informing what they believe they are capable of enduring. Bandura (1978) noted that self-efficacy informs emotional responses to challenging situations, and that those with high self-efficacy tend to manage stress in healthier ways, an observation also made of resilience by McLafferty, Mallet, and McCauley (2012). Some researchers have examined resilience as an attribute that manifests as self-efficacy (Boardman et al., 2011). Maddux (2013) also connected self-efficacy with the ability to adapt to changing and stressful circumstances. Measures of self-efficacy should be strong predictors of resilience (Bandura, 2006; Martínez-Martí & Ruch, 2017). With the two constructs being so intertwined, it is a challenge for measures of resilience to show that what they are assessing is truly resilience above and beyond self-efficacy.

Resilience manifests both as a trait, enduring for a longer period, and as a state, more temporal and circumstantial (Child & Medvedev, 2024). One may be generally more resilient to stressful situations, but have little resilience to a specific stressor. However, examinations of multiple scales used to evaluate resilience within the same people at different timepoints showed little variation by time point or person, lending strength to the idea that resilience can be measured as a generally stable trait (Child & Medvedev, 2024). While generally stable across evaluative measures, when looked at over a lifespan, resilience is a "mutable, scalable, dynamic and fluid characteristic…and is largely context dependent". (Denckla et. al. 2020, p.3; Herman et. al. 2011)

In fact, resilience can be fostered through experience and intervention. Research into resilience in children has found several factors that support its development. An environment that is loving and supportive, that avoids repeated exposures to uncontrollable stress and trauma, such as abuse or war, promotes resilience in developing children (Wu et al., 2013). Additionally, experiences overcoming manageable life challenges can promote lifelong resilience, much like mastery experiences promote self-efficacy (Bandura, 1977; Wu et al., 2013). Interventions that impact skills like problem-solving and developing autonomy can also increase resilience (Waxman et al., 2003). Resilience also predicts more positive coping methods, particularly in high stress situations (McLafferty et al., 2012; Zeuger et. al., 2022). Researchers have argued that interventions and public policy that support fostering resilience are crucial for long-term national wellbeing (Herrman et. al. 2011). However, there are potential sources of resilience that are not easily accessed through intervention. There are known neurobiological markers and epigenetic factors that contribute to an individual's resilience to stress, including abnormal amygdala activity (Wu et al., 2013; Denckla et al., 2020).

Resilience is also conceptualized as both a characteristic of an individual and a process an individual engages with in order to recover from stressful or traumatic events. As a process, someone is able to acknowledge the adversity they are facing, and either endure it without suffering, being unaffected by the adversity, or they are affected but able to either "bounce-back" to pre-adversity levels of functioning or exhibit post-traumatic growth to return to higher than pre-adversity levels of functioning (Ayed, Toner, and Priebe, 2019; Bonanno et al., 2015; Crowe et al., 2016). Characteristics of resilience include a set of attributes that one uses to protect oneself from harmful outcomes, such as human, hope, and motivation. The resilience that comes from one's surroundings and social context can also be accessed as needed (Ayed, Toner, and Priebe, 2019). Grafton et. al. (2019) identified what they call the *3 R's* of resilience: resistance, recovery, and robustness. Resilience measurements have been designed to assess those attributes to explain different internal sources of resilience (Asheim et al., 2020).

Measuring resilience

Many measures of resilience are self-reported scales in which an individual is gauging how resilient they believe they would be in a specific scenario. This has led to criticism that such measures are actually assessing coping self-efficacy, something that is a direct predictor of resilience, but not resilience itself (Denckla et al., 2020; Wu et al., 2013). Resilience may be less of a stable trait and more an outcome of internal processes that occur as a response to stress. “If we believe resilience is a process rather than a trait, then to assess the presence of resilience, you have to wait for some kind of stressful experience to happen and then see how people respond; only then can a researcher ascertain if the overall process indicates resilience or not” (Denckla et. al. 2020, pg. 4).

Measurement of resilience is difficult to do without touching on one’s beliefs about their own ability, largely due to the fact that researchers cannot ethically induce trauma. Much of the early psychological work on measuring resilience examines children who were traumatized by abuse or by war, and who later return (or do not return) to pre-trauma levels of functioning (Luthar & Zigler, 1991). In order to expand research on resilience, many researchers began utilizing measures of life stress and adjustment to life stress, rather than trauma specifically. (Luthar & Zigler, 1991). Lazarus (1984) argued that it is better to look at everyday life “hassles” rather than traumatic events as indicators of stress, given that they are more frequent and less heterogeneous. Therefore, resilience is primarily measured by looking at the determinants of resilience (Asheim, 2020).

Connor and Davidson (2003) conceptualized resilience as a measure of stress coping ability, rather than reaction and recovery from traumatic events, and developed a measure to assess resilience for the purpose of treatment in anxiety and depression. The Connor-Davidson Resilience scale (CD-RISC-25) in its original form contains 25 items, and across multiple groups, the researchers found that higher scores on this measure were associated with greater improvement in response to treatment. An exploratory factor analysis (EFA) was performed by Connor and Davidson (2003) who identified five factors: (1) personal competence, high standards, and tenacity, (2) trusting one’s instincts, tolerance of negative affect, and strengthening effects of stress, (3) positive acceptance of change, (4) control, and (5) spiritual influences. The sample that was used to derive the original CD-RISC was 77% White and had a mean age of 43.8 years (Connor & Davidson, 2003), warranting questions about its validity when applied to undergraduate students from underrepresented populations. Connor and Davidson (2003) could not effectively speak to the veracity of their measure about specific racial minorities, as they grouped all other groups as "non-white".

A later EFA of the CD-RISC-25 by Campbell-Sills and Stein (2007) found that the measure had an unstable factor structure across samples with the same demographic make-up. In contrast with Connor and Davidson (2003), their sample was closer in age to the undergraduate population served by the research training programs of interest to this study. However, more than two-thirds of their sample was White, with the second largest group being Hispanic, constituting only 13.4% (Campbell-Sills & Stein, 2007). The researchers were able to identify 10 questions (items 1, 4, 6, 7, 8, 11, 14, 16, 17, and 19 from the original scale) that were not only psychometrically stable and contained minimal redundancy but also strongly positively correlated with outcomes from the original 25-item measure (r= .92). Further exploration of this 10-item measure (CD-RISC-10) identified a factor structure containing two factors the researchers labeled *hardiness* and *persistence*. However, *persistence* was determined by a single item on the measure, leading many other researchers to interpret the CD-RISC-10 as a single-factor measure of resilience (Davidson, 2020).

In addition to the 25 and 10-item versions, a highly abridged 2-item version (CD-RISC-2) was developed for use alongside other measures to predict recovery time for PTSD patients (Davidson, 2020). Comparisons of the three versions of the CD-RISC lend theoretical support for the 10-item version. One study found that between the 25, 10, and 2-item versions of the measure, only the CD-RISC-10 demonstrated convergent and divergent validity (Kuiper et al., 2019). Another study examining resilience in competitive athletes found that the 10-item version was psychometrically superior to the 25-item version and more useful than the 2-item version (Gonzales et al., 2016). The CD-RISC 10 has since been validated across populations of diverse individuals and in multiple languages (Davidson, 2020; Wolly & Jacobs, 2023). However, the acceptance of a unifactorial model has been called into question. In addition to the two-factor structure first identified by Campbell-Sills and Stein (2007), Aloba et al. (2016) found improved model fit and factor structure stability using a 2-factor hierarchical solution with *resilience* as its first-order variable composed of *toughness* and *motivation*, with factor loadings more evenly distributed than those found by Campbell-Sills and Stein (2007). This interpretation of the CD-RISC-10 is supported by subsequent research; however, it has not yet been approved in the official CD-RISC Manual, leaving the need for further testing (Davidson, 2020; Smith et al., 2018).

This study provides the opportunity to evaluate the CD-RISC-10 in a population that is markedly different than the samples used to derive both the original CD-RISC and the CD-RISC-10. Specifically, this study addresses what the structure of these measures is when completed by a sample primarily composed of Hispanic women.

Factor analysis and SEMs

Confirmatory factor analysis (CFA) is a type of SEM analysis that allows researchers to examine the relationship between observed and latent variables that cannot be directly measured (Jöreskog, 1964). CFA tests how well data fits an existing or proposed model with previously dictated pathways. In psychological tests, the goal of the test is often to measure a variable that cannot be directly observed (i.e., a latent variable) by examining several variables that can be observed. The strength of the relationship between the latent variable and the observed variable is expressed as a factor loading value between -1.0 and 1.0 and is interpreted similarly to a Pearson correlation coefficient or regression coefficient. The square of these factor loadings can be interpreted as percentages of test variance (Cronbach, 1971). However, highly correlated variables have been shown to yield standardized factor loadings lower than -1.0 and greater than 1.0 (Brown, 2015; Devlieger & Rosseel, 2023). Confirmatory factor analysis, as an aspect of structural equation modeling (SEM), also yields statistics that indicate the fit of the proposed model with a hypothetical model that perfectly matches the observed data. (Brown & Moore, 2012).

While CFA can be used to confirm basic SEMs composed of well-understood latent variables, the fact that the cross-loading of items is constrained to 0 in this analysis limits its ability to adequately describe more complex attributes. Many academics argue that over-reliance and over-use of CFA in modeling motivation is to blame for the issues of replicability and reliability in the study of motivation (Marsh et al., 2011). Large-scale analysis of published papers also suggests that methodological robustness, including revising previous models, is key in creating replicable data (Youyou, Yang, and Uzzi, 2023).

Exploratory factor analysis (EFA) is similar to CFA. Both allow researchers to examine interactions between latent and observed variables. However, the key distinction is the constraints put on CFA. Items in a CFA can only map onto the latent variables to which they are already theoretically connected. However, an EFA allows items to cross-load, and can assess relationships between items that may have been overlooked in previous structural equation models. Typically, an EFA will be performed to determine the suggested structure of a construct, and then that structure can be tested using CFA.

The decision about whether to use an EFA or a CFA is primarily rooted in the researcher's degree of certainty about the model they are examining (Gorsuch, 2014). If there is a robust and trusted theory regarding the structure, one might elect to utilize a CFA (Mulaik, 2014). However, one must also remember to be critical of existing models. Even when theory provides a model, its universal applicability is far from guaranteed. Often, an EFA will be used first to establish what items load heavily on what factors, before revising the model constraints and rerunning a CFA on the revised model (Gorsuch, 2024). However, EFAs capture more complexity than rigid SEMs produced by CFAs.

Exploratory structural equation modeling (ESEM) integrates the strengths of EFA into a structural equation model in which the confirmatory checks of CFA can still occur, but without the need for factor loadings to be constrained (Asparouhov & Muthén, 2009; Marsh et. al., 2009). Allowing factors to cross-load is particularly useful when variables in the model may have more than one source of variance (Morin et. al., 2013). This flexibility is important when examining psychological constructs that are not easily measured or structured. Using ESEM, a researcher can examine cross-loading amongst conceptually connected variables (as one would with an EFA) but also be able to produce goodness of fit statistics, incorporate covariates, and look at potentially correlated residuals (Alamer, 2022; Marsh & Alamer, 2024).

Typically, items that correlate strongly with the latent variable intended to be measured, but not other latent variables, are retained for use on the measure. This gives the interpreter a sense of the criterion validity of a model. Suppose items do not correlate with their intended latent variables or correlate highly with another. In that case, it may indicate that the content of those items should be examined. High inter-item correlations often reflect inadequate construct sampling (Clark & Watson, 1995). This means items may not represent the full range of complexity within the measured construct. However, there are times when highly correlated items are retained due to distinctions in the utilization (Brown & Moore, 2012; Cronbach, 1971). For example, working memory, attention, and information processing speed are often highly correlated and likely share variance between related factors (Friedman & Miyake, 2017; Kail & Hall, 1994; Salthouse 1996). However, each of those attributes is interpreted in distinct ways. Difficulties with working memory may indicate a need for more scaffolded learning, attention issues may indicate a need for behavioral or pharmaceutical intervention, and slow processing speed may necessitate more time for learning (Alloway & Alloway, 2010). Including overlapping items can be justified if they contribute to different interpretations and decision-making aspects.

