Response Shift Bias: An Examination of Measurement Invariance in Self-Reported Change

Katherine Nelson Daniels

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
RESPONSE SHIFT BIAS: AN EXAMINATION OF MEASUREMENT INVARIANCE IN SELF-REPORTED CHANGE

by

Katherine Nelson Daniels

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Leadership, Research and Technology Western Michigan University April 2018

Doctoral Committee:

Edward Brooks Applegate, Ph.D., Chair
Kim Nimon, Ph.D.
Jessaca Spybrook, Ph.D.
David Hartman, Ph.D.
RESPONSE SHIFT BIAS: AN EXAMINATION OF MEASUREMENT INVARIANCE IN SELF-REPORTED CHANGE

Katherine Nelson Daniels, Ph.D.
Western Michigan University, 2018

Traditional pre-test (TpT)/post-test (PT) and retrospective pre-test (RpT)/post-test (PT) designs are used to collect data on self-reported measures to assess the magnitude of change that occurs from interventions. If measurement invariance does not exist across the measurement occasions within these research designs, it is inappropriate to compare mean group differences that result from the intervention and derive inferences about change. The theory of response-shift suggests a subject’s understanding of a subject matter at TpT may not be the same as their understanding of a subject matter at PT, and that as a result, the construct measured at these measurement occasions may not be the same, or may not have the same structural components (factor loadings and scale). RpT/PT research designs have been suggested as an alternative to TpT/PT research designs to control for response-shift bias. Unfortunately, measurement invariance is rarely investigated in either of these research designs, it is merely assumed. Given this, it is important to understand the extent to which both TpT/PT and RpT/PT research designs demonstrate measurement invariance in various contexts and the impact measurement invariance may have on effect size estimates. The principle aim of this dissertation study was to examine the theory of response-shift bias by testing if an instrument administered in a TpT/PT design evidences the same structural meaning as the same instrument administered in a RpT/PT design, and the extent to which the observed (raw) scores obtained in the context of these designs are the
same as the latent means. This study examines self-reported change of communication reticence using a longitudinal measurement invariance model. In this study, the measurement occasions in the TpT/PT research design are invariant to the level of strong. This indicates that there is not a shift in the understanding of the construct in either research design. However, the measurement occasions in the RpT/PT research design were only found to be invariant to the level of weak. A partial invariance analysis of the RpT/PT model revealed that when the intercept of an indicator associated with knowledge was unconstrained, strong invariance was achieved.

In the TpT/PT research design an effect size analysis revealed that the raw scores underestimate the effect, however the difference between the raw and latent means is not statistically significantly different. This study would suggest, that in this context, the TpT/PT and RpT/PT research designs resulted in the explication of the same construct at pre-test and post-test and given this both research designs could be used to derive valid inferences about the constructs being measured. Because measurement invariance in the RpT/PT was not found beyond the level of weak, only the TpT/PT research design could be used to derive valid inferences about the magnitude of pre-test and post-test scores, unless the constraint for on the intercept for the knowledge indicator was relaxed. Hence in this context, only the TpT/PT research design should be used to derive valid inferences about the magnitude of change that resulted from the intervention.
ACKNOWLEDGMENTS

I would like to thank Dr. Edward Brooks Applegate, who throughout my graduate work, pushed me to challenge myself and then helped see me through difficult times. Thank you for not giving up on me when “life got in the way of life.”

To my parents, who, from the time I was born, instilled in me the value of education, modeled educational attainment, and supported me in the pursuit of my personal educational goals, thank you. It is not possible for me to adequately articulate the extent to which my education has transformed my views and broadened my perspectives. I am eternally grateful for the educational opportunities you provided me with early in my life, which laid the foundation for all subsequent learning.

To my family, Mark, Kyleigh, and Micah—thank you for sharing a big part of your life with this work over the last several years.

To my friends and colleagues who nudged me forward and offered words of encouragement at the times I needed it most, thank you.

I would also like to thank the other members of my committee, Dr. Kim Nimon, Dr. Jessaca Spybrook, and Dr. David Hartman. I sincerely appreciate your expertise and thoughtful feedback. I would especially like to thank Dr. Nimon for introducing me to the theory of response-shift when I did not yet have the measurement vocabulary to articulate this phenomenon that I was observing in practice and was fascinated by.

Katherine Nelson Daniels
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... ii  

LIST OF TABLES ................................................................................................................ vi  

LIST OF FIGURES ................................................................................................................ viii  

CHAPTER  

I. BACKGROUND ...................................................................................................................... 1  
   Validity ............................................................................................................................... 4  
   Research Design ............................................................................................................... 5  
   Measurement Invariance ................................................................................................. 6  
   Statement of the Problem ............................................................................................... 8  
   Research Questions ........................................................................................................ 8  
   Significance of the Research ......................................................................................... 10  

II. REVIEW OF THE LITERATURE ..................................................................................... 12  
   Validity ............................................................................................................................. 12  
      Validity Judgments ....................................................................................................... 13  
      Sources of Validity Evidence ..................................................................................... 15  
   Research Design ............................................................................................................ 17  
      Descriptions of Experiments ...................................................................................... 17  
      Research Designs and Measuring Change .................................................................. 19  
      Threats to Validity in the Context of Research Design ............................................ 21  
   Measurement Invariance ............................................................................................... 27
Table of Contents—Continued

CHAPTER

Measurement Invariance................................................................. 60
Research Question 1 ........................................................................... 62
Research Question 2 ........................................................................... 71

V. CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS.................. 76
  Conclusions......................................................................................... 76
  Discussion........................................................................................... 78
  Limitations ......................................................................................... 81
  Recommendations ............................................................................. 82

REFERENCES ......................................................................................... 84
LIST OF TABLES

1. Threats to Statistical Conclusion Validity ................................................................. 23
2. Threats to External Validity .......................................................................................... 24
3. Construct Validity .......................................................................................................... 24
4. Threats to Internal Validity ........................................................................................... 26
5. RS Subscales and Items .............................................................................................. 46
6. Missing Data at Measurement Occasions .................................................................... 48
7. Multivariate Normality Tests (N = 211) ..................................................................... 54
8. Baseline CFA Model Fit (N = 211) .............................................................................. 56
9. Descriptive Statistics for Reticence Scale and Subscales (N = 211) ......................... 59
10. t-Test Analysis of RS Scale Scores (N = 211) .............................................................. 61
12. Satorra-Bentler Scaled Chi-Square Difference and Delta(CFI) for RpT/TpT Models ...... 64
13. TpT/PT Measurement Invariance Models for the RS .................................................... 65
14. Satorra-Bentler Scaled Chi-Square Difference Tests and Delta (CFI) for TpT/PT ........ 66
15. RpT/PT Measurement Invariance Models for the RS .................................................... 67
16. Satorra-Bentler Scaled Chi-Square Difference Tests and Delta(CFI) for RpT/PT ........ 68
17. Weak to Strong Partial Invariance for the Unconstrained RpT/PT Model .................... 70
18. t-Test Analysis of RS Scale Scores in the TpT/PT Research Design (N = 211) .......... 73
20. t-Test Analysis of RS Scale Scores in the RpT/PT Research Design (N = 211) .......... 74
List of Tables—Continued

LIST OF FIGURES

1. Simple CFA model............................................................................................................. 30
2. RS structural model........................................................................................................... 56
3. TpT estimated model (unstandardized solution). ............................................................. 57
4. RpT estimated model (unstandardized solution).............................................................. 57
5. PT estimated model (unstandardized solution)................................................................. 58
6. Longitudinal CFA model. ................................................................................................. 62
CHAPTER I

BACKGROUND

In the age of accountability, there are increasing requirements to demonstrate progress in meeting established outcomes and benchmarks. This has prompted an increase in formal and informal research, evaluation, and assessment that utilizes experimental, quasi-experimental, or pre-experimental research designs to advance knowledge about the effectiveness of educational and human services interventions. The ultimate goal of this research is to obtain data from which valid inferences may be derived about change in knowledge, skills, behaviors, or attitudes in order to demonstrate impact from an intervention.