While stronger factor loadings are typically the justification for accepting a model, a proposed use of factor analysis is to examine what *does not* correlate with the latent variables. Chamberlain (1965) and Cronbach (1971) suggest that factor analysis should be used to test a counterhypothesis rather than a hypothesis. If no highly correlated factors of a measure are applied to the counterhypothesis, there is more support for the original hypothesis.

CHAPTER III

Methods

Procedure

The data being examined in this study came from eight research training programs at the University of Arizona that target undergraduate students from underrepresented populations. Potential participants are identified through multiple forms of outreach, including web-based advertisements, outreach through university listservs, open information sessions hosted by multiple colleges, and direct outreach to undergraduate classes. Interested people submitted general applications, which were sorted and selected to be interviewed by representatives of whichever training program most matched their research interests. The data being evaluated for this work was collected between 2016 and 2025 as part of a longitudinal evaluation of the effectiveness and outcomes of these programs. The data was taken at the pre-intervention time point and represents the state of the participants' research self-efficacy and resilience before starting their research training program.

Fellows were contacted via email and prompted to complete a series of surveys using the Research Electronic Data Capture (REDCap) web-based platform. Participants received multiple reminders to complete the surveys, but completion was not compulsory. There were no negative consequences for the fellows who chose not to respond. As the measures were voluntary, not all participants completed all measures. Therefore, the number of participants is not consistent across measures. Additionally, not all respondents provided demographic information at the first time point, as demographics were gathered through a separate voluntary survey. Of the 200 participants who responded to the CD-RISC, 92 volunteered demographic information. Of the 173 participants whose data from the RSES is being used, 98 provided demographic information.

Participants

These research training programs specifically target students who are underrepresented in their fields and/or the student body as a whole. As such, they differed in key ways from the total demographic make-up of the university. Respondents for both measures were primarily Hispanic, female, and first-generation college students (Table 1). Compared to the most recent enrollment demographics of the University of Arizona undergraduates, this sample had nearly double the proportion of Black students as well as American Indian or Alaskan Native students, in addition to more than double the proportion of Hispanic students (University Analytics & Institutional Research, 2024).

Table 1: Demographic information (%)

	CD-RISC	RSES
Race/Ethnicity		
Hispanic	59.8	60.2
White	17.4	16.3
Black	13.0	14.3
American Indian/Alaskan Native	6.5	6.1
Asian	6.5	6.1
Hawaiian/Pacific Islander	1.1	1.0
Gender Identity		
Female	71.7	73.5
Male	25.0	23.5
Nonbinary	3.3	3.1
First Generation	54.4	56.12

Note: RSES (*n*=98), CD-RISC (*n*=92)

Participant responses were provided for the present study as secondary data for psychometric analysis. This work was determined to be *not human subjects research* by a local-site institutional review board.

Intervention

The training programs from which this sample comes largely focus on apprenticeship and practical experience. Fellows participate in courses designed to teach them the basics of research design and manuscript preparation. They also attend multiple workshops that provide professional development around presenting research and applying to graduate programs. The central component of these programs is the apprenticeship aspect, in which fellows are paired with a research mentor who holds a PhD and actively performs research. The fellows assist in ongoing research and design and execute their own study under the guidance of a more senior researcher. At the conclusion of their program, the fellows present their findings as a traditional oral presentation and during a poster session. Fellows are subsequently encouraged to find mechanisms for scientific and non-scientific dissemination of their findings.

Instruments

Connor-Davidson Resilience scale - 10 Item (CD-RISC-10)

The CD-RISC-10 (Campbell-Sills & Stein, 2007; Davidson, 2020) was used to assess the baseline resilience of the participants. This version of the CD-RISC retains 10 of the original 25 items and asks participants, rating on a 5-point Likert scale (0-4), to consider how a series of statements applies to any stressful situations they encountered over the last month. If no stressful situations occurred within the last month, participants are asked to respond according to how they believe they would have felt. Participants rate the statements as either not true at all (0), rarely true (1), sometimes true (2), often true (3), and true nearly all the time (4). Each statement addresses aspects of responses to hardship that are theoretical determinants of resilience, such as “having to cope with stress can make me stronger” and “I am not easily discouraged by failure”. (Appendix A).

Research Self-Efficacy Scale (RSES)

A version of the 53-item Research Self-Efficacy Scale (RSES; Greeley et al.,1989; Bieschke et al., 1996) was adapted and modified by the evaluation team for the research training programs to address specific domains of research self-efficacy that are most relevant to the programs' intended outcomes. The original RSES used 53 items to examine research self-efficacy through four latent variables representing different stages of the research process: (1) early tasks, (2) conceptualization, (3) implementation, and (4) presenting findings. Of those factors, Bieschke et al. (1996) determined that only early tasks, conceptualization, and implementation accounted for unique variance in whether someone is interested in pursuing research as a career. Because motivating interest in a research career is an expressed goal of these training programs and because presentation efficacy was determined through practical assessments, the team that initially collected this data utilized only questions from those three subscales. Removing the *presenting findings* items from the measure produced an abridged 38-item measure.

The modified RSES provided participants with a list of behaviors required at different stages in the research process, asking them to rate on a scale of 0-100 (no confidence - complete confidence) their degree of confidence in their ability to successfully perform the behavior. These items included the ability to "follow ethical principles of research", "synthesize current literature", "choose appropriate research design", and "organize collected data for analysis" (Appendix B).

Data Analysis

Data from the CD-RISC-10 and the modified RSES were analyzed using a combination of SPSS 29 (IBM Corp, 2022) and Microsoft Excel (Microsoft Corporation, 2021). Structural equation models and model fit indices were produced using MPlus version 8.8 (Muthén & Muthén, 2022). Both measures were evaluated for normality at the item level and scale level. For the CD-RISC, the scale level analysis used sum scores from the CD-RISC-10 as a composite describing resilience (Davidson, 2020). For the modified RSES, the distribution of the three selected sub-scores described by Bieschke et. al. (1996) was examined. Skewness and kurtosis at both the item and scale levels were largely within acceptable ranges for the CD-RISC-10 and the modified RSES (Kline, 2023). However, two outliers were identified in the RSES data, with extremely low scores more than three standard deviations from the mean. Additionally, the time it took these two participants to complete the measure was substantially low, indicating they read and responded to items at a rate of less than two seconds per item. While it is not impossible that they were able to process information rapidly, brevity combined with uniform responses suggests the responses were not legitimate. Those two participants were removed from the RSES data (n=173). While those students are of interest at a programmatic level, it is necessary to remove them from the data set when evaluating the psychometric properties of the scale. Internal reliability at a scale level for both measures was determined using *Cronbach's alpha* (α), which examines the proportion of total variance that reflects shared variance among items. Complete descriptive statistics for both the items and the scale/subscales for the CD-RISC-10 and modified RSES are provided in the next chapter.

To determine if the single-factor structure of the CD-RISC-10 fit this data, an initial confirmatory factor analysis (CFA) was performed, designating all items to load onto the single factor of *resilience*. Following the CFA, the potential for a 2-factor model as described by Aloba et. al. (2016) and Smith et. al. (2018) was explored, first utilizing CFA along the pathways described in previous research and subsequently using exploratory structural equation modeling (ESEM) both to examine the model fit of the 2-factor model and to examine the potential for cross-loading between items.

The modified RSES (*n*=173) underwent a similar process. First, CFA was used to examine the model fit of the theoretical model that was assumed when the data was first gathered, followed by an ESEM to determine where instabilities in the factor structure may arise. Adjustments were made to the measure to examine potentially more stable factor structures, and further ESEM was used to examine the revised model as well.

Following the individual assessment of measures, an ESEM containing all latent variables and items from both the CD-RISC-10 and modified RESES was performed to examine the potential for cross-loading between items on a measure that evaluates two constructs that are theoretically closely related.

As a gauge of model fit, multiple statistics were used. First, Chi-Squared (χ^2) was used as a general indicator of the difference between the observed covariance matrix and the model-predicted one. Large and statistically significant values for χ^2 indicate poor model fit. However, the test used to determine this statistic assumes large sample sizes. This can lead the test to amplify small differences and may lead to the false rejection of a model. Therefore, additional statistics are necessary for making a decision about model fit (Kline, 2023).

The Comparative Fit Index (*CFI*) utilizes the χ^2 but adjusts for model complexity and sample size. CFI is calculated by subtracting degrees of freedom (*df*) from χ^2 values for both the target and null models that assume unrelated variables. The value from the target model is divided by the values for the null model and the result is subtracted from 1. CFI values closer to 1 indicate better model fit. Generally, values greater than .9 indicate acceptable model fit (Hu & Bentler, 1999).

The Tucker-Lewis Index (*TLI*) also compares an existing model to a null model, but favors less complex models and provides a more conservative estimate of model fit than the CFI while being less sensitive to sample size than χ^2. Similar to CFI, the TLI statistic is interpreted as indicating better model fit the closer it is to 1, and values greater than .9 indicate acceptable model fit (Hu & Bentler, 1999).

Root Mean Square Error of Approximation (*RMSEA*) was also used as an indicator of model fit. RMSEA is able to adjust for sample size as well as model complexity and provides an estimate of how well the model would fit the population per degree of freedom. This statistic represents the square root of the difference between the model's χ^2 and degrees of freedom, scaled by sample size. Lower values for RMSEA indicate better model fit, with values between .05 and .08 indicating acceptable model fit. However, RMSEA values less than .05 are considered the cut-off for close fit (Hu & Bentler, 1999).

The final statistic used in this study to evaluate model fit is the Standardized Root Mean Square Residual (*SRMR*). Unlike other statistics used for determining model fit, the SRMR compares standardized observed covariances between two variables with expected covariances, producing the average standardized difference between the covariances predicted by the model and the covariances observed in the data. SRMR, unlike RMSEA, CFI, TLI, and χ^2, is less sensitive to sample size and model complexity. Like RMSEA, values less than .08 indicate acceptable model fit, while values less than .05 indicate ideal model fit.

Each of these statistics is calculated for the existing and explored models for the CD-RISC-10 and modified RSES and is interpreted collectively to provide a comprehensive understanding of how the proposed models fit the data being analyzed.

CHAPTER IV

Results

Descriptive Statistics

CD-RISC-10

Descriptive statistics for the 10-item version of the CD-RISC are reported in Table 1. All items were checked for normal distribution by evaluating skewness and kurtosis. Values for skewness ranged from -.33 (item 8) at the highest and -1.07 (items 6 and 9) at the lowest. Kurtosis values ranged from .45 (item 9) to -.74 (item 8). All scores for skewness and kurtosis fall within acceptable ranges for univariate normality (Kline, 2023). Mean scores for each item ranged from 3.50 (*SD*=.66) to 2.67 (*SD*=1.04) while the scale mean was 31.1 (*SD*=5.83), indicating that the sample overall displayed moderately high levels of resilience. Internal consistency was high (α=.88), confirmed by moderately strong positive but statistically significant inter-item correlations (Table 3). Most inter-item correlations were significant at the $p<.001$ level. Correlations between items 3 and 8 (r=.23) as well as between 3 and 10 (*r=.21)* were significant at the $p<.01$ level. The correlation between items 3 and 7 was a weak positive correlation (r=.16) was significant at the $p<.05$ level.

Table 2: Descriptive Statistics for CD-RISC-10

Item	*m*	*SE*	*SD*	Skewness	Kurtosis
cdrisc1	3.39	.05	.64	-.57	-.62
cdrisc2	3.18	.05	.76	-.52	-.46
cdrisc3	3.06	.06	.90	-.58	-.41
cdrisc4	3.01	.06	.88	-.60	.08
cdrisc5	3.14	.06	.80	-.55	-.44
cdrisc6	3.50	.05	.66	-1.07	.40
cdrisc7	2.88	.06	.88	-.34	-.45
cdrisc8	2.67	.07	1.04	-.33	-.74
cdrisc9	3.23	.06	.92	-1.07	.45
cdrisc10	3.06	.06	.85	-.66	.11
Total	31.1	.41	5.83	-.38	-.41

Note: *n*=200

Table 3: Inter-item correlation matrix for CD-RISC-10

Item	1	2	3	4	5	6	7	8	9	10
1	1.0									
2	.64***	1.0								
3	.39***	.34***	1.0							
4	.41***	.44***	.38***	1.0						
5	.51***	.50***	.30***	.51***	1.0					
6	.49***	.51***	.30***	.36***	.48***	1.0				
7	.42***	.46***	.16*	.43***	.41***	.41***	1.0			
8	.47***	.48***	.23**	.45***	.55***	.42***	.54***	1.0		
9	.48***	.55***	.29***	.50***	.44***	.54***	.46***	.64***	1.0	
10	.43***	.42***	.21**	.40***	.41***	.37***	.30***	.51***	.53***	1.0

Note: n=200, all values represent Pearson correlations (r), *p<.05 **p<.01, ***p<.001

RSES

Descriptive statistics for the modified RSES are presented in Table 3. Measures of skewness ranged from .02 (item 38) to -1.83 (item 2), showing a slightly negative skew across all items except item 38. Values for kurtosis were more variable, ranging from 4.48 (item 9) to -1.00 (item 38). Items 2, 4, 9, and 21 were more platykurtic than all other variables in the data set. Despite the higher-than-desirable levels of kurtosis for these items, they do fall within the acceptable range for the purposes of factor analysis utilizing robust estimators (Kline, 2023). Mean scores for the RSES items ranged from 92.49 (*SD*=11.42) to 44.97 (*SD*=3.03), showing a large amount of variance across the items on this scale. As the scale is intended to capture distinct domains of research tasks, that is not particularly surprising. However, a further examination of this variability is included in the *Discussion* section. Scores of skewness and kurtosis for the three subscales found by Bieshcke et. al. (1996) to account for the most unique variance in whether participants pursuing future research were all within the -1/+1 range, meeting assumptions for univariate normality. Mean scores for those values indicate greater efficacy for the tasks involved in the early stages of research (*m*=81.58, *SD*=12.78) with progressively lower efficacy and higher variance for the conceptualization (*m*=74.35, *SD*=14.43) and implementation (*m*=65.58, *SD*=2.11) stages.

Table 4: Descriptive Statistics for RSES

Variable	*m*	*SE*	*SD*	Skewness	Kurtosis
rses1	81.62	1.30	17.11	-1.07	1.09
rses2	92.49	.87	11.42	-1.83	3.36
rses3	82.25	1.22	16.07	-1.06	1.19
rses4	85.14	1.27	16.66	-1.40	3.11
rses5	75.84	1.54	2.26	-.79	.59
rses6	72.20	1.59	2.93	-.59	.25
rses7	70.69	1.70	22.40	-.66	.19
rses8	79.19	1.41	18.60	-1.16	1.87
rses9	82.20	1.43	18.83	-1.74	4.48
rses10	84.39	1.14	14.99	-.77	.05
rses11	80.69	1.33	17.51	-1.01	1.33
rses12	64.80	1.72	22.69	-.61	-.06
rses13	64.22	1.76	23.13	-.66	.13
rses14	73.35	1.43	18.87	-1.18	2.71
rses15	73.47	1.43	18.76	-.59	.04
rses16	75.20	1.42	18.70	-.53	-.15
rses17	75.61	1.32	17.36	-.64	.18
rses18	74.86	1.35	17.8	-.79	1.03
rses19	75.32	1.43	18.79	-.93	1.50
rses20	71.62	1.65	21.67	-.81	.46
rses21	81.79	1.29	16.90	-1.34	2.90
rses22	69.02	1.62	21.29	-.86	.77
rses23	66.76	1.63	21.38	-.71	.86
rses24	75.43	1.44	18.94	-.57	-.21
rses25	70.00	1.72	22.64	-.89	.91
rses26	63.99	1.74	22.87	-.58	.33
rses27	58.79	2.30	3.26	-.44	-.63
rses28	62.72	2.16	28.43	-.63	-.32
rses29	57.23	2.28	3.05	-.41	-.78
rses30	65.26	2.04	26.88	-.84	.16
rses31	64.10	2.0	26.30	-.73	-.03
rses32	61.50	2.29	3.16	-.64	-.59
rses33	68.38	2.01	26.47	-.96	.38
rses34	72.49	1.85	24.31	-1.14	1.23
rses35	76.53	1.78	23.44	-1.22	1.39
rses36	67.98	1.97	25.90	-.77	.10
rses37	66.36	2.00	26.26	-.87	.31
rses38	44.97	2.28	3.03	.02	-1.00
Concept	74.35	1.10	14.43	-.44	-.05
Implement	65.58	1.53	2.11	-.57	-.22
Early Task	81.58	.97	12.78	-.57	.11

Note: n = 173

Internal reliability as measured by *Cronbach's alpha* was extraordinarily high (α=.97). Scores this high may not indicate internal reliability as much as they are indicative of redundancy in the items. Too many items asking the same question in ways that are not substantially different from one another can produce high inter-item correlations that inflate *alpha* scores (Kline, 2023; Streiner, 2003). The 15 largest and 15 smallest inter-item correlations are shown in Tables 4 and 5, respectively. As shown in Table 4, there are multiple strong positive correlations between multiple items. Correlations such as that between items 12 and 13 (r=.94, p<.001), 23 and 23 (r=.89, p<.001), and 29 and 32 (r=.88, p<.001) may indicate redundancy rather than reliability. While most items correlated significantly at the p<.05 level, there were five combinations that failed to correlate significantly. Item 38 was not meaningfully correlated with items 2, 3, or 4. Item 36 was not significantly correlated with Item 2, and Item 20 was not significantly correlated with Item 2. The lack of significant correlation here may be an indicator that those questions are addressing fundamentally different latent variables within research self-efficacy.

Table 5: Top 15 Inter-Item Correlations for modified Research Self-Efficacy Scale

Item 1	Item 2	Correlation (*r*)
13	12	.94***
23	22	.89***
32	29	.88***
34	33	.87***
28	27	.85***
31	30	.83***
33	32	.82***
30	29	.82***
33	31	.81***
32	31	.79***
37	36	.78***
31	28	.77***
29	28	.77***
32	30	.77***
33	30	.77***

Note: n = 173. Significance levels: *p < .05, **p < .01, ***p < .001.

Table 6: Bottom 15 Inter-Item Correlations for RSES

Item 1	Item 2	Correlation (*r*)
8	2	.20**
29	4	.20**
37	2	.20**
32	4	.20**
38	17	.20**
35	5	.20**
34	4	.19*
36	17	.19*
19	2	.19*
36	5	.17*
38	4	.15
38	3	.15
36	2	.15
38	2	.14
20	2	.13

Note: n = 173. Significance levels: *p < .05, **p < .01, ***p < .001.

CD-RISC-10: CFA

To begin the examination of the CD-RISC-10 factor structure, an initial CFA was performed assuming a unifactorial model, treating each item as a component of general resilience (Figure 1). Factor loadings ranged from .42 to .76 and were all significant at the p<.001 level. Item 3 had the smallest factor loading and explained the smallest, however still statistically significant, proportion of variance in overall resilience (R^2=.18, p < .01). Table 6 reports all standardized factor loadings and their R^2 values. Multiple statistics were used to examine model fit. A *chi-square* test initially indicated poor model fit (χ^2= 85.8, *df*=35, p<.05). RMSEA was closer to values indicating fit, but was also outside of the range considered acceptable (*RMSEA*=.085). However, the statistics used for model fit all indicate acceptable levels of model fit when compared to a null model (*CFI*=.94, *TLI*=.92, *SRMR*=.04). This indicates that the model is not perfect, with small discrepancies being flagged by the *chi-square* and *RMSEA* tests, however it explains the data far better than a model in which the variables were not related at all.

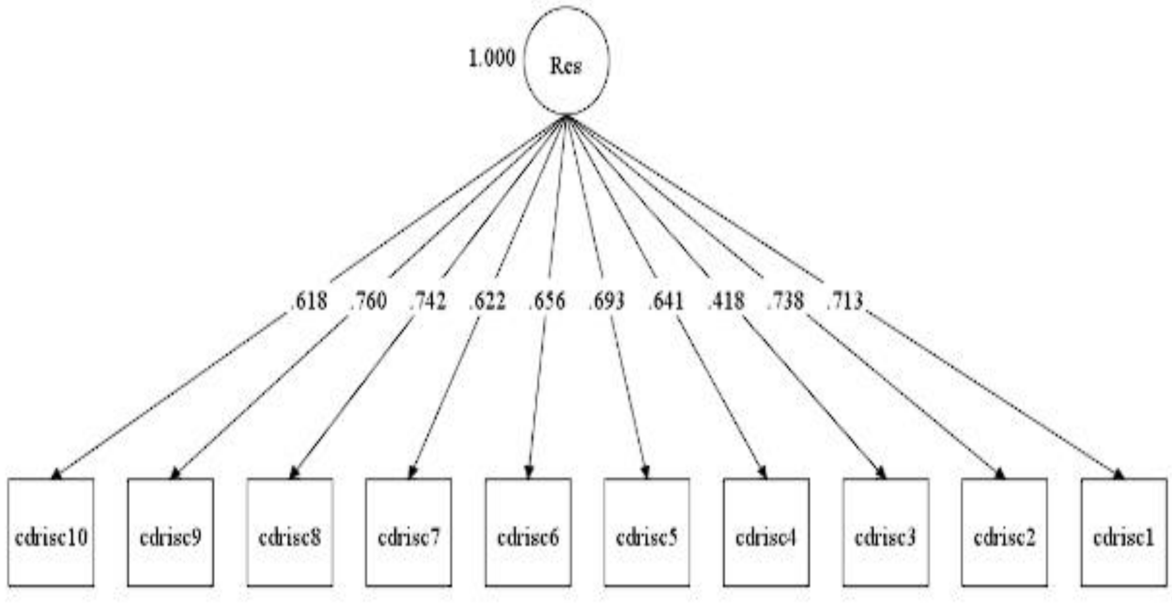
Table 7: Single Factor CFA Standardized Factor Loading and R^2 Estimates

Item	λ	*SE*	R^2
1	.71***	.04	.51***
2	.74***	.04	.55***
3	.42***	.06	.18**
4	.64***	.05	.41***
5	.69***	.04	.48***
6	.66***	.05	.43***
7	.62***	.05	.39***
8	.74***	.04	.55***
9	.76***	.04	.58***
10	.62***	.05	.38***

Note: *n*=200, ***p*<.01, ****p*<.001

Figure 1:

CD-RISC-10 Single Factor CFA loading



CD-RISC-10: 2-factor CFA

In addition to examining the single-factor version endorsed by Davidson (2020) and others, a CFA was performed assuming two factors. The original work that produced the CD-RISC-10 indicated two factors, *persistence* and *hardiness*. However, the *persistence* factor was determined by a single question (Campbell-Sills & Stein, 2007). The 2-factor model tested in this study is the structure reported by Aloba et. al. (2016) and Smith et. al. (2018). This model identifies *resilience* as a second-order factor determined by the first-order factors *toughness* and *motivation* (Figure 2). For *toughness*, standardized factor loadings ranged from .44 (item 3) to .78 (item 2). *Motivation* factor loadings ranged from .63 (items 7 and 8) to .79 (item 9). All loadings were significant at the p<.001 level. As with the previous model, item 3 loads weakly and accounts for the least variance (λ=*.44,* R^2=.19, p < .01). Table 7 reports all standardized factor loadings and their R^2 values for this measure. As expected, a *chi-square* test indicated poor model fit (χ^2=71, *df*=33, p<*.05*). However, his structure showed slightly improved model fit compared to the unifactorial model under statistics less sensitive to size and complexity (*RMSEA*=.076, *CFI*=.95, *TLI*=.94, *SRMR*=.04).