When measuring self-reported change in attitudes or behavior, an area of substantive concern is the variability of effect size estimates yielded from traditional pre-test/post-test (TpT/PT) and retrospective pre-test/post-test (RpT/PT) research designs (also referred to as within designs). Some suggest that self-reported measurements are prone to bias (Klatt & Taylor-Powell, 2005; Sibthorp et al., 2007) and many researchers have questioned the validity of the inferences derived from such measures (Albanese et al., 2006; Bardella, Janosky, Elicki, Ploof, & Kolarik, 2005; Barnsley et al., 2004; Cantrell, 2003; Davis et al., 2006; Eva et al., 2004; Evans, Leeson, Newton John, & Petrie, 2005; Manthei, 1997; Pratt et al., 2000; Vnuk, Owen, & Plummer, 2006; Wayne et al., 2006). Given this, research has been conducted and theories have been developed that explore and define validity concerns associated with research designs that purport to measure self-reported change.

Measuring change in the context of a TpT/PT or RpT/PT research design requires the conceptualization of a construct, as well as time and consideration of the timing of the measurement occasions (Little, 2013), which is more complex than often realized. The
researcher must take this complexity, as well as factors related to convenience, timing of access to the population, budget, and resources into consideration when making decisions about the research design they will employ. The decisions made in the research planning phases have critical implications for the implementation of the research design, the subsequent findings, and ultimately the validity of the inferences derived from the study.

Considerations about the research design that impact the internal structure of the data are of particular importance when measuring change because comparisons of the constructs or latent variable across measurement occasions often occur through analyses that compare scores collected at different time points. Comparisons of mean scores derived from a set of indicators at different times is only appropriate and justified when it has been determined that the structure of the data is invariant across the measurement occasions. It is the invariance across measurement occasions that provides credence to the validity of the inferences about change derived from the study. Measurement invariance of an observed score exists when the estimate of an individual’s observed score does not depend on the measurement occasion, i.e., the meaning of the observed score is not dependent on the measurement occasion. Measurement invariance must be differentiated from structural invariance. In structural invariance, the level of expression of the trait does not change over measurement occasions.

Several studies have been conducted to better understand the validity of inferences derived from research about change, and have resulted in theories and considerations for TpT/PT and RpT/PT research designs. These theories include response-shift theory (Howard & Dailey, 1979; Howard et al., 1979), personal recall theory (Schwartz & Rapkin, 2004), implicit theory of change (Lam & Bengo, 2003; Norman, 2003), and impression management theory (Paulhus, 2002; Tedeschi, Schlenker, & Bonoma, 1971). Of particular relevance to this study is response-
shift theory; the other theories will be covered briefly in Chapter II.

Response-shift theory is concerned with the equivalence of a construct at each measurement occasion in TpT/PT or RpT/PT research designs. If a lack of measurement invariance exists over the measurement occasions, a comparison of the mean group differences cannot be inferred to represent a change in the construct over time regardless of the design, TpT/PT or RpT/PT. Unfortunately, measurement invariance is rarely investigated by the researcher, bringing into question the validity of effect size estimates measured in a longitudinal or within subject design. Thus, it is important to understand the extent to which TpT/PT and RpT/PT designs demonstrate measurement invariance.

Howard and his colleagues (Howard & Dailey, 1979; Howard et al., 1979) offered a concrete example of response-shift theory. Workshop participants might believe at TpT that they are "average" leaders. The intervention changes their understanding of the skills involved in being a leader or the construct of leadership as it was internalized by participants. This change in participants’ understanding of leadership is a change in the understanding of the construct leadership, and is illustrative of a lack of construct invariance. If after the workshop is completed participants believe they were actually “below average” leaders prior to the workshop, their TpT score would not reflect this. If after the workshop, based on their new understanding of leadership, they felt they were an “average” leader, the TpT and PT scores would be the same. Unfortunately, these ratings are based on a different understanding of the leadership construct. Erroneous inferences would be drawn if it is concluded that the study participants had not benefitted from the workshop. Whenever such shifts in understanding occur, conventional self-report TpT/PT designs are unable to accurately gauge treatment effects (Howard & Dailey, 1979). In this way, response-shift theory is an explanation for a lack of measurement invariance.
In order to understand response-shift theory and its implications for research designs, it is imperative to understand how valid inferences are derived from a study, and the role and function of research design, measurement, and statistical analysis in this inference. Research methods are not specifically aligned with any one aspect of validity or source of validity evidence. Instead, research design choices have numerous consequences for validity, which in some cases, are unanticipated.

Validity

Ensuring the validity of the inferences derived from the conclusions of the research study is critically important if the study conclusions are to be considered as part of the growing body of knowledge. The systematic collection of educational or psychological data for the purpose of deriving inferences about a particular phenomenon is broadly conceptualized as a test. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education have jointly published Standards for Educational and Psychological Testing in 1985, 1999, and 2014. The purpose of the Standards is to “provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses,” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 1). The Standards are meant to apply to all research designs, measurement, and data analysis activities. According to the Standards (American Educational Research Association et al., 2014), interpretation of inferences requires professional judgment, however, it is critical to “ensure that relevant issues are addressed” (p. 1). The relevant issues (Standards) for validity are centered around three themes or clusters: Establishing Intended Uses and Interpretations, Issues Regarding Samples and Settings used in
Validation, and Specific forms of Validity Evidence. Common sources of evidence to support validity arguments include: evidence based on test content, evidence based response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence for validity consequences of testing. The clusters and the sources of evidence listed in the Standards identify the aspects of research design, measurement, and analysis evidence needed to support appropriate inferences, for example, that change occurred, from the data collected. Not surprisingly, the validity evidence required to assess the appropriateness of the derived inferences differs as a function of the design elements of the study.

**Research Design**

There are three research design families commonly used when measuring change. The three common design families are: pre-experimental pre-test/post-test design, quasi-experimental pre-test/post-test design with a control group, and full experimental pre-test/post-test with a control group and randomization. Each of these design families includes measurement activities and data analysis strategies that require validity evidence to ensure that appropriate inferences are derived from the conclusion(s) of the study. In order to assess change in any research design, the measurement procedures must be capable of detecting change, and the magnitude of the change must be able to be identified and demonstrated through the appropriate statistical analysis. While the validity evidence necessary to ensure appropriate inferences are derived for studies with various types of research designs differs, the evidence for common measurement and analysis strategies used may be the same.

Research studies may measure change via TpT/pT or RpT/pT designs. In the TpT/PT design, the pre-test measurement occurs before the intervention and the post-test is administered after the intervention is completed. The desire of the researcher is to quantify the magnitude of
the observed change. Hence, change is estimated by a PT - TpT = gain score, which can be converted to an effect size (ES). A valid inference of change from this measurement strategy and data analysis is contingent upon the existence of measurement invariance in the trait across the TpT and PT measurement occasions. It is the invariance of the constructs over measurement occasions that allows for the valid interpretation of change. If the construct is not invariant, the interpretation of the ES as a structural (e.g., mean) change may not be valid (Bray, Maxwell, & Howard, 1984), thus calling into question any inferences derived about the magnitude of the observed change. The use of the retrospective pre-test (RpT) design has been proposed as an alternative to the PT. RpTs distinguish themselves from TpTs by the timeframe in which they are administered and the relationship to the PT and intervention experience. In the RpT, the PT and the RpT are administered at the end of the intervention. A variety of articles (Betz & Hill, 2005; Allen & Nimon, 2007; Moore & Tananis, 2009) have emerged in the literature that demonstrate the usefulness of RpT designs to measure of change. In some instances, the RpT provides the researcher or evaluator with a cost-effective and convenient alternative to the TpT/PT design.

**Measurement Invariance**

Given that any comparison of the same constructs across measurement occasions assumes that the measurements are factorially invariant (Little, 2013), measurement invariance is a fundamental aspect of evaluating change in a construct (Brown, 2015). According to Little (2013), measurement invariance is “probably the most important empirical question to address in any analysis that involves more than one group and/or more than one time point” and “is one of the most misunderstood concepts” (p. 137). The definition for *measurement invariance* is as follows: an observed score is considered measurement invariant if an individual’s probability of
an observed score does not depend on their group membership, conditional on their true score (Mellenburgh, 1989; Meredith, 1993; Meredith & Millsap 1992). This definition can be applied to TpT/PT or RpT/PT research designs in which case an observed score is considered measurement invariant if an individual’s probability of an observed score does not depend on the measurement occasion.