Figure 2.
CD-RISC 2-Factor CFA

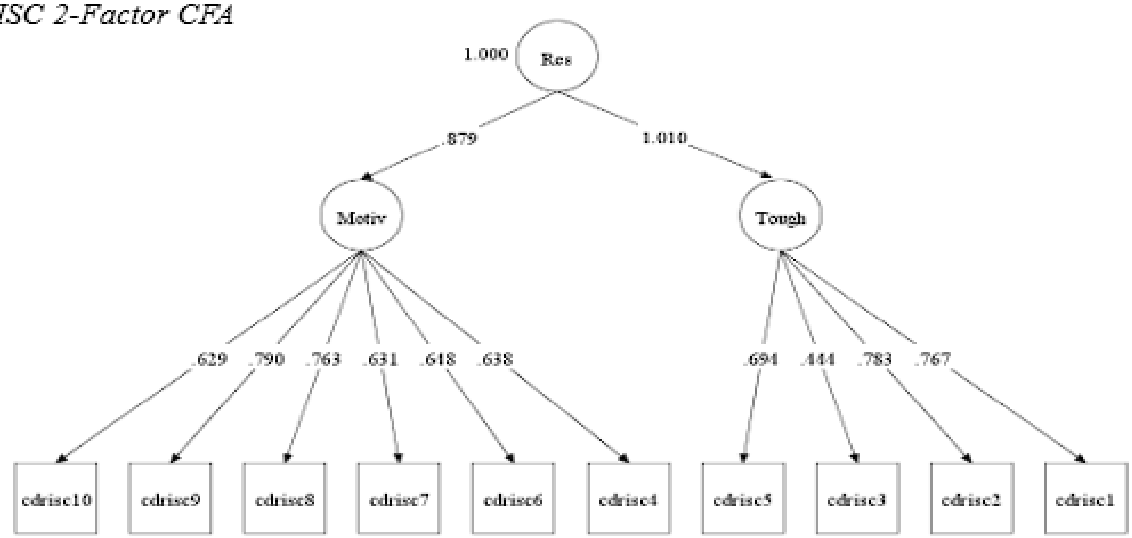


Table 8: 2-Factor CFA Standardized Factor Loading and R^2 Estimates

Subscale	Item	λ	*SE*	R^2
Toughness				
	1	.77***	.04	.59***
	2	.78***	.03	.61***
	3	.44***	.06	.19***
	5	.69***	.04	.48***
Motivation				
	4	.64***	.05	.41***
	6	.65***	.05	.42***
	7	.63***	.05	.40***
	8	.77***	.04	.58***
	9	.79***	.03	.63***
	10	.63***	.05	.40***
Resilience				
	Toughness	1.01	17.40	
	Motivation	.88	19.99	

Note: n=200 , ***p<.001

CD-RISC-10: 2-Factor ESEM

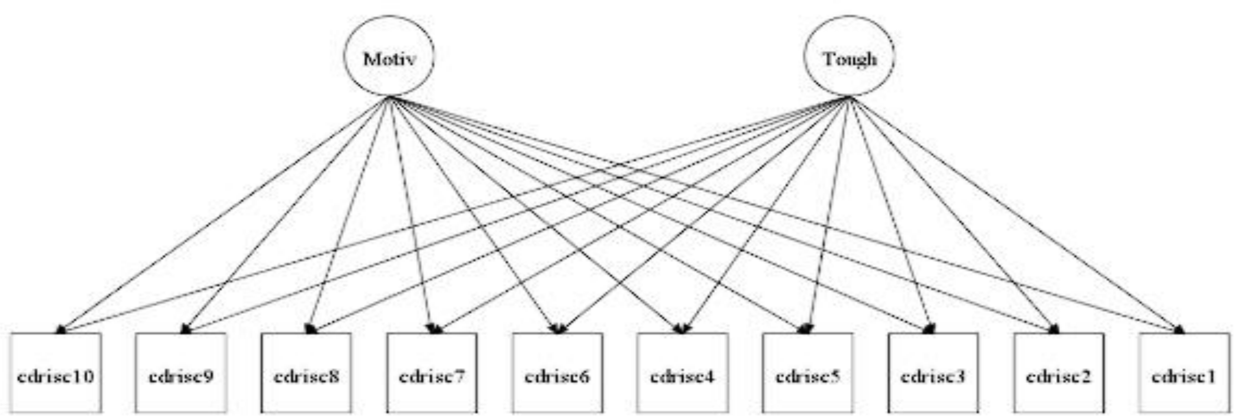
After testing two existing conceptual models for resilience as assessed by the CD-RISC-10 with results mixed on model fit, exploratory structural equation modeling was used to determine possible interactions between variables that are not accounted for in those existing models. This model designated theoretical pathways based on the 2-factor model but allowed for cross loading between items that determine the first-order latent variables (Figure 3). Interestingly, all items, not just the items designated for the latent variable in the 2-factor model, loaded significantly on the *toughness* variable (Table 9.) On the other hand, only items 7-9 loaded significantly on the *motivation* variable. Further, although items 7 (λ=.37, p<.05) and 9 (λ=.45, p<.05) loaded onto *motivation*, the strength and statistical significance of their loading onto *toughness* was greater. Of the items theoretically connected to *motivation* only item 8 (λ=.67, p<.001) acted as expected when not constrained. Despite those factor loadings, the two-factor model that allows for cross loading between items yielded improved model fit statistics (χ^2=49.4, p<.05; *RMSEA*=.067, *CFI*=.97, *TLI*=.95, *SRMR*=.03).

Table 9: *CD-RISC: Standardized Factor loadings for ESEM*

Subscale	Item	λ	*SE*
Toughness			
	1	.78***	.05
	2	.78***	.03
	3	.44***	.06
	4	.57***	.11
	5	.63***	.10
	6	.64***	.08
	7	.52***	.13
	8	.54***	.21
	9	.63***	.15
	10	.51***	.12
Motivation			
	1	.05	.24
	2	.09	.25
	3	-.04	.17
	4	.27	.21
	5	.29	.15
	6	.16	.22
	7	.37*	.17
	8	.67***	.17
	9	.45*	.23
	10	.35	.21

Note: n=200 ,*p<.05, ***p<.001

Figure 3
CD-RISC-10 2-Factor ESEM



Modified RSES: CFA

To begin an investigation of the modified RSES, an initial CFA utilizing the factor structure established by Bieschke et al. (1996) was performed (Figure 4). Table 9 presents factor loadings for items on their pre-designated pathways. Table 10 shows the strength of the relationships between latent variables. Standardized factor loadings for *early tasks* ranged from .4 (item 2, R^2=.16) to .81 (item 8, R^2=.66). For the *conceptualization* variable, loadings ranged from .62 (item 20,R^2=.42). For the *implementation* variable, standardized factor loadings ranged from .46 (item 9, R^2=.21). All factors loaded on their intended latent variables at p<.001. While all subscales correlated positively, the strength of the correlation between *early tasks* and *conceptualization* (r=.90, p<.001) draws into question how well this measure distinguishes between those stages of research. Despite items loading as intended, all statistics used for model fit indicated this data fit the described model poorly (χ^2=2478.1, p<.05; *RMSEA*=.131, *CFI*=.69, *TLI*=.67, *SRMR*=.09).

Figure 4
Factor Structure for modified RSES CFA

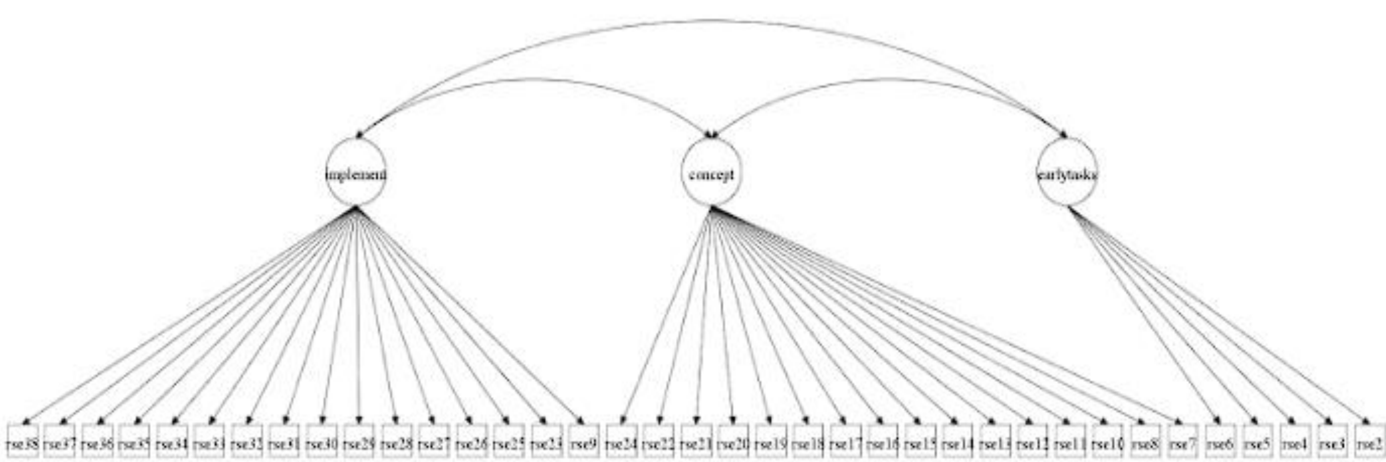


Table 10: 3-Factor CFA Standardized Factor Loading and R2 Estimates for RSES

Subscale	Item	λ	SE	R^2
Early Tasks				
	2	.40***	.07	.16***
	3	.81***	.03	.66***
	4	.71***	.04	.51***
	5	.64***	.05	.41***
	6	.65***	.05	.42***
Conceptualize				
	7	.64***	.05	.41***
	8	.75***	.04	.57***
	10	.67***	.04	.45***
	11	.68***	.04	.46***
	12	.78***	.03	.61***
	13	.77***	.03	.59***
	14	.80***	.03	.64***
	15	.80***	.03	.64***
	16	.86***	.02	.74***
	17	.71***	.04	.51***
	18	.78***	.03	.61***
	19	.64***	.05	.41***
	20	.62***	.05	.38***
	21	.65***	.05	.42***
	22	.71***	.04	.51***
	24	.79***	.03	.62***
Implementation				
	9	.46***	.06	.21***
	23	.70***	.04	.49***
	25	.69***	.04	.48***
	26	.73***	.04	.54***
	27	.80***	.03	.64***
	28	.84***	.02	.70***
	29	.86***	.02	.75***
	30	.86***	.02	.75***
	31	.89***	.02	.79***
	32	.87***	.02	.76***
	33	.89***	.02	.80***
	34	.84***	.02	.71***
	35	.56***	.05	.31***
	36	.61***	.05	.37***
	37	.73***	.04	.54***
	38	.63***	.05	.39***

Note: n=173, ***p<.001

Table 11: Correlation between subscales in modified RSES

	Early Tasks	Conceptualization	Implementation
Early Tasks	1.0		
Conceptualization	.90***	1.0	
Implementation	.57***	.72***	1.0

Note: *n*=173, ****p*<.001, values represent Pearson's *r*

Modified RSES: ESEM

To further examine the factor structure of this modified version of the RSES, a second analysis using ESEM was performed. Like the previous ESEM performed on the CD-RISC-10, this model was designed to designate pathways but not to constrain the ability of items to load onto latent variables other than the ones they are theoretically aligned with (Figure 5). For ease of reading, the factor loadings for each latent variable are reported in their own table. As shown in Table 11, the only item that retained its previously designated structure was item 6 (λ=.62, p<.001). In fact, item 6 was the only item to significantly load onto *early tasks* at all. Items 12, 13, and 34 showed values well outside standardized ranges (outside of -1 to 1), indicating that the model is likely not correctly specified and that the *early tasks* variable is poorly distinguished.

The *conceptualization* variable was also poorly defined (Table 13). Like the *early tasks* variable, only one item maintained its pathway, item 24, and managed to load significantly (λ=.62, p<.001). Items 12 and 13 again produced estimates far outside of what is normal for factor analysis, further suggesting model misspecification.

The *implementation* variable maintained much more of its structure than the previous latent variables (Table 13). Of the 16 items originally used to evaluate this stage, 12 of them retained significant loadings (p<.05). Item 9 (λ=.62, p=.49), item 25 (λ=.16, p=.65), item 34

(λ=.59, *p*=.17), and item 35 (λ=.24, *p*=.31) failed to load not only on their previously designated variable, but on any variable at all. Beyond the predetermined pathways examined in the CFA, allowing for cross-loading of items produced significant factor loadings from items intended for *early tasks* and *conceptualization* onto the *implementation* variable. From the *early tasks* side, item 3 (λ=-.52, p<.05) and item 4 (λ=-.33, p<.05) loaded significantly in a negative direction. This indicates that those questions may be more of a reverse-scored indication of efficacy in the implementation stage of research than they are a positively directed indicator of efficacy in the early stages. Item 22 (λ=.33, p<.05) seemed to be a slight indicator of efficacy at the *implementation stage* rather than the *conceptualization stage*, as previously indicated by the CFA. While inconsistent with the determined factor loadings, the *implementation* variable did not produce the same Heywood cases (loadings outside of -1 to 1) that were produced when examining the other two variables.

The statistics for model fit, while slightly improved relative to the rigid CFA, also indicated poor model fit (χ^2=1621.8, p<.05; *RMSEA*=.11, *CFI*=.77, *TLI*=.72, *SRMR*=.05).

Figure 5
Factor Structure for modified RSES ESEM

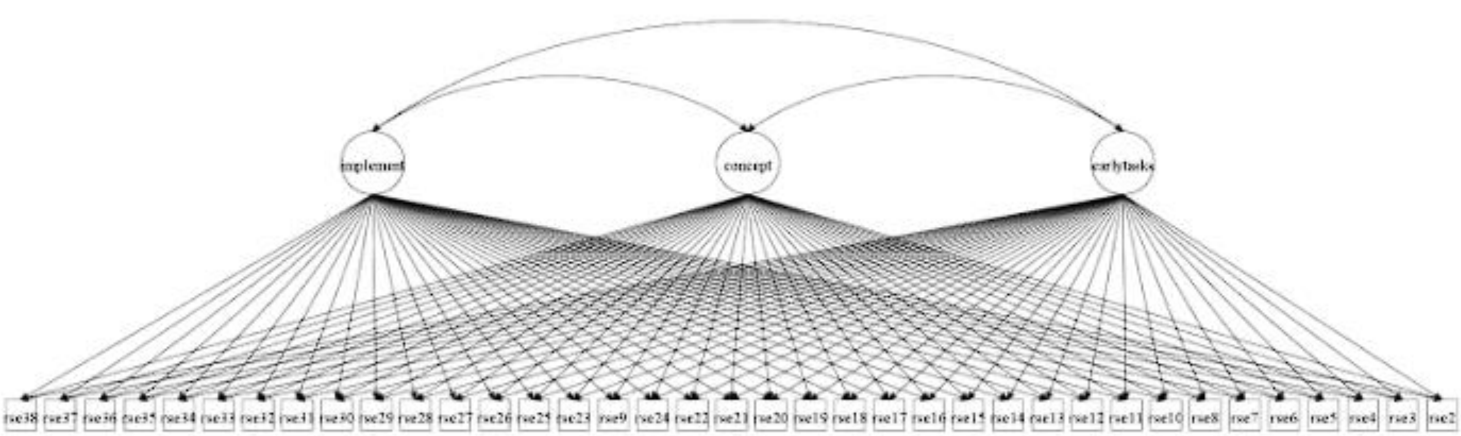


Table 12: Unconstrained Factor Loadings for Early Stages variable in RSES ESEM

Subscale	Item	λ	*SE*
Early Stages	RSE2	.30	.44
	RSE3	-.24	.74
	RSE4	.28	.51
	RSE5	.29	.43
	RSE6	.62***	.06
	RSE7	-.25	.72
	RSE8	-.69	1.08
	RSE9	-1.05	1.47
	RSE10	-.49	.91
	RSE11	.59	.75
	RSE12	3.33	3.12
	RSE13	3.64	3.32
	RSE14	.87	.74
	RSE15	.69	.73
	RSE16	.51	.52
	RSE17	.15	.49
	RSE18	.17	.59
	RSE19	.15	.57
	RSE20	-.01	.46
	RSE21	-.82	1.19
	RSE22	.08	.47
	RSE23	.38	.55
	RSE25	-.98	1.22
	RSE26	.02	.53
	RSE27	.35	.54
	RSE28	-.01	.66
	RSE29	.24	.65
	RSE30	-.24	.79
	RSE31	-.20	.77
	RSE32	-.14	.69
	RSE33	-.56	1.00
	RSE34	-1.36	1.53
	RSE35	-.46	.85
	RSE36	-.10	.49
	RSE37	-.65	.81

Note: n=173, ***p<.001,

Table 13:Unconstrained Factor Loadings for Conceptualization variable in RSES ESEM

Subscale	Item	λ	*SE*
Conceptualization	RSE2	.04	.50
	RSE3	1.38	.81
	RSE4	.56	.57
	RSE5	.26	.51
	RSE7	.88	.81
	RSE8	1.61	1.17
	RSE9	1.92	1.59
	RSE10	1.33	1.02
	RSE11	.08	.82
	RSE12	-2.92	3.36
	RSE13	-3.32	3.52
	RSE14	-.03	.80
	RSE15	.22	.80
	RSE16	.49	.58
	RSE17	.83	.55
	RSE18	.68	.71
	RSE19	.50	.66
	RSE20	.54	.56
	RSE21	1.51	1.32
	RSE22	.39	.51
	RSE23	-.03	.63
	RSE24	.78***	.03
	RSE25	1.56	1.32
	RSE26	.39	.60
	RSE27	-.44	.61
	RSE28	-.06	.73
	RSE29	-.33	.72
	RSE30	.13	.91
	RSE31	.13	.89
	RSE32	.08	.80
	RSE33	.64	1.15
	RSE34	1.56	1.65
	RSE35	.81	.86
	RSE36	.18	.55
	RSE37	.77	.86

*Note: n=173, ***p<.001*

Table 14: Unconstrained Factor Loadings for Implementation variable in RSES ESEM

Subscale	Item	*λ*	*SE*
Implementation	RSE2	.085	.16
	RSE3	-.52*	.22
	RSE4	-.33*	.16
	RSE5	.01	.17
	RSE7	.02	.22
	RSE8	-.21	.32
	RSE9	-.29	.42
	RSE10	1.33	1.02
	RSE12	.54	.93
	RSE13	.64	.99
	RSE14	-.07	.20
	RSE15	-.16	.20
	RSE16	-.19	.15
	RSE17	-.36	.15
	RSE18	-.06	.20
	RSE20	.09	.18
	RSE21	-.04	.33
	RSE22	.33*	.16
	RSE23	.47*	.19
	RSE25	.16	.35
	RSE26	.41*	.16
	RSE27	.90***	.16
	RSE28	.90***	.15
	RSE29	.95***	.16
	RSE30	.93***	.19
	RSE31	.94***	.19
	RSE32	.91***	.17
	RSE33	.79***	.27
	RSE34	.59	.43
	RSE35	.24	.24
	RSE36	.53***	.16
	RSE37	.59***	.22
	RSE38	.63***	.06

*Note: n=173, *p<.05, ***p<.001*

Distinguishing Resilience from Research Self-Efficacy

As self-efficacy and resilience are closely related constructs, and because the CD-RISC-10 asks questions primarily about how individuals *believe they would* respond to an event, it is important to determine whether the measures of self-efficacy and resilience used for evaluation of these programs distinguish themselves from one another while maintaining a positive correlation. To explore the possibility of significant overlap between the CD-RISC-10 and modified RSES, an exploratory structural equation model was constructed that incorporates the latent variables and items for each construct and allows them to freely interact.

Encouragingly, examining the measures this way produced no significant cross-loading between variables intended for *resilience* and its subscales with the subscales used to assess research self-efficacy. Correlations between resilience and the RSES latent variables were weak, positive, and statistically significant. To add support to their differentiation, all RSES subscales were more strongly correlated with each other than with resilience (Table 15).

Table 15: Correlation Matrix (*r*) for RSES subscales and Resilience total

	Resilience	Conceptualization	Implementation
Resilience	1		
Conceptualization	.30***	1	
Implementation	.30***	.74***	1
Early Tasks	.29***	.76***	.52***

Note: *n=173, ***p<.001*

Summary of Results

Resilience (CD-RISC-10)

Confirmatory factor analysis identified both the single-factor model and two-factor model as having acceptable levels of model fit as assessed by the statistics, less sensitive to sample size and complexity, with slightly improved model fit found for the 2-factor model. However, when items were permitted to cross-load in the ESEM, only the *toughness* variable was well defined. All items, including those theoretically bound to the *motivation* variable, loaded significantly on *toughness.* Most of the items that composed the *motivation* variable failed to load as predicted. Two of those that did, items 7 and 9, had stronger loadings on the *toughness* factor. Item 7 states, "Under pressure, I stay focused and think clearly", while item 9 is "I think of myself as a strong person when dealing with life's challenges and difficulties" (Appendix A). The ESEM indicates that while those items do speak to levels of *motivation* that influence resilience, they are more at home among questions like "I am able to adapt when changes occur" (item 1) and "I can deal with whatever comes my way" (item 2). While results like this are not uncommon when there is large overlap between first and second order constructs that are closely connected conceptually, the overloading of the *toughness* factor on resilience in the 2-factor CFA (λ=1.01) indicates that *toughness* almost perfectly explains *resilience*, confirming that they are not manifesting as distinct constructs in this data (Kline, 2023).

Interestingly, item 3, "I try to see the humorous side of things when I am faced with problems", underperformed in both models. It was also only very slightly correlated with any of the other items on the measure (Table 4). This could indicate that humor is not playing the same role in contributing to resilience as other topics on the measure. It may be that humor is a component of resilience that is not adequately represented in the CD-RISC-1.

Ultimately, the outcome of both the CFA and ESEM supports utilizing this data as a singular representation of general resilience that does not clearly distinguish components. While this information can be used to speak to changes in general resilience that may be the outcome of these training programs, conclusions about exactly how resilience is being influenced cannot be drawn without a more sensitive, potentially expanded measure.

Originally, these models were estimated using a maximum likelihood (ML) estimator, treating the variables as though they were continuous (Rhemtulla et. al, 2012; Robitzsch, 2020). However, the ordinal structure of the Likert-scale questions in the CD-RISC-10 suggests that using weighted least squares maximum likelihood (WLSMV) estimators that are better suited for categorical data would improve model fit (Li, 2016; Savalei, 2021). Fit indices generated using the WLSMV estimator are scaled indices that are calculated using a χ^2 that is based on polychoric correlation matrices rather than covariance matrices, meaning that meaningful direct numeric comparisons cannot be made between those models and ones utilizing the ML estimator (Li, 2016; Rhemtulla et. al, 2012). However, the pronounced improvement in CFI and TLI with very little movement on RMSEA suggests that treating these items as categorical captures their true complexity better than if they are assumed to be continuous (Table 16).

Table 16: Progression of model fit indices for the CD-RISC-10

	χ^2	*df*	*RMSEA*	*CFI*	*TLI*
Single Factor CFA	85.8***	35	.09	.94	.92
2 Factor CFA	71.0***	33	.08	.95	.94
ESEM	45.6*	26	.06	.97	.95
Single Factor CFA (categorical)	95.8***	35	.09	.98	.97
ESEM (categorical)	50.1**	26	.07	.99	.98

Note: *n*=200, * p>.05, ** p>.01, ***p>.001

Attempts to conduct a chi-square difference test ($\Delta \chi^2$) between the two-factor categorical CFA and the two-factor categorical ESEM models were unsuccessful due to singular matrix issues. Specifically, the latent factors *of toughness* and *motivation* exhibited an extremely high correlation, resulting in model identification problems and violating the assumptions required for nested model comparisons. Consequently, the two factors were interpreted as one merged construct of resilience. Per recommendations for handling singular matrices and highly correlated latent variables in categorical modeling, model fit for the one-factor structure was evaluated based on approximate fit indices (CFI, TLI, RMSEA, and SRMR) rather than chi-square difference testing (Brown, 2015; Marsh et al., 2014).