**Confirmatory factor analysis (CFA)** is a type of structural equation modeling used to identify the relationships between observed measures or indicators and latent variables or factors (Brown, 2015). The CFA framework is hypothesis driven, which makes it ideal for evaluating method effects or invariance of the factor model across measurement occasions (Brown, 2015). In the context of evaluating measurement invariance, CFA is used to assess three levels of factorial invariance: (1) configural invariance, (2) weak invariance, and (3) strong invariance. **Configural invariance** requires that the relationships between the indicator and its construct have the same pattern of fixed and free loadings at each measurement occasion. The configurally invariant model is primarily used as a baseline model to evaluate the degree to which different levels of factorial invariance are supported by the data (Little, 2013). In addition to the equality constraints of configural invariance, **weak invariance** requires equality in the loadings of the indicators across all measurement occasions. The factor loadings are not allowed to vary across measurement occasions. **Strong invariance** adds the additional requirement that the intercepts of the indicators are constrained to be equal across measurement occasions.

If measurement invariance is demonstrated (through strong invariance), establishing that the factors measure the same constructs in the same way at each measurement occasion, it is possible to compare the structural parameters of the factors, specifically the factor variances and covariances, as well as the latent means (Brown, 2015). The equality of a factor variance
examines whether the amount of variance within group (in this case within the measurement occasion) differs across the measurement occasions. Conceptually, this is the evaluation of the extent to which the different measurement occasions draw from different ranges of the underlying construct to respond to the indicators of that construct (Brown, 2015). The equality of factor means evaluates whether the levels of the underlying construct differ across measurement occasions. Comparisons of the factor variances across measurement occasions is only meaningful if the factor loadings are invariant (Brown, 2015). Comparisons of the factor covariances are meaningful if both the factor loadings and factor variances are invariant (Brown, 2016). Equality of the latent means across measurement occasions requires invariant factor loadings, and indicator intercepts.

**Statement of the Problem**

Traditional TpT/PT as well as RpT/PT designs are used to collect data on self-reported measures in order to assess the magnitude of observed change. If measurement invariance does not exist over time, for example between TpTs and PTs, it is inappropriate to compare the mean differences between these measurement occasions. Unfortunately, measurement invariance is rarely investigated in the research design, it is merely assumed. Given this, it is important to understand the extent to which TpT/PT and RpT/PT research designs demonstrate construct equivalence.

**Research Questions**

The principle aim of this dissertation study was to test if an instrument administered in a TpT/PT design evidenced the same structural meaning as the same instrument administered in an RpT/PT design. If measurement invariance exists within these two designs, then researchers and evaluators have flexibility in the design they employ. Utilizing an RpT/PT design would be
more efficient, saving time and effort and perhaps costing less than employing a TpT/PT design. The RpT/PT would also have utility when a researcher or evaluator does not have access to the population under study prior to the intervention. This study was designed to evaluate measurement invariance of a communication reticence instrument in TpT/PT and RpT/PT designs and the implications of these designs on calculations of effect sizes from which inferences about change are derived. Measurement invariance was examined through the evaluation of three hierarchal levels of factorial invariance (configural, weak, and strong). This research was designed to address two research questions. The first research question focuses on invariance between the TpT and RpT measurement occasions and with the TpT/PT and RpT/PT designs. The specific research questions were as follows:

RQ 1: What is the extent of measurement invariance in a communication reticence scale?

RQ1.1a Configural Invariance: Is the number of factors and the factor loading pattern of the TpT and RpT measurement occasions equivalent?

RQ1.1b Weak Invariance: Are the factor loadings of the TpT and RpT measurement occasions equivalent?

RQ1.1c Strong Invariance: Are the factor loadings and intercepts of the TpT and RpT measurement occasions equivalent?

RQ1.2a Configural Invariance: Are the number of factors and the factor loading pattern of the TpT and PT measurement occasions equivalent?

RQ1.2b Weak Invariance: Are the factor loadings of the TpT and PT measurement occasions equivalent?

RQ1.2c Strong Invariance: Are the factor loadings and intercepts of the TpT and PT measurement occasions equivalent?
RQ1.3a Configural Invariance: Are the number of factors and the factor loading pattern of the RpT and PT measurement occasions equivalent?
RQ1.3b Weak Invariance: Are the factor loadings of the RpT and PT measurement occasions equivalent?
RQ1.3c Strong Invariance: Are the factor loadings and intercepts of the RpT and PT measurement occasions equivalent?

The second research question is concerned with the estimation of the effect size of the latent and observed means from TpT/PT and RpT/PT research designs, and the subsequent implications for interpretation about the magnitude of change attributed to any intervention effect. The second research question was specifically as follows:

RQ2: Do the observed mean scores and the latent mean scores yield the same effect sizes in TpT/PT and RpT/PT research designs?

RQ2.1: Is the magnitude of the effect size of the observed mean and the latent mean in a TpT/PT research design is the same?
RQ2.2: Is the magnitude of the effect size of the observed mean and the latent mean in a RpT/PT research design is the same?

**Significance of the Research**

Obtaining accurate and meaningful data that reflects observed change in human service and educational program participants involves making decisions about the most appropriate research design methodology to employ given the focus of the research and the availability of resources. Ideally, research and evaluation activities are not disruptive to the program, and efficiently use time and resources, including those dedicated to data collection, analysis, and reporting (Cooke, 1998). In some cases, tight time constraints (Bamberger et al., 2004), limited
access to the population being studied, or other constraints impact the available options for research methods. Despite these constraints, the desire is always to be able to draw valid inferences about the impact of the intervention. By continuously monitoring and improving the efficacy of research designs, measurement activities, and data analysis strategies, researchers can assess the magnitude of observed change with greater precision and sensitivity (Bray et al., 1984). Researchers such as Hill and Betz (2005) and Nimon et al. (2010) have suggested that there is further need for well-designed research that evaluates the circumstances under which TpT/PT and RpT/PT research designs should be utilized. They suggest further research should be designed to utilize a measurement instrument with previously established psychometric properties (Nimon et al., 2010) and use CFA to explore and validate the factor analytic structure across each of the three measurement occasions (TpT, PT, RpT) to better understand possible sources of measurement error (Hill & Betz, 2005; Nimon et. al., 2010). This study contributes to the body of literature related to the validity inferences derived about change in research designs that utilize TpT/PT and RpT/PT measurements. It addresses areas of deficit identified by Nimon et al. (2010) and Hill and Betz (2005), including utilizing an instrument with known psychometric properties and conducting a CFA to examine the factor analytic structure across measurement occasions. Ultimately, this study will aid researchers in understanding the concerns associated with designing research that evaluates change and the how particular design choices may impact the inferences derived about change, and therefore any subsequent inferences about the impact of an intervention, derived from the study.
CHAPTER II
REVIEW OF THE LITERATURE

This chapter outlines the literature and theories pertinent to understanding the aspects of research design, measurement, and statistical analysis that underlie the validity of inferences associated with measurement of change. The literature is introduced in the context of the applicable theoretical frameworks about validity and research design, and then the chapter specifically outlines how CFA must be used to evaluate measurement invariance. Finally, examples from the literature are brought in to demonstrate the theory of response-shift bias and the implications for research designs intending to measure change.

Validity

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014), validity is currently defined as the extent to which theory and evidence support the interpretation of test scores in the context of the use for which the test was developed. Standish, Cook, and Campbell (2002) provide a similar definition, stating that validity “refers to the approximate truth of an inference” (p. 34). Given these definitions, validity is arguably the most fundamental consideration in research design, measurement, and analysis. Without the demonstration of the validity of inferences derived from any research, the findings of the study have no meaning.

The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014) have jointly published *Standards for Educational and Psychological Testing* in order to provide a framework through which validity may be established and assessed. Specifically, the purpose of the standards is to identify and explain criteria for the development and evaluation of tests and testing practices as well as to
provide guidelines for the evaluation of the validity of interpretations derived from the test based on the established use (American Educational Research Association et al., 2014). The Standards are centered on three themes or clusters: (1) Establishing Intended Uses and Interpretations, (2) Issues Regarding Samples and Settings used in Validation, and (3) Specific forms of Validity Evidence. According to the Standards, the label test is used to describe traditional tests as well as scales and inventories used to measure attitudes, interests, and dispositions, and the standards are meant to guide the evaluation of such instruments (American Educational Research Association et al., 2014). Therefore, test is used universally in the discussion of validity throughout this study.