Research Self-Efficacy (modified RSES)

While initial descriptive statistics and indicators for internal reliability suggested strong internal consistency (α=.97), through examination of inter-item correlations and factor structure, it became apparent that the unusually high *alpha* coefficient was the result of redundancies within the content rather than reliability (Table 5). For example, item 12, “Decide when to quit searching for related research/writing” and item 13, “Decide when to quit generating ideas based on your literature review” extremely strongly correlated (*r*=.94), suggesting that only one of those questions is needed to address the basic idea of knowing when to move on from the idea-generating process. Similarly, items 22 and 23, “choose an appropriate research design,” and “choose methods of data collection” (*r*=.89) cover the concept of study planning and design. While strong item correlations are not undesirable in a measure that is designed to address a unified construct, the items need to be different enough in substance so as not to be redundant. Beyond inter-item correlations, the correlations between theorized subscales are also a source of concern that the constructs are ill-defined, particularly between the *early tasks* and *conceptualization* variables (r = .90). Indeed, ESEM revealed many concerns about the structural stability of the 3-factor model used for interpreting data from this modified version of the RSES.

Under CFA constraints, all items appeared to load onto their theorized constructs. However, all model fit indicators demonstrate that even though items seem to fit their factors, the model does not explain the relationships observed in the data well. ESEM was used to explore potential sources of misfit. When the constraints of CFA are scaled back to allow for cross-loading in the ESEM, not only does the model fit not improve to acceptable levels, but the factor loadings fall apart from their intended constructs. *Early tasks* and *conceptualization* maintained only one of their items each. For the *early stages* variable, only Item 6, “find needed articles which are not available in your library,” loaded as predicted. Theoretically, this makes sense that this question is included as an early-stage process. Likewise, for the conceptualization variable, item 24, “be flexible in developing alternative research strategies,” does make sense to include as part of conceptualizing research. However, it should be noted that conceptually close questions, such as Item 22, “choose an appropriate research design,” did not load onto this factor, but onto *implementation*.

The *implementation* subscale is the only subscale that was adequately captured, although more items loaded onto it than were initially theorized. This suggests that the information taken from this modified measure is really only accessing self-efficacy around the implementation of research, and that questions believed to be predictors of earlier stages are better captured alongside the implementation stage. Significant revision of this measure will be needed if the goal is to capture multiple sub-domains of research self-efficacy.

All this combined information suggests that a full revision of the measure is likely necessary to capture additional domains of research self-efficacy beyond the implementation of research. It could also indicate the removal of certain items is appropriate before interpreting outcomes.

Table 17: Approximate fit indices for Research Self Efficacy Scale Models

	RMSEA	*CFI*	*TLI*	*SRMR*
CFA	.13	.69	.67	.09
ESEM	.11	.77	.72	.05

Resilience apart from Self-Efficacy

One encouraging component of this study was the determination that these measures do not meaningfully overlap one another. An initial concern was that because the CD-RISC-10 asks questions about how participants believe they would respond to a potentially hypothetical scenario, it was assessing self-efficacy around their participation in the program rather than resilience. Perhaps aided by the domain-specificity of self-efficacy as a construct, that does not appear to be the case. Items from each measure did not significantly load onto one another when examined using ESEM. While there are concerns with the factor structure of the modified RSES, none of its subscales correlated with resilience as measured by the CD-RISC-10 beyond r=.3. This slight but statistically significant (p<.001) adheres to the assertion that resilience and self-efficacy are related but theoretically distinct (Bandura, 2007).

CHAPTER V.

Discussion

The purpose of this study is to determine the psychometric properties of two measures used in the evaluation of outcomes for research training programs that target underrepresented students. These programs strive to foster resilience and increase self-efficacy around performing research tasks. Towards those ends it is crucial to have a trustworthy way to gauge the baseline status of those constructs to compare post-intervention. CFA was used to examine how well extant factor structures explained the relationship between items in this data, while ESEM was used to detail some of the potential problems with those factor structures. The results of those analyses have implications for the interpretation of the results of the resilience scale and suggest major restructuring of the research self-efficacy scale.

The first question this dissertation seeks to answer is whether a two-factor model for resilience will fit the data from our sample better than the one-factor model that is traditionally used. As hypothesized, the two-factor model did produce improved model fit statistics. However, improved model fit statistics do not necessarily indicate that this is the model through which the data gathered by the CD-RISC-10 should be interpreted. Examining the factor structure highlighted a conceptual imbalance in which all items strongly loaded onto toughness, and the only two items to load onto motivation also loaded more strongly onto toughness. This loading structure suggests that the CD-RISC-10 disproportionately captures toughness-related aspects of resilience, perhaps at the cost of examining the role motivation may play in resilient behavior. Because of this over-emphasis on toughness, the two-factor model structure was unstable, demonstrated by the >1.0 factor loading between toughness and resilience. A factor loading above 1 is a sign of multicollinearity as well as factor indistinguishability (Brown, 2015;

Devlieger & Rosseel, 2023; Marsh et. al., 2014). The items present in this form of the CD-RISC appear to narrowly capture resiliency and lack the construct validity needed to support a two-factor model.

When considering this measure in the context of assessment for research training programs for underrepresented undergraduate students, program evaluators need to consider what aspects of resilience they are trying to foster. Toughness as an attribute likely does not cover the full ability of an individual to resist and recover (Grafton et. al., 2019). Nor does it access the neurobiological markers indicating individual stress resilience (Wu et. al., 2013). If this version of the CD-RISC informs changes to program structure, they may neglect to cultivate aspects of motivation or purpose that strengthen resilient behavior. If the goal is to bolster resilience more comprehensively, it will be necessary to use a more exhaustive resilience measure.

The next question this dissertation addresses is whether the modified version of the RSES that was intended included only the conceptualization, early tasks, and implementation-oriented questions from the original RSES adequately assessed those latent variables. As hypothesized, model fit indices produced by CFA revealed unacceptable model fit, despite item loadings that aligned with the theorized factor structure. Using ESEM to allow for cross-loading to examine if the rigidity of traditional CFA was obscuring a better fitting structure produced only slightly improved model fit statistics, failing to meet the threshold for acceptable model fit. Interestingly, while model fit indices slightly improved, albeit not to acceptable standards, the factor structure largely deteriorated. Only questions intended to address the implementation stage maintained their theorized loadings, while other items showed significant cross-loading onto unintended factors or failed to load at all meaningfully.

The uneven loading of factors combined with a number of exceptionally high inter-item correlations strongly suggests the 38 questions on the modified RSES do not have the construct validity to claim to adequately evaluate the three distinct stages of research that are of interest (Brown & Moore, 2012; Clark & Watson, 1995). Compounding concerns about construct validity, the internal validity was extremely high (α=.97). While a strong alpha is typically desirable, when it is accompanied by high inter-item correlations it likely indicates redundancy rather than reliability (Clark & Watson, 1995; Kline, 2023; Streiner, 2003). In this case, it appears the items on the modified RSES overly represent the implementation stage of research, resulting in an artificially high reliability and deviated factor structure. This has serious implications for its usefulness as a measure to inform changes made to training programs that seek to bolster research self-efficacy.

Only a small portion of the research process is captured in the items on this measure. Undoubtedly, an effective training program will be able to impact the self-efficacy of its participants to perform all necessary tasks involved in the research process, including planning and development. As it stands, this measure is asking several relevant questions addressing their self-efficacy around performing research, but more questions that do not adequately address any domains of research self-efficacy that Bieschke et al. (1996) intended to address.

One potential explanation for the disjointed factor structure is that all respondents were novice researchers who had not yet begun their apprenticeship. As their understanding of the research process improves, so does their ability to adequately assess their belief that they are competent in performing research (Brancolini and Kennedy, 2017). Some participants may not yet be able to distinguish between research stages, which might influence their responses.

Retesting this model using data taken from the end of the apprenticeship and/or the years following it may produce different results.

Another goal of this dissertation is to determine whether the constructs of research self-efficacy and resilience are adequately differentiated from one another when these measures are used in tandem. Initially, it was hypothesized that because the CD-RISC asks questions primarily regarding beliefs about how one might respond, it was assessing self-efficacy, and there was expected to be significant cross-loading between items from the CD-RISC and the RSES. This turned out not to be the case. At an item level, there was no significant cross-loading between the two measures, supporting the idea that they address fundamentally distinct constructs. This separation provides some preliminary evidence of discriminant validity (Campbell & Fiske, 1959). The CD-RISC's emphasis on adaptability, personal strength, and stress coping in a general capacity overlaps very little with highly domain-specific questions concerning research self-efficacy (Connor & Davidson, 2003; Bieshcke et. al., 1996). Despite differentiating, the aggregate scores from each measure were significantly positively correlated, indicating that individuals with higher general resilience also tend to report greater confidence in their ability to perform research. This relationship aligns with theoretical expectations and supports their tandem use to more completely examine the impact of research training programs (Banudra, 1997, 2007; Waxman et al., 2003).

This dissertation explores the utility of ESEM as a more flexible and informative approach to assessing measurement validity for psychological constructs. Across measures, CFA produced theoretically consistent loadings that masked deeper issues with construct validity, including significant item redundancy and inability to discriminate between latent constructs. ESEM was able to uncover these issues and prevent program evaluators from interpreting the results of these measures in a way that their factor structure does not support. When looking at the overlap between measures, ESEM was able to provide evidence of discriminant validity as well as correlated composite scores. This study highlights the capacity of ESEM to uncover structural weaknesses that traditional CFA may obscure or overlook, especially when considering measures that have been modified and contain conceptual overlap (Alamer, 2022; Marsh & Alamer, 2024). Using ESEM in this dissertation underscored its value as a tool for construct validation that enhances theory by offering a more nuanced and empirically sensitive view of how psychological measures function in practice (Asparouhov & Muthén, 2009; Brown, 2015; Marsh et. al., 2014).

Conclusion

In this study, exploratory structural equation modeling (ESEM) proved to be an invaluable tool that identified instabilities in factor structures that would have otherwise been accepted if utilizing only confirmatory factor analysis (CFA). For both measures, a basic confirmatory analysis yielded stable and significant factor loadings for each theorized domain. In the case of the CD-RISC-10, the same was true of a 2-factor model that produced improved model-fit statistics. For the modified RSES, CFA also found that factor loadings held for their intended factors; however, model fit was far from acceptable. Using ESEM to unpack the results of each of these analyses produced valuable insights that shape not only how we interpret these results in context but also how they can be modified such that they maintain usefulness for their intended purposes.

Through ESEM, this study determined that, while slightly improving model fit, a two factor structure for resilience is not as stable when it comes to its factor structure, and that for the purposes of evaluating change in resilience, the CD-RISC-10 does not distinguish between factors well enough in this sample to be to determine if the programs being evaluated impact *resilience* through its theorized latent constructs. ESEM was also able to establish that the theorized factor structure of the questions on the original 53-item RSES do not hold in this modified version of the scale that intends to assess only three of the four originally identified domains. Therefore, this data should be interpreted cautiously, with a limited set of questions that can speak only to a single aspect of research self-efficacy.

Limitations

While this study produced meaningful and useful information, its potential impact and the depth of its analysis were limited by several factors. Firstly, limited demographic information was available about the participants contained in the data, limiting the ability of the study to generalize or identify trends based on demographic information that may shift the outcome of the analysis. This study primarily speaks to the structure of these measures when utilized to assess primarily Hispanic women who are first generation college students. Therefore, this study does not accurately capture the construct across the broader spectrum of underrepresented students.

Davidson (2020) asserts the stability of the CD-RISC across gender, ethnicity, and language. Smith et al. (2018) and Gonzales et al. (2016) found invariance between men and women on the CD-RISC-10 specifically. Other studies have identified small but significant differences in the way participants of different genders respond to the measure (Wolly & Jacobs, 2023). Likewise, there was no way to account for how resilience or self-efficacy may change in the context of socio-economic status. It is also possible that an increased sample size would help further define ill-determined factor structures for each of the measures.

Future research

The outcome of this study, in particular of the ESEMs, insists that the modified Research Self-Efficacy scale be further restructured to account for its evident limitations. For example, a condensed version of the measure, utilizing only those factors that loaded significantly, could be used as a valid measure of general research self-efficacy. Alternatively, if the goal is to produce a more nuanced picture of how research self-efficacy may evolve, one could begin with a re-examination of the 53-item version and then use exploratory factor analysis to determine a theoretically better fitting number of factors, before testing that new structure with ESEM to better define relationships.

This study provides evidence for the utility of localized psychometric tests beyond confirming a factor structure theorized while examining a separate population. Further, it demonstrates the capability of ESEM to identify issues and challenge assumptions about existing measures. Without applying ESEM to these measures, researchers would make statements about constructs they are not truly measuring, and evaluators would suggest changes to programs designed to impact those poorly measured constructs.

While the focus of this study was psychometric, the high mean scores and demographic make-up of responses on the CD-RISC-10 indicate important factors when thinking about how resilience manifests and is bolstered in young Hispanic women. While personal resilience is linked to academic success for Hispanic women, that resilience largely originates from navigating cultural tensions and gender expectations, and is formed by leveraging family support, involvement with community, and self-determination (Para, 2007). A systematic review by Cardoso & Thompson (2018) discerned that resilience is a dynamic, multidimensional process, informed by multiple domains. Positive ethnic identity, spiritual beliefs, and biculturalism were found to be important determinants of resilience and buffer the psychological distress of discrimination (Bartoszek et. al. 2020; Cardoso & Thompson, 2018; Craney & Watson, 2024). The integration of those determinants into Hispanic culture may explain why Hispanic women have been found to have higher resilience compared to other ethnic groups despite being disadvantaged (D'Agostino & Waldrop, 2014; Emdur, 2022).

The data in this study aligns with the observations of previous research (D'Agostino & Waldrop, 2014; Emdur, 2022). The multiple sources of resilience rooted in community and culture highlight the need for culturally informed mentorship. Future work can examine the role of culture in research apprenticeships. Additionally, it may be that the experiences that shape resilience in Hispanic women who elect to participate in a research apprenticeship may be distinct from the experiences of others.

An additional direction for future work is to refine a measure of resilience that is specific to the challenges faced by academics and researchers. Resilience is largely examined in the context of health outcomes, rather than achievement (Connor & Davidson, 2023; Denckla et. al. 2020). The kinds of challenges to which researchers have to be resilient are specific. Does a measurement generated based on health outcomes adequately capture one's resilience to professional criticism or rejection? The creation and validation of a scale specific to resilience to the challenges encountered as a professional researcher would better aid in assessing the outcomes of research training programs.

APPENDIX A

CONNOR-RESILIENCE SCALE

Please indicate how much you agree with the following statements as they apply to you over the last month. If a particular situation has not occurred recently, answer according to how you think you would have felt.

	not true at all (0)	rarely true (1)	sometimes true (2)	often true (3)	true nearly all the time (4)
1. I am able to adapt when changes occur.	○	○	○	○	○
2. I can deal with whatever comes my way.	○	○	○	○	○
3. I try to see the humorous side of things when I am faced with problems.	○	○	○	○	○
4. Having to cope with stress can make me stronger.	○	○	○	○	○
5. I tend to bounce back after illness, injury, or other hardships.	○	○	○	○	○
6. I believe I can achieve my goals, even if there are obstacles.	○	○	○	○	○
7. Under pressure, I stay focused and think clearly.	○	○	○	○	○
8. I am not easily discouraged by failure.	○	○	○	○	○
9. I think of myself as a strong person when dealing with life's challenges and difficulties.	○	○	○	○	○
10. I am able to handle unpleasant or painful feelings like sadness, fear, and anger.	○	○	○	○	○

APPENDIX B

<u>Research Self Efficacy Scale</u>

Think about your level of confidence in your ability to perform each behavior listed and place a number in the blank to the right of the item indicating the degree of confidence in your ability to successfully perform that behavior. Use the following scale to make your ratings.

0 10 20 30 40 50 60 70 80 90 100
No Confidence Moderate Confidence Complete Confidence

1. How confident are you in your overall ability to complete a significant project? _____
2. Follow ethical principles of research. _____
3. Brainstorm areas in the literature to read about. _____
4. Conduct a computer search of the literature in a particular area. _____
5. Locate references by manual search. _____
6. Find needed articles which are not available in your library. _____
7. Evaluate journal articles in terms of the theoretical approach, experimental design and data analysis techniques. _____
8. Participate in generating collaborative research ideas. _____
9. Work interdependently in a research group. _____
1. Discuss research ideas with peers. _____
11. Consult senior researchers for ideas. _____
12. Decide when to quit searching for related research/writing. _____
13. Decide when to quit generating ideas based on your literature review. _____
14. Synthesize current literature. _____
15. Identify areas of needed research, based on reading the literature. _____
16. Develop a logical rationale for your particular research idea. _____
17. Generate researchable questions. _____
18. Organize your proposed research ideas in writing. _____
19. Effectively edit your writing to make it logical and succinct. _____
2. Present your research idea orally or in written form to an advisor or group. _____
21. Utilize criticism from reviews of your idea. _____
22. Choose an appropriate research design. _____
23. Choose methods of data collection. _____
24. Be flexible in developing alternative research strategies. _____
25. Choose measures of dependent and independent variables. _____
26. Choose appropriate data analysis techniques. _____
27. Obtain approval to pursue research (e.g., approval from Human Subject's Committee/IRB, Animal Subject's Committee, special approval for fieldwork, etc.). _____
28. Obtain appropriate participants/general supplies/equipment. _____
29. Train assistants to collect data. _____
3. Perform experimental procedures. _____
31. Ensure data collection is reliable across trial, raters, and equipment. _____
32. Supervise assistants _____
33. Attend to all relevant details of data collection. _____
34. Organize collected data for analysis. _____
35. Use computer software to prepare texts (word processing). _____
36. Use computer software to generate graphics. _____
37. Use a computer for data analysis. _____
38. Develop computer programs to analyze data. ___

REFERENCES

Alamer, A. (2022). Exploratory structural equation modeling (ESEM) and bifactor ESEM for construct validation purposes: Guidelines and applied example. *Research Methods in Applied Linguistics, 1*, 100005. https://doi.org/1.1016/j.rmal.2022.100005

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, 106(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Aloba O, Olabisi O, Aloba T. (2016) The 10-Item Connor–Davidson Resilience Scale: Factorial Structure, Reliability, Validity, and Correlates Among Student Nurses in Southwestern Nigeria. *Journal of the American Psychiatric Nurses Association*.22(1):43-51. doi:1.1177/1078390316629971

American Psychological Association. (2021). Apology to People of Color for APA's Role in Promoting, Perpetuating, and Failing to Challenge Racism, Racial Discrimination, and Human Hierarchy in U.S. *Council Policy Manual*. Retrieved from https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/41899596/2c430da8-cf1c-4f8a-b57a-e6a8f799e6d4/resolution-racism-apology.pdf

Appel, M., & Kronberger, N. (2012). Stereotypes and the achievement gap: Stereotype threat prior to test taking. *Educational Psychology Review*, *24*, 609-635.

Asheim, G. B., Bossert, W., D'Ambrosio, C., & Vögele, C. (2020). The measurement of resilience. *Journal of Economic Theory*, 189, 105104. https://doi.org/1.1016/j.jet.202.105104

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/1.1080/10705510903008204

Ayed, N. Toner, S., and Priebe, S. (2019) Conceptualizing resilience in adult mental health literature: A systematic review and narrative synthesis. Psychology and Psychotherapy: Theory, Research and Practice. 92, 299-341. The British Psychological Society

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*.

Bandura, A. (1978). Reflections on self-efficacy. *Advances in behaviour research and therapy*, *1*(4), 237-269.

Bandura, A. (1981). Self-referent thought: A developmental analysis of self-efficacy. *Social cognitive development: Frontiers and possible futures*, *200*(1), 239.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American psychologist*, *37*(2), 122

Bandura, A. (1983). Self-efficacy determinants of anticipated fears and calamities. *Journal of personality and social psychology*, *45*(2), 464

Bandura, A.(1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice–Hall.

Bandura, A. (1996). Failures in self-regulation: Energy depletion or selective disengagement?. *Psychological Inquiry*, *7*(1), 20-24.

Bandura, A. (1997). *Self-efficacy: The exercise of control* (Vol. 604). Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents*, *5*(1), 307-337.

Bandura, A., & Adams, N. E. (1977). Analysis of self-efficacy theory of behavioral change. *Cognitive therapy and research*, *1*(4), 287-31.

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41, 586–598.

Bartoszek, L.A., Jacobs, W., & Unger, J.B. (2020) Correlates of Resilience in Hispanic Young Adults. *Family and Community Health*. 43(3):p 229-237. *DOI:* 10.1097/FCH.0000000000000261

Bieschke, K. J., Bishop, R. M., & Garcia, V. L. (1996). The Utility of the Research Self-Efficacy Scale. *Journal of Career Assessment*, 4(1), 59-75. https://doi.org/1.1177/106907279600400104

Blair, J., Czaja, R. F., & Blair, E. A. (2013). *Designing surveys: A guide to decisions and procedures*. sage publications.

Blackman, D. (2022). Conditioned suppression and the effects of classical conditioning on operant behavior. *Handbook of operant behavior,* 340-363.

Boardman, F., Griffiths, F., Kokanovic, R., Potiriadis, M., Dowrick, C., & Gunn, J. (2011). Resilience as a response to the stigma of depression: A mixed methods analysis. *Journal of affective disorders*, 135(1-3), 267-276.

Bonanno, G. A., Romero, S. A., & Klein, S. I. (2015). The temporal elements of psychological resilience: An integrative framework for the study of individuals, families, and communities. *Psychological Inquiry*, 26, 139–169. https://doi.org/1.1080/1047840x.2015.992677

Brancolini, K. R. and Kennedy, M. (2017). The development and use of a research self-efficacy scale to assess the effectiveness of a research training program for academic librarians. *Library and Information Research*, 41(124), 44-84.

Brodzinsky, D. M. (1993). On the use and misuse of psychological testing in child custody evaluations. *Professional Psychology: Research and Practice,* 24(2), 213–219. https://doi.org/1.1037/0735-7028.24.2.213

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). *Guilford Press*

Butz, A. R., Byars-Winston, A., Leverett, P., Branchaw, J., & Pfund, C. (2018). Promoting STEM trainee research self-efficacy: A mentor training intervention. *Understanding Interventions Journal*, *9*(1).

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. https://doi.org/10.1037/h0046016

Campbell-Sills, L., & Stein, M. B. (2007). Psychometric analysis and refinement of the Connor–Davidson resilience scale (CD-RISC): Validation of a 10-item measure of resilience. Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies, 20(6), 1019-1028.

Cardoso, J. B. & Thompson, S.J. (2018) Common Themes of Resilience Among Latino Immigrant Families: A Systematic Review of Literature. *Families in Society: The Journal of Contemporary Social Services.* Alliance for Children and Families. DOI: 10.1606/1044-3894.4003

Child, L.E. & Medvedev, O.N.(2024) Investigating state and trait aspects of resilience using Generalizability theory. *Curr Psychol* 43, 9469–9479 https://doi.org/1.1007/s12144-023-05072-4

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor–Davidson Resilience Scale (CD-RISC). Depression and Anxiety, 18, 76–82.

Cowen, E. L., Work, W. C., Wyman, P. A., Parker, G. R., Wannon, M., & Gribble, P. (1992). Test comparisons among stress-affected, stress-resilient, and nonclassified fourth-through sixth-grade urban children. Journal of Community Psychology, 20(3), 200-214.