Validity Judgments

While validity is intimately tied to the idea of truth (Standish et al., 2002), validity judgments are not absolute; various degrees of validity may be deduced through the collection of evidence (p. 35). Establishing validity requires judgments to be made. There are multiple sources of evidence that may be used to evaluate or establish the validity of an interpretation of test scores for a specified use. The various sources of evidence contribute to the overall, unitary concept of validity. Therefore, validity is established through the collection of various evidences that support the intended interpretation of the data for the proposed use. The evidence necessary to establish validity varies and is dependent on the proposition that underlies the proposed interpretation for a specific use, as well as the assumptions of the research design, measurement, and analysis techniques used to derive the test scores. A comprehensive validity argument integrates various strands of evidence into an account that explains how the evidence supports the inferences for the intended use of the test scores. Evidence may include prior research about the research design, measurement instrument and/or selected analysis strategy being used, and/or
evidence related to the study underway. Hence, establishing validity is an ongoing process, not a terminal activity. Research evidence is provided to support the inferences being drawn at that time; subsequent studies may elaborate on that existing evidence. In addition to establishing the efficacy of the findings of a study, the collection of validity evidence may also lead to the refinement of the definition of the construct, changes to the testing process, or identify additional areas for study.

Historically and for practical purposes, the unitary concept of validity was and continues to be broken up into a subset of related components, which are roughly aligned with three clusters identified in the Standards. Cook and Campbell (1979) identified four related subcomponents: statistical conclusion validity, internal validity, construct validity, and external validity. Statistical conclusion validity is concerned with the extent to which statistics have been appropriately used to infer covariation between the independent and dependent variables. Internal validity refers to whether the relationship between the independent and dependent variables resulted from a causal relationship. External validity refers to generalizations about the extent to which the inferences about the cause/effect relationship hold up over variation in people, settings, intervention variables, and measurement variables. And, construct validity evaluates the extent to which the research operations and extracted constructs were aligned. For the purposes of this study, the discussion about validity evidence and threats to validity are organized around these four subsets of the unitary concept of validity.

Understanding validity is extremely important given the implications for the effective design of research studies, particularly those that seek to understand change based on causal attribution from an intervention to an outcome. Shadish et al. (2002), identified four questions that are pertinent to researchers interpreting causal studies:
1. How large and reliable is the covariation between the presumed cause and effect?

2. Is the covariation causal, or would the same covariation have been obtained without the intervention?

3. Which general constructs are involved in the persons, settings, interventions, and observations used in the experiment?

4. How generalizable is the locally embedded relationship over varied persons, interventions, observations, and settings (p. 39)?

Clearly, these questions are interrelated; however, asking them separately helps to identify the appropriate validity evidence given the purpose of the study and inferences the researcher hopes to derive.

**Sources of Validity Evidence**

Common sources of evidence to support validity arguments include: *evidence based on test content, evidence based response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence for validity consequences of testing* (American Educational Research Association et al., 2014). *Evidence based on test content* is concerned with the relationship between the content of the test and the construct it is intended to measure. It is critically important that the content of the test is intentionally aligned with the intervention being studied. Evidence based on test content also refers to the way in which the test is structured, including the format of the items or questions and how the test is administered and/or scored. As with all sources of validity evidence, the appropriateness of the content is related to the specific inferences being derived from the test scores. Given this, if an existing test is being used in a different context or for a different purpose than for which it was originally developed, it is especially important to evaluate the content for the new purpose. Evidence about
content may be used to address differences in the meaning or interpretation of test items across subgroups or measurement occasions. In this instance, the extent to which the items are being interpreted consistently across subgroups or measurement occasions is of particular concern.

*Evidence based on response processes* is concerned with the interpretation of the cognitive processes engaged by test takers when they are responding to items. Theoretical and empirical evidence about the response process can provide information about the fit or alignment between the actual and assumed thought processes engaged while responding to items on a test. This evidence is most often derived from individual responses and/or documentation established through the observations such as eye movement or response time of test takers. Individuals may also be questioned about their thought processes and response strategies, which can yield information about the definition of the construct. A review of the variability of response patterns may lead to questions about the test format or content. Additionally, a review of differences of responses to various sections of the test in comparison to other objective measures may provide some evidence about construct alignment and/or the difference in the understanding of the vocabulary or test content, and therefore, different understandings and perspectives about the constructs being measured by subgroups, or at different measurement occasions.

*Evidence based on relationships to other variables* provides information about the extent to which the construct is related to other variables. Assessing evidence in this area is relevant when it has been determined that the items are related to other variables. This may include measuring criteria the test is expected to predict in addition to other tests hypothesized to measure the same constructs. This evidence ensures that the relationships with other variables are consistent with the constructs underlying the proposed test score interpretations. If the evidence is intended to assess the same or similar constructs, it is referred to as *convergent*
evidence. If the evidence is intended to measure different constructs, it is called *divergent evidence*.

*Evidence based on internal structure* identifies the extent to which the test items and components conform to the construct being measured. This source of evidence is concerned with the conceptual framework of the test and the how the item interrelationships represent the assumptions of the framework. The evidence used to establish this aspect of validity is dependent on how the test is being used. Some studies, such as the current study, are focused on the consistency of measurement over time or across subgroups. This study attempts to meet appropriate validity standards to ensure the fidelity of the research, but also to contribute to the literature with regard to evidence based on internal structures.

**Research Design**

An *experiment* is defined as “a study in which an intervention is deliberately introduced to observe its effects” (Standish et al., 2002, p. 12). There are four different kinds of studies that are commonly recognized in educational and social science research: *randomized experiment*, *quasi-experiment*, *natural experiment*, and *correlational studies*.

**Descriptions of Experiments**

*Natural experiments* have a “naturally” occurring contrast between the intervention and a comparison condition (Fagan, 1990; Meyer, 1995; Standish et al., 2002). In these studies, researchers are often examining the phenomena of interest retrospectively; hence interventions cannot often be manipulated, and plausible inferences have to be constructed through data collection about known occurrences.

In *correlational studies*, also known as *passive observational designs* or *non-experimental designs*, the cause and effect or relationship between the intervention and the
outcome are presumed, measured, and studied, but not manipulated. Given this, the design elements of the study do not allow the researcher to develop counterfactual inferences. Instead, alternative explanations are usually measured individually and tested statistically. With this type of study, it is difficult to establish causation, particularly when alternative interpretations are not known prior to measurement and data collection, a difficult condition to meet.

In a randomized experiment, interventions are assigned to experimental units (people) by chance with the goal of obtaining two or more groups that are probabilistically similar to each other on the average so that the observed differences between the groups at the end of the study may be attributed to the intervention, establishing causal inferences (Standish et al., 2002). The randomized experiment is often referred to as the gold standard for studies concerned with evaluating the outcome of an intervention because there are stronger assurances that the counterfactuals can be controlled; hence, a greater opportunity to infer causal attributions.

Quasi-experiments have the same purpose and most of the same attributes as randomized experiments; however, by definition, they lack random assignment. Instead, assignment to conditions is by self or administrative selection. In quasi-experiments, the intervention can be manipulated and occurs before the effect is measured. Quasi-experimental design features create less compelling support for counterfactual inferences (Standish et al., 2002). Control groups may differ in systematic ways (other than the presence of the intervention), which causes concern that the systematic differences may be alternative explanations for the intervention effect, and therefore, must be ruled out in order to derive valid inferences about the effect of the intervention. In quasi-experimental designs, while the goal is still to draw causal attributions about and intervention, this is not possible. Instead, the experimenter must use logic, design, and measurement to assess the potential to derive valid inferences about the effect of the
Research Designs and Measuring Change

Research studies measuring change primarily utilize experimental or quasi-experimental pre-test/post-test research designs. The three most common designs are: (1) quasi-experimental pre-test/post-test designs without control groups, (2) quasi-experimental pre-test/post-test designs with control groups, and (3) experimental pre-test/post-test designs with a control group and randomization. Measuring constructs using any of these designs requires careful consideration and planning in order to ensure the validity of inferences derived from the study. Key considerations include the operational definition of the construct (construct under or over representation), research design methodologies (such as sampling), and measurement design.

Quasi-experimental design is commonly used in social science and educational research to measure change based on an intervention. There are many reasons why quasi-experimental designs may be used. In cases where the requirements for experimental design such as a control group and random assignment are not possible due to ethical concerns, timing of access to the population of study, or lack of resources, a quasi-experimental design is desirable. A causal inference derived from a quasi-experiment must meet the basic requirements for all causal relationships; that is, cause (intervention) must precede effect (outcome), cause must covary with effect, and alternative explanations for the causal relationship must not be plausible. There are three principles used in quasi-experimental design to show that alternative explanations for the impact of the intervention are not plausible.

First, identification and study of plausible threats to internal validity is required. When threats to internal validity are known, they can be studied to identify the extent to which it is likely that they explain intervention-outcome covariation. The second principle is the primacy of
control by design. This principle is primarily concerned with using research design elements to prevent the confounding of a threat to validity with intervention effects or to provide evidence about the plausibility of those threats. An alternative to design controls is statistical controls that attempt to remove confounds from effect estimates using statistical adjustments after data has been collected. The relationships between the research design and statistical controls must all be considered in the design of the research study. The third principle for reducing the plausibility of alternative causal explanations in quasi-experimentation is coherent pattern matching. In coherent pattern matching, a complex predication is made about a causal hypothesis that that few alternative explanations could match. The more complex the pattern that is successfully predicted, the less likely that alternative explanations could generate the same conclusion, increasing the likelihood that the change is related to the intervention. This design method is often used when resources need to be devoted to concerns related to validation of constructs and external validity, or to address practical issues such as funding, ethics, and timing of access to the population under study such as when the intervention being studied began prior to the initiation of the study.

There are varying designs associated with quasi-experiments aimed at measuring change. In the one-group post-test only design, a single group is studied only once after an intervention. This design is problematic because there is no comparison or contrast or indication of difference, which would facilitate interpretation of the findings, i.e., change. This lack of evidence makes it very difficult, if not impossible to interpret change. Instead, the evidence produced creates a baseline or benchmark of evidence, which is implicitly compared with other events that have been casually observed and remembered. Hence, the inferences observed are based on general expectations about what the data would have been, if the intervention had not occurred. Of
additional concern is the lack of standardization with which future data may be collected and compared to the baseline data. Without consistency in the measurement and data collection process, comparison may not be valid.

The one-group pre-test/post-test design is characterized by pre- and post-measures of an intervention. Often called a *within-subjects design*, this single-cohort longitudinal design allows the researcher to evaluate change. In this design, the desire of the researcher is for observed change to represent the change due to the intervention; however, this change is confounded with time-of-measurement effects. Estimates of the constructs’ means, variances, or correlations in the sample at the measurement occasion may reflect the experiences of the cohort, leading to questions about the inferences derived from the findings of the study. There are design elements that can be added to the one-group pre-test/post-test design to improve the validity of causal inference. These strategies include using a double pre-test, using a nonequivalent dependent variable, adding a removed intervention, and adding a repeated intervention (with or without a retrospective pretest). In designs using self-reported measures, adding an objective measure to the design can help contribute to the evidence used to establish the validity of the inferences derived from the findings.

**Threats to Validity in the Context of Research Design**

The strength of a research study designed to measure change is determined by the extent to which casual inference is demonstrated. Although most research is localized, it is the desire of the researcher to connect experimental results to theories with broad conceptual applicability, which requires generalization of the linguistic level of the constructs rather than at the level of operations used to represent the constructs in the experiment (Standish et al., 2002). Perceptual and cognitive stability that is fostered by generalizations is valued by humans (Standish et al.,
2002). However, a conflict exists between the localized, specific inferences derived from individual studies and the generalized goals the research hopes to attain. Cronbach (1982) describes how each experiment consists of (1) units that receive the experiences being contrasted, (2) the intervention, (3) the observations made on the unit (outcomes), and (4) of the settings in which the study was conducted; essentially the actual people, interventions, measures, and settings sampled in the experiment. Cronbach and his colleagues identified two problems with generalization: (1) generalizing to the domain about which the question is being asked, and (2) generalizing to units (persons), interventions, variables, and settings not directly observed. Hence, in order to derive inferences that are deemed valid and generalizable, the aforementioned elements of the research design must be well conceived and the consequences for decisions made in each element must be known in the design phase of the research.

The threats to the validity have historically been described in four categories: (1) threats to statistical conclusion validity; (2) threats to external validity; (3) threats to construct validity; and (4) threats to internal validity. This framework is used to describe the considerations for the evidence needed to support the inferences derived from various research designs.

**Threats to statistical conclusion validity.** Statistical conclusion validity is concerned with two related statistical inferences that affect covariation of causal inferences: (1) Do the presumed cause and effect covary? (2) To what extent or the magnitude is the covariation? Type I or type II errors may occur with regard to the first inference, and over or underestimation of the magnitude of the covariation and/or the degree of the confidence about that magnitude warrants is relevant to the second inference. Specifically, Standish et al. (2002) identified the following nine threats to statistical validity presented in Table 1 (p. 45).
Table 1

Threats to Statistical Conclusion Validity

1. **Low statistical power**: An insufficiently powered experiment may incorrectly conclude that the relationship between treatment and outcome is not significant.
2. **Violated Assumptions of Statistical Tests**: Violations of statistical test assumptions can lead to either overestimating or underestimating the size and significance of an effect.
3. **Fishing and the Error Rate Problem**: Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
4. **Unreliability of Measures**: Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
5. **Restriction of Range**: Reduced range on a variable usually weakens the relationship between it and another variable.
6. **Unreliability of Treatment Implementation**: If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.
7. **Extraneous Variance in the Experimental Setting**: Some features of an experimental setting may inflate error, making detection of an effect more difficult.
8. **Heterogeneity of Units**: Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
9. **Inaccurate Effect Size Estimation**: Some statistics systematically overestimate or underestimate the size of the effect.

**Threats to external validity**. External validity is concerned with how causal relationships hold up over variations in persons, settings, treatments, and outcomes. Estimates of the extent to which a causal relationship holds over these variations are similar conceptually to tests of statistical interactions. Standish et al. (2002) identified five threats to external validity presented in Table 2 (p. 87).
Table 2

Threats to External Validity

1. **Interaction of the Causal Relationships with Units**: An effect found with certain kinds of units might not hold if other kinds of units have been studied.

2. **Interaction of the Causal Relationship Over Treatment Variations**: An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.

3. **Interaction of Causal Relationships with Outcomes**: An effect found on one kind of outcome observation may not hold if other outcome observations are used.

4. **Interactions of the Causal Relationship with Settings**: An effect found in one kind of setting may not hold if other kinds of settings were to be used.

5. **Context-Dependent Mediation**: An exploratory mediator of a causal relationship in one context may not mediate in another context.

**Threats to construct validity.** Threats to validity related to inferences about constructs are concerned with the match between the study operations and the constructs used to describe the operations. There may be issues with explication of the constructs or the sampling or measurement design. The construct may also be under or over represented in the study.

Standish et al. (2002) identified several threats to validity of inferences about constructs that have been shown to occur frequently in research, as presented in Table 3 (p. 73).

Table 3

**Construct Validity**

1. **Inadequate Explication of Constructs**: Failure to adequately explicate a construct may lead to incorrect inferences about the relationship between operation and constructs.

2. **Construct Confounding**: Operations usually involve more than one construct, and failure to describe all the constructs may result in incomplete construct inferences.

3. **Mono-Operation Bias**: Any one operationalization of a construct both under represents the construct of interest and measures irrelevant constructs, complicating inference.

4. **Mono-Method Bias**: When all operationalizations use the same method (e.g., self report), that method is a part of the construct actually studied.

5. **Confounding Constructs with Levels of Constructs**: Inferences about the constructs that best represent study operations may fail to describe the limited levels of the construct that were actually studied.
<table>
<thead>
<tr>
<th></th>
<th><strong>Intervention Sensitive Factorial Design:</strong> The structure of a measure may change as a result of intervention, change that may be hidden if the same scoring is always used.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Reactive Self-Report Changes:</strong> Self-reports can be affected by participant motivation to be in intervention condition, motivation that can change after assignment is made.</td>
</tr>
<tr>
<td></td>
<td>** Reactivity to the Experimental Situation:** Participant responses reflect not just interventions and measures, but also participants’ perceptions of the experimental situation, and those perceptions are a part of the intervention construct actually tested.</td>
</tr>
<tr>
<td></td>
<td><strong>Experimenter Expectancies:</strong> The experimenter can influence participant responses by conveying expectations about desirable responses, and those expectations are a part of the construct actually tested.</td>
</tr>
<tr>
<td></td>
<td><strong>Novelty and Disruption Effects:</strong> Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine, a response that must then be included as a part of the intervention construct description.</td>
</tr>
<tr>
<td></td>
<td><strong>Compensatory Equalization:</strong> When intervention provides desirable goods or services, administrators, staff, or constituents may be provide compensatory goods or services to those not receiving intervention, and this action must then be included as part of the intervention construct description.</td>
</tr>
<tr>
<td></td>
<td><strong>Compensatory Rivalry:</strong> Participants not receiving intervention may be motivated to show they can do as well as those receiving intervention, and this compensatory rivalry must then be included as a part of the intervention construct description.</td>
</tr>
<tr>
<td></td>
<td><strong>Resentful Demoralization:</strong> Participants not receiving a desirable intervention may be so resentful or demoralized that they may respond more negatively than otherwise, and this resentful demoralization must be included as part of the intervention construct description.</td>
</tr>
<tr>
<td></td>
<td><strong>Intervention Diffusion:</strong> Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.</td>
</tr>
</tbody>
</table>

**Threats to internal validity.** Internal validity refers to the minimum (design structure and attributes) without which any experiment is uninterpretable (Campbell & Stanley, 1963). The term internal validity refers to the inferences about whether the observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were measured (Shadish et al., 2002). Causal order can be difficult to establish in quasi-experimental design. Threats to internal validity are other plausible causes to believe that the relationship from A to B is not causal. Shadish et al. (2002) identified nine threats to internal validity, as presented in Table 4 (p. 55).
Table 4

Threats to Internal Validity

1. **Ambiguous Temporal Precedence**: Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. **Selection**: Systematic differences over conditions in respondent characteristics that could also cause the observed effect.
3. **History**: Events occurring concurrently with treatment could cause the observed effect.
4. **Maturation**: Naturally occurring changes over time could be confused with a treatment effect.
5. **Regression**: When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with treatment effect.
6. **Attrition**: Loss of respondents to treatment or to measurement can produce artificial effects if that loss is systematically correlated with conditions.
7. **Testing**: Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confounded with a treatment effect.
8. **Instrumentation**: The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.
9. **Additive and Interactive Effects of Threats to Internal Validity**: The impact of a threat can be added to that of another threat or may depend on the level of another threat.

This study is primarily concerned with measurement invariance in TpT/PT and RpT/PT research designs given the phenomena of response-shift effect. Specifically, it is concerned with examining the extent to which the following threats to validity are of concern in these types of research designs: (a) violated assumptions of statistical tests; (b) inaccurate effect size estimation; (c) inadequate explication of constructs; (d) intervention sensitive factorial design; and (e) instrumentation.

This study is primarily concerned with threats to validity that result in measurement invariance across measurement occasions. Fundamentally, measurement invariance is about *instrumentation*, the threat to validity that is concerned that the nature of a measure may change over time in a way that could be confused with a treatment effect. The concept of measurement invariance is specifically concerned with estimation of effect sizes in TpT/PT and RpT/PT research designs; hence, the *violated assumptions of statistical tests* and *inaccurate effect size*
estimation threats to validity are both relevant. If the statistical assumption of measurement invariance is not met in calculations of mean differences across measurement occasions, the violated assumptions of statistical tests threat to validity will not be addressed. The second consideration this study addresses related to these two threats to validity is the extent to which simple statistics such as a dependent t-test, which calculates the difference between observed scores at each measurement occasion, yields an effect size estimate that represents the magnitude of the difference that exists in the latent trait. Measurement invariance is also concerned with the stability of constructs over time; hence inadequate explication of constructs and intervention sensitive factorial design are both relevant to this study. Measurement invariance at the levels of configural, weak, and strong have varying meanings about the structural stability of the measures over time and implications for the inferences that may be derived about the change that occurred as a result of an intervention.

Measurement Invariance

Measurement invariance is a fundamental aspect of evaluating change in a construct (Brown, 2015) using TpT/PT and TpT/RpT research designs. According to Little (2013), factorial invariance is “probably the most important empirical question to address in any analysis that involves more than one group and/or more than one time point” and “is one of the most misunderstood concepts” (p. 137). Any comparison of the same constructs across times or groups assumes that the measurement occasions are factorially invariant (Little, 2013), but it is rarely known if invariance exists at the time of the measurement occasion or statistical analysis. It is this lack of knowledge about the invariance across the measurement occasions, which is the cause for the questions about the validity.

Brown (2006) discusses three types of change, identified in previous literature, which
may be encountered in TpT/PT and RpT/PT research designs: alpha, beta, and gamma change. Alpha change is the true score change in a construct given a constant conceptual domain and constant measurement (measured change in the latent means a post-pre change is interpreted as a rise of fall in the latent construct amount) (Brown, 2015). Alpha change in TpT/PT or RpT/PT research designs occurs when measurement invariance exists between measurement occasions. Beta change occurs when the construct remains constant but the measurement properties of the indicators of the construct are inconsistent across the measurement occasions (Brown, 2015). In this instance, the numerical values across the measurement occasions are not on the same measurement scale, (e.g., a post-pre change cannot be interpreted as a rise or fall in the latent construct amount but this inequality can be corrected). Gamma change occurs when the meaning of the construct changes between measurement occasions (e.g., a post-pre change cannot be interpreted as a rise or fall in the latent construct amount and this inequality cannot be corrected), for example, the number of indicators that represent a factor changes (Brown, 2015). When measurement is not invariant over measurement occasions, it is potentially misleading to analyze and interpret the change in the observed measures to represent a change in the latent constructs. Given this, it is necessary to examine measurement invariance prior to even simple analyses such as repeated measures t-tests.

**Definition of Measurement Invariance**

A definition for measurement invariance (MI) that has been provided by many researchers (Mellenburgh, 1989; Meredith, 1993; Meredith & Millsap, 1992) is that an observed score is considered measurement invariant if an individual’s probability of an observed score does not depend on their group membership, conditional on the true score. This definition can be extended to the context of TpT/PT or RpT/PT research designs in which case an observed score
is considered measurement invariant if an individual’s probability of an observed score does not depend on the measurement occasion. A statistical definition for measurement invariance is:

The observed random variable $Y$ is said to be measurement invariant with respect to selection on $G$, if $F\left(y \mid \eta, g\right) = F\left(y \mid \eta\right)$ for all $(y, \eta, g)$ in the sample space, where $Y$ denotes an observed random variable with realization $y$; $H$ denotes the latent variable (i.e., factor) with realization $\eta$ that is measured by $Y$, or underlies $Y$; $G$ denotes a random variable with realization $g$ that functions as a selection of a subpopulation from the parent population by application of a selection function $s(g)$, $0 \leq s(g) \leq 1$. (see Meredith, 1993, p. 528)

**Confirmatory Factor Analysis**

*Confirmatory factor analysis* (CFA) is a type of structural equation modeling used to identify the relationships between observed measures or indicators and latent variables or factors (Brown, 2015), and is a principle analytical method for investigating invariance. The CFA framework is hypothesis driven, which makes it ideal for evaluating methods effects or invariance of the factor model across measurement occasions or research confederates (Brown, 2015). According to Brown (2015), "the fundamental intent of factor analysis is to determine the number and nature of the latent variables or factors that account for the variation and covariation among a set of observed measures, commonly referred to as indicators" (p. 10). The common factor model (CFM) identified by Thurstone (1947), lays the conceptual framework for CFA.

The CFM postulates that in a set of observed measures, each indicator is a linear function of one or more common factors and one unique factor (Brown, 2015). Hence, all CFA models have factor loadings, unique variances, and factor variances. Unique variance is typically
presumed to be measurement error. CFA models may include error covariances, which suggests that the two indicators covary for reasons other than the shared influence of the latent factor. In an unstandardized solution, a factor variance expresses how similar or different the sample participant’s relative standing on the latent trait is from the other participants (the dispersion of the factor). When a CFA solution consists of two or more factors, a factor covariance is usually specified to estimate the relationship between the latent dimensions.

In the traditional conventions of factor analysis, the latent factor is depicted by a circle or oval, and the ratings or indicators (items) are represented as squares or rectangles, as shown in Figure 1. Unidirectional arrows represent the factor loadings (*lambda*), which are regression slopes that predict the indicators from the latent factor. The arrows are used to relate the unique variances (*epsilon*) to the indicators. Matrices are represented in factor analysis by upper case Greek letters, and specific elements of the matrices are denoted by lowercase letters.

![Figure 1. Simple CFA model.](image-url)
In CFA analysis, the researcher specifies the number of factors and the pattern of indicator-factor loadings in advance. At that time, pertinent information is identified, such as the independence or covariance of the factors and indicator unique variances (Brown, 2015). The factor solution that is assumed based on a priori knowledge is evaluated in terms of how well it produces the sample correlation (covariance) matrix of the measured variables. Given this, CFA requires a strong empirical or conceptual foundation to guide the specification and evaluation of the factor model (Brown, 2015).

**CFA model identification.** Estimation of the CFA model is based on the model-data fit; that is, how well the model imposed structure estimates the observed variance-covariance pattern. Model imposed goodness of fit is determined by how adequately both the measurement and structural components in complex models are specified. Given this, a key aspect of CFA evaluation is the parameters of the measurement model to reproduce the observed relationships among the indicators (Brown, 2015). Hence, in order to estimate the parameters in CFA, the measurement model must be identified. A model is identified if it is possible to obtain a unique set of parameter estimates for each parameter in the model whose values are unknown on the basis of known information (Brown, 2015). Model identification includes the difference between the number of freely estimated model parameters and the number of pieces of information in the input variance-covariance matrix. Brown (2015) provided the following guidelines for model identification:

1. Regardless of the complexity of the model, latent variables must be scaled by either specifying the marker indicators or fixing the variance of the factor (usually to a value of 1).

2. Regardless of the complexity of the model, the number of pieces of information in the input matrix must equal or exceed the number of the freely estimated model parameters, e.g.,
positive \textit{d.f}.

3. In the case of one-factor models, a minimum of three indicators is required to evaluate model-data fit. When three indicators are used, the one-factor solution is just-identified and goodness of fit evaluation does not apply, although this model can still be evaluated in terms of the interpretability and strength of its parameter estimates. When four or more indicators are used, the model is over identified and goodness of fit can be used in the evaluation of acceptability of the solution.

4. In the case of models that entail two or more factors and two indicators per latent construct, the solution will be over identified, provided that every latent variable is correlated with at least one other latent variable and the errors between indicators are uncorrelated. However, because such solutions are susceptible to empirical underidentification, a minimum of three indicators per latent variable is recommended.

\textbf{CFA sample size.} Varying recommendations and guidelines for the necessary sample size for confirmatory factor analysis have been proposed over the years. These guidelines are typically stated in terms of the minimum necessary sample size, \( N \), or the maximum ratio of \( N \) to the number of variables being analyzed, \( p \) (MacCallum et al., 1999). Historically, recommendations for absolute minimum sample size for confirmatory factor analysis have ranged from 100 to 250 (MacCallum et al., 1999). Recommendations for the \( N:p \) have been made for the ratio to be anywhere from three to 10 (MacCallum et al., 1999). However, in CFA there are two components: (1) the stability of the variance-covariance matrix of the indicators (e.g., the stability of the correlation matrix of the indicator variables) and (2) the model (imposed) structure must be over identified, e.g., positive degrees of freedom. These two components dictate the sample size necessary to complete the CFA analysis.
**Estimation of CFA model parameters.** Maximum likelihood is by far the most widely used estimator in CFA research (Brown, 2015). Key assumptions of ML include that (1) the sample size is sufficiently large; (2) the indicators have been measured on continuous scales, or approximate continuous data; and (3) the distribution of the indicators is multivariate normal (Brown, 2015). When the data are non-normal, ML analysis can result in biased standard errors and poor model fit; thus, it is preferred to use robust maximum likelihood (MLR), particularly with small and medium sample sizes (Brown, 2015). Robust estimation in MLR still assumes the data follow a multivariate normal distribution. But also that the data have more or less kurtosis than would otherwise be common in a normal distribution. Under MLR, model fit statistics are adjusted based on an estimated scaling factor. Using MLR in MPLUS has the advantage that if the data happen to be normally distributed, no adjustment is made.

**Measurement Invariance Models**

Longitudinal measurement models are best evaluated using measurement and structural invariance models, hence this study was designed to utilize a within group confirmatory factor analysis to evaluate configural, weak, and strong measurement invariance, and if appropriate, structural invariance (equality of the means, variances, and covariances) between the traditional pre-test and the retrospective pre-test. Different levels of factorial invariance hold if the change in the model fit from a lower level of invariance to a higher level of invariance is not statistically significant.

**Configural invariance.** The first step in analysis of invariance across measurement occasions is to establish that the same factor structure is present. This is known as *equal form* (Brown, 2015), or *configural invariance* (Little, 2013). At this step, the relationship between the indicators and the construct should have the same pattern of fixed and free loadings at each
measurement occasion—all estimates should occur at both measurement occasions, and the fixed 0 (nonestimated parameter) should be the same at both measurement occasions. No constraints are placed on any of the parameter estimates, and only the same pattern of factor loadings is expected to be the same (Little, 2013). Correlated errors are specified if it is anticipated that covariance will exist between repeated measures (Brown, 2015). Evaluation criteria for whether configural invariance holds are: (1) does the pattern look the same, and (2) basic model fit statistics (Little, 2013). A key feature of the configurally invariant model is that each element of the SEM equation is estimated uniquely for each measurement occasion. In the fixed factor method, the variance of the constructs are fixed to 1.0. This is the baseline model used to evaluate the extent to which the different levels of factorial invariance are supported by the data. Evidence of the equal form (cf. gamma change) or configural invariance allows for the subsequent tests of measurement invariance to proceed.

**Weak factorial invariance.** The next analysis, known as *weak factorial invariance*, tests the equality of factor loading over the measurement occasions. Testing for weak factorial invariance involves the same parameters as configural invariance, with the additional specification that the factor loadings from the separate measurement occasions are mathematically equal. The intercepts are freely estimated at each point. When the loadings are equated across time, the scaling constraint at the second measurement occasion is no longer needed, regardless of the scaling method. Constraints at subsequent measurement occasions are no longer needed because the scale for the estimates of the loadings and the variance of the constructs at those measurement occasions are determined by the scaling constraint that is in place at the first measurement occasion. Constraining the factor loadings to be equal across the measurement occasions standardizes the metric of the latent constructs variances so that they are
operationally defined in the same way. This means that the relative strength of each indicator is the same across measurement occasions, which essentially forces the loadings to be estimated as the optimal balance within and across the measurement occasions. When the loadings are mathematically equal across time, the scale for interpreting the variance and covariance estimates is the same. Another way of saying this is that the estimates of the factor loadings are now based on a common metric. Using the fixed factor method of scaling, the fixed variance at the second measurement occasion is no longer needed to set the scale for the factor loadings, and instead is a freely estimated parameter (Little, 2013). Evidence of beta change or weak invariance allows for the subsequent test of measurement invariance to proceed. This model becomes the model to which the subsequent model (alpha) will be compared.

**Strong factorial invariance.** Testing for strong factorial invariance is similar to testing for weak factorial invariance; however the focus is on the observed means and estimated intercepts of the indicators (Little, 2013). Each corresponding indicator is specified to be mathematically equal across measurement occasions. When the fixed factor method of scaling is used, the means of the latent constructs are fixed at zero to provide the scale for the mean level information (intercepts and latent means) at both measurement occasions. Because the latent mean at the first measurement occasion is fixed to zero, the difference between the first measurement occasion and the subsequent measurement occasion is the difference from zero. Evidence of strong invariance is also known as *alpha change.*

**Validity of Self-Reported Measures**

**Validity Concerns With Self-Reported Measures**

Klatt and Taylor-Powell (2005) and Sibthorp et al. (2007) caution that self-reported measurements in both RpT/TP and TpT/PT research designs are prone to bias. Findings from
studies concerned about the validity of self-reported measures suggest that when measurement error exists in TpT and RpT ratings, it is greatly influenced by several contextual factors including time interval between TpT and PT, item and anchor wording, the target of recall, motivational and personality factors, instructions given at the time of administration, the attitude of the individual administering the questionnaire, and the content or focus of the experience or intervention (Betz & Hill, 2005). Other theories generated from contextual factors associated with self-reporting include personal recall theory (Schwartz & Rapkin, 2004; Schwarz, 2007), which shows that respondents tend to use estimation strategies instead of recall and count methods, and social desirability bias (Krosnick, 1999), which leads to underreporting of undesirable traits or beliefs and over reporting of socially desirable traits. Another similar theory includes acquiescence (Krosnick & Fabrigar, 1998), which is hypothesized to lead to inflated responses. Impression management theory (Paulhus, 2002; Tedeschi, Schlenker, & Bonoma, 1971) suggests that respondents believe that the appearance of improvement will make them look good and report in a way that demonstrates improvement, regardless of their actual experience. Cognitive dissonance theory suggests that if participants do not perceive they gained anything from an intervention, they may reconstruct their experiences to avoid feeling as though they wasted the time or effort they invested (Hill & Betz, 2005).

Lam and Bengo (2003) identified three potential sources of bias in TpT/PT research designs: carryover, interference, and sensitization. Carryover effect is concerned with the potential for respondents to correct for errors they made at the TpT measurement occasion. Interference effect hypothesizes that there is a negative bias in the PT ratings as a result of repetitive testing. Sensitization suggests that the PT measurement exposes the participants to the intervention, highlighting particular aspects, which may infer importance. Each of these theories
is concerned with the aforementioned threats to validity about effect size estimates, violated assumptions of statistical tests and inaccurate effect size estimation.

**Response-Shift Theory**

Response-shift theory has been identified as one explanation for construct invariance that can lead to the over or underestimation of intervention effects. This phenomena was first brought to light in the 1970s and 80s by Howard and his colleagues when they reported a variety of studies that demonstrated that the post-test scores of a training were significantly lower than the pre-test scores (Howard & Dailey, 1979; Howard et al., 1979). Howard, Ralph et al. (1979) laid the preliminary foundations for response shift theory through an evaluation of an Air Force communication skills training program aimed at reducing dogmatism. In this study, a TpT/PT research design was used to measure change in self-reported dogmatism. Findings indicated an apparent decrease in dogmatism after participating in the training; however, Howard and his colleagues subsequently interviewed workshop participants and concluded that their perceived level of dogmatism that was initially reported (prior to the workshop) was not accurate. The interviews revealed that participants had overestimated their perceived levels of dogmatism prior to the workshop, and hence it was determined that the initial inferences that dogmatism had decreased as a part of the training were not valid.

Howard, Ralph et al. (1979) replicated this first study using a TpT/PT for one group, and in another group using the RpT research design. In the second study, Howard, Ralph et al. (1979) found that the difference between the TpT and RpT scores reflected a decrease in dogmatism, corroborating the qualitative data gathered in the first study, whereas no difference was found between the TpT and PT scores. In a third study, aimed at studying stereotypical gender roles and androgyny, Howard, Ralph et al. (1979) again employed a TpT, PT, RpT
research design. In this study, women were also randomly assigned to control or experimental groups. The intervention was designed to promote androgyne through the development of skills typically stereotyped as masculine. In order to monitor the effectiveness of the intervention, self-report and objective measures of assertiveness, sex-role orientation, and attainment of individual goals were obtained. The same measures were taken with both the experimental and control groups, allowing for comparisons. Findings suggested that the effects of response shift were evident for the treatment subjects, but not for the control groups. The objective measures of change correlated more highly with PT/RpT than the TpT/PT measures of change, suggesting that the PT/RpT research design would result in inferences that are more valid than the TpT/PT research design. Given the findings from these studies, Howard and his colleagues theorized that participants develop awareness and judgments of their behavior both prior to and after the intervention as a function of the knowledge gained during the intervention. More specifically, their understanding of the construct changed as a function of the intervention, not just the level of the construct.

In an attempt to further explore response shift in the context of cognitive variables, Bray, Maxwell, and Howard (1984) conducted a study about the perception of change in knowledge of learning theory in an educational psychology class. For this study, a TPT/RpT/PT research design was used (both traditional and retrospective pre-tests were administered). An objective measure (an essay exam) was also employed. The findings demonstrated higher correlation between the RpT and the objective measure than between the objective measure and both the TpT and PT.

Based on this research response shift theory (sometimes now called response shift bias) was developed. The theory of response shift challenges the assumption that in TpT/PT research
designs in which self-reported change is being measured, a respondent’s conceptual understanding of the construct being assessed remains stable across measurement occasions. If the understanding of the construct changes, the PT ratings reflect this shift in construct understanding in addition to the changes that might have occurred from participation in an intervention. This change is called a response-shift effect (Bray & Howard, 1984). Response-shift effect has been defined as the “the treatment group subject’s awareness of treatment relevant constructs or dimensions” (Bray & Howard, 1984) and the “recalibration of the participant’s internal metric from the beginning to the end of the program” (Sibthorp et al., 2007). Many researchers have argued that response-shift effect is an unavoidable consequence or even goal of experiences and interventions, given that in many instances the primary purpose of the intervention itself is to change one’s awareness or understanding of particular construct(s). Under this assumption, if an intervention meets this goal, it will necessarily alter one’s internalization of the construct and its attributes or indicators; thus, inferences derived from the measurement of change will not be valid due to response-shift effect.

Since the development of the theory of response shift effect, numerous studies have been conducted in order to further explore this phenomena. This research elaborates on the difference in scores associated with RpT and TpT measurement, leading to hypotheses that support the theory of response shift effect and consequently further elucidate the validity of both TpT/PT and RpT/PT research designs in various contexts.

In a study with psychology students, Hoogstraten (1982) found that the RpT/PT research design yielded a larger effect size estimate than the TpT/PT research design, and that the RpT responses were better aligned with the objective measure administered than the TpT. Sprangers and Hoogstraten (1989) did find that a behavioral pre-test that requires actual performance by
study participants can reduce response-shift in TpT/PT research designs. Other evidence suggests that even the promise that an objective measure will be administered may impact response shift bias (Spangers & Hoogstraton, 1987, 1988b). Aiken and West (1990) contributed to the body of literature about RpTs by suggesting their use to detect bias in self-report measures, including experience limitation, condition justification, altered states, and self-presentation. Gutek and Winter (1992) concluded that response shift bias is a threat to validity over time, and suggested using RpT measurements as a way to address threat when measuring attitudes. Skeff et al. (1992) also found that RpT measurements were more closely aligned with objective measures.

Manthei (1997) administered RpT, TpT, and PT measures to evaluate master’s level counselors, and conducted follow up interviews with participants based on three response patterns: (1) individuals with RpT and TpT scores that were almost identical; (2) individuals with scores higher on the TpT than the RpT; and (3) individuals with lower scores on the TpT than the RpT. The interviews revealed a difference in the relationship between the level of knowledge subjects had about the subject matter prior to taking the class. The individuals from the first group (TpT and RpT equal) reported that they already had a high level of knowledge about the subject matter prior to beginning the course and that their experience reinforced that knowledge. The individuals from the second group (TpT higher than RpT) identified that they systematically overestimated their beginning knowledge and skill level and that the course helped them see their deficiencies more clearly. These findings seemed to be reinforced by Sibthorp et al. (2007), who found that if a participant’s understanding of the constructs being studied are well established and stable, the underlying metric will not change and response shift bias is not likely to occur (Sibthorp et al., 2007).
In order to evaluate the effectiveness of the Oregon Healthy Start (OHS) program, Pratt et al. (2000) used an RpT design as one of many measures, and found that the RpT/PT measurement yielded improvement on all seven indicators (all), whereas the TpT/PT measurement only indicated improvement on four of seven indicators. Cantrell (2003) found that in an assessment of student teacher self-efficacy beliefs, students indicated that prior to teaching students, they felt more confident in their ability to explain concepts to students than when they actually started teaching. The individuals from the third group (TpT lower than RpT) indicated that the course had given them a greater understanding of their beginning skill level. In a study about gender and self-efficacy, Moore and Tananis (2009) found a significant effect size using both TpT/PT and RpT/PT research designs, but also reported that the RpT scores were significantly lower than the PT scores. Around that same time, Pelfrey and Pelfrey (2009) advocated for the use of RpT/PT research designs to control for response-shift effects.

As RpT/PT research designs were increasingly being seen as a way to address concerns about response shift effect, researchers began to speculate about how subjects respond to retrospective questions. Taylor, Russ-Eft, and Taylor (2009) suggested that respondents start with an inflated sense of change in the PT measurement and from that point, retrospectively lower their RpT score in order to maintain consistency with their assumption of change. Some researchers have cautioned that implicit theory of change (Lam & Bengo, 2003; Norman, 2003), and the flawed nature of personal recall can lead participates to assume positive change based on interventions (Ross, 1989). The RpT measures may also be biased if the length and specificity of time period time affects the recall process (Pratt, McGuigan, & Katzev, 2000).

These studies have demonstrated that there is cause for concerns about validity in TpT/PT and RpT