Craney, R. S., & Watson, L. B. (2024). Pride and prejudice: The relations among multiracial microaggressions, resilience, and psychological distress for Latine/White multiethnic individuals. *Journal of Latinx Psychology.* Advance online publication. https://doi.org/10.1037/lat0000262

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. https://doi.org/1.1037/h0040957

Cronbach, L.J. (1971) Test Validation. R.L. Thorndike (Ed.) Educational Measurement (Second Edition, p.443-507) American Council on Education.

Crowe, A., Averett, P., & Glass, J. S. (2016). Mental illness stigma, psychological resilience, and help seeking: What are the relationships? Mental Health and Prevention, 4(2), 63–68. https://doi.org/1.1016/j.mhp.2015.12.001

D'Agostino, L., & Waldrop J. (2014). The associations among maternal resiliency, perceptions of touch, and reports of infant touch. The University of Florida Undergraduate Research Journal, 7(1), 1-14.

Davidson, J. R. T. (2020). Connor-Davidson Resilience Scale (CD-RISC) Manual. Unpublished manuscript. Available from http://www.cd-risc.com

Deci, E. L., Ryan, R. M., (1985). Conceptualizations of intrinsic motivation and self-determination. *Intrinsic motivation and self-determination in human behavior*, 11-4.

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of personality and social psychology*, *53*(6), 1024.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. Educational psychologist, 26(3-4), 325-346.

Denckla, C. A., Cicchetti, D., Kubzansky, L. D., Seedat, S., Teicher, M. H., Williams, D. R., & Koenen, K. C. (2020). Psychological resilience: an update on definitions, a critical appraisal, and research recommendations. *European Journal of Psychotraumatology*, 11(1), 1822064. https://doi.org/1.1080/20008198.202.1822064

Emdur, R.D. (2022) Hispanic Women and Resilience: A qualitative study. [Doctoral dissertation, Immaculata University].

Evans, R. (Interviewer). (1964). *B.F. Skinner interviewed by Richard Evans* [Interview]. Cummings Center for Historical Psychology. Internet Archive. https://archive.org/details/bf-skinner-interviewed-by-richard-evans-1964

Feltz, D. L., & Mugno, D. A. (1983). A replication of the path analysis of the causal elements in Bandura's theory of self-efficacy and the influence of autonomic perception. *Journal of Sport and Exercise Psychology*, *5*(3), 263-277.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, 86, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Ford, D. (2004) Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations. *The National Research Center on the Gifted and Talented.*

Forester, M., Kahn, J. H., & Hesson-McInnis, M. S. (2004). Factor Structures of Three Measures of Research Self-Efficacy. *Journal of Career Assessment, 12*(1), 3–16. https://doi.org/1.1177/1069072703257719

Garcia, C. M. and Hamilton, K. D. (2023) "History Untold: A Historical Review of Psychological Harm of Racialized Minorities," *Psychology from the Margins: Vol. 5, Article 1.*

Garmezy, N. (1972). Invulnerable children: The fact and fiction of competence and disadvantage. American Journal of Orthopsychiatry, 42, 328.

Grafton, R.Q., Doyen, L., Béné, C., Borgomeo, E., Brooks, K., Chu, L., Cumming, G.S., Dixon, J., Dovers, S., Garrick, D., Helfgott, A., Jiang, Q., Katic, P., Kompas, T., Little, L.R., Matthews, N., Ringler, C., Squires, D., Steinshamn, S.I., Villasante, S., Wheeler, S., Williams, J., Wyrwoll, P.R., (2019). Realizing resilience for decision-making. *Nature Sustainability* 2, 907–913.

Greeley, A. T., Johnson, E., Seem, S., Braver, M., Dias, L., Evans, K., Kincade, E., & Pricken, P. (1989). [Research Self-Efficacy Scale]. Unpublished scale presented at the conference of the Association for Women in Psychology, Bethesda, MD.

Greene, R. R., Galambos, C., & Lee, Y. (2004). Resilience theory: Theoretical and professional conceptualizations. *Journal of Human Behavior in the Social Environment*, 8(4), 75–91. https://doi.org/1.1300/J137v08n04_05

Gonzales, S.P., Moore, W.G., Newton, M., Galli, N. (2016) Validity and reliability of the Connor-Davidson Resilience Scale (CD-RISC) in competitive sport. *Psychology of Sport and Excercise*, 23, 31-39

Goulart, E. L., Weymer, A. S. Q., & Moreira, V. R. (2022). The influence of self-efficacy on training effectiveness in cooperative organizations. Revista de Administração da UFSM, 15(2), 331–353. https://doi.org/1.5902/1983465967213

Hemmings, B., &Kay, R. (2010). Research self-efficacy, publication output, and early career development.International Journal of Educational Management, 24(7), 562-574.doi: 1.1108/09513541011079978

Hemmings, B., & Kay, R. (2016). The relationship between research self-efficacy, research disposition and publication output. *Educational Psychology: An International Journal of Experimental Educational Psychology, 36*(2), 247-361

Herrman, H., Stewart, D. E., Diaz-Granados, N., Berger, E. L., Jackson, B., & Yuen, T. (2011). What is resilience?. *The Canadian Journal of Psychiatry*, *56*(5), 258-265.

Hu, L. & Bentler, P.M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal*, 6:1, 1-55, DOI: 1.1080/10705519909540118

IBM Corp. (2022). IBM SPSS Statistics for Windows (Version 29.0) [Computer software]. IBM Corp.

Johnson, B. (2008). Teacher-student relationships which promote resilience at school: A micro-level analysis of students' views. British Journal of Guidance & Counselling, 36(4), 385-398.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183-202.

Kail, R. V., & Hall, L. K. (1994). Processing speed, naming speed, and reading. *Developmental Psychology*, 30(6), 949–954. https://doi.org/10.1037/0012-1649.30.6.949

Kline, R. B. (2023). Principles and practice of structural equation modeling (5th ed.). The Guilford Press.

Kuiper, H., van Leeuwen, C. C. M., Stolwijk-Swüste, J. M., et al. (2019). Measuring resilience with the Connor–Davidson Resilience Scale (CD-RISC): Which version to choose? *Spinal Cord*, 57(5), 360–366. https://doi.org/1.1038/s41393-019-0240-1

Lazarus, R. S. (1984). Puzzles in the study of daily hassles. *Journal of Behavioral Medicine*, 7(4), 375-389

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of vocational behavior*, *45*(1), 79-122.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Luthar, S. S., & Zigler, E. (1991). Vulnerability and competence: a review of research on resilience in childhood. *The American journal of orthopsychiatry*, 61(1), 6–22. https://doi.org/1.1037/h0079218

Maddux, J. E. (Ed.). (2013). Self-efficacy, adaptation, and adjustment: Theory, research, and application. *Springer Science & Business Media*.

Marsh, H., & Alamer, A. (2024). When and how to use set-exploratory structural equation modelling to test structural models: A tutorial using the R package lavaan. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–18.

Marsh, H., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 439–476. https://doi.org/1.1080/10705510903008220

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Martin, A. J., & Marsh, H. W. (2006). Academic resilience and its psychological and educational correlates: A construct validity approach. *Psychology in the Schools*, 43(3), 267-281

McLafferty, M., Mallet, J., & McCauley, V. (2012). Coping at university: The role of resilience, emotional intelligence, age and gender. *Journal of Quantitative Psychological Research*, 1, 1-6.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749

Microsoft Corporation. (2021). Microsoft Excel (Version 16.0) [Computer software]. https://office.microsoft.com/excel

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 395–436). IAP Information Age Publishing.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. Educational researcher, 25(1), 20-29.

Muthén, L. K., & Muthén, B. O. (1998–2022). Mplus User's Guide (8th ed.). Muthén & Muthén.

Niemeyer, H., Knaevelsrud, C., van Aert, R. C. M., & Ehring, T. (2023). Research Into Evidence-Based Psychological Interventions Needs a Stronger Focus on Replicability. *Clinical psychology in Europe*, 5(3), e9997. https://doi.org/1.32872/cpe.9997

Parra, M. O. (2007). Sociocultural, resilience, persistence and gender role expectation factors contribute to the academic success of hispanic females (Order No. 3274948).[Doctoral dissertation, Kansas State University] ProQuest Dissertations and Theses Global.

Pasupathy, R., & Siwatu, K. O. (2014). An investigation of research self-efficacy beliefs and research productivity among faculty members at an emerging research university in the *USA. Higher Education Research & Development*, 33(4), 728-741.

Rhemtulla, B., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. https://doi.org/10.1037/a0029315

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. In *Frontiers in education* (Vol. 5, p. 589965). Frontiers Media SA.

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational psychologist*, *40*(1), 53-63.

Sanders, J. D., & Katz, S. (2013) The overuse and misuse of psychological testing: why less is more. *American Journal of Family Law*, 26(4), 221-225.

Savalei, V. (2021) Improving Fit Indices in Structural Equation Modeling with Categorical Data. *Multivariate Behavioral Research*, 56(3), 390-407. https://doi.org/10.1080/00273171.2020.1717922

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of educational psychology*, *73*(1), 93.

Schunk, D. H. (2012). Social cognitive theory. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1. Theories, constructs, and critical issues* (pp. 101–123). American Psychological Association. https://doi.org/1.1037/13273-005

Seginer, R. (2008). Future orientation in times of threat and challenge: How resilient adolescents construct their future. International Journal of Behavioral Development, 32(4), 272-282.

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology,* 26(6), 978–986. https://doi.org/1.1037/0012-1649.26.6.978

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual review of psychology*, *67*(1), 415-437.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science,* 103(2684), 677–68.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Terman, L. M., & Merrill, M. A. (1937). Measuring intelligence: A guide to the administration of the new revised Stanford-Binet tests of intelligence. Houghton Mifflin

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. *Teachers College Record*, 19(7), 16-24.

Thorndike, E. L. (1927). The law of effect. *The American journal of psychology*, 39(1/4), 212-222.

University Analytics & Institutional Research. (2024). *Enrollment.* University of Arizona. Retrieved April 1, 2025 from https://uair.arizona.edu/content/enrollment

Van Zyl, L. E., Ten Klooster, P. M., & Rothmann, S. (2022). A practical tutorial and convenient online tool for estimating exploratory structural equation modeling (ESEM) with Mplus. *Frontiers in Psychiatry*, 12, 795672. https://doi.org/1.3389/fpsyt.2021.795672

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. *Harvard University Press*.

Wakefield, H., & Underwager, R. (1993). Misuse of psychological tests in forensic settings: Some horrible examples. *American Journal of Forensic Psychology*, 11(1), 55–75.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes. *Psychological science*, 29(7), 1159–1177. https://doi.org/1.1177/0956797618761661

Waxman, H. C., Gray, J. P., & Padron, Y. N. (2003). Review of research on educational resilience. Center for Research on Education, Diversity & Excellence.

Weinberg, R., Gould, D., & Jackson, A. (1979). Expectations and performance: An empirical test of Bandura's self-efficacy theory. *Journal of Sport and Exercise Psychology*, *1*(4), 320-331.

Wolly, A.I., & Jacobs, I. (2023) Validity and reliability of the German versions of the CD-RISC-10 and CD-RISC-2. *Current Psychology: Research and Reviews 52(5), 3437-3448*. New York. DOI:1.1007/s12144-021-01670-2

Wu, G., Feder, A., Cohen, H., Kim, J. J., Calderon, S., Charney, D. S., & Mathé, A. A. (2013). Understanding resilience. *Frontiers in behavioral neuroscience*, *7*, 1.

Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, 37(1), 215-246.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, *25*(1), 82-91.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting. *American Educational Research Journal*, *29*(3), 663-676. https://doi.org/1.3102/00028312029003663

Zolkoski, S. M., & Bullock, L. M. (2012). Resilience in children and youth: A review. *Children and Youth Services Review*, 34(12), 2295-2303.

Zueger, R., Niederhauser, M., Utzinger, C., Annen, H., & Ehlert, U. (2022). Effects of resilience training on mental, emotional, and physical stress outcomes in military officer cadets. *Military Psychology*, 35(6), 566–576. https://doi.org/1.1080/08995605.2022.2139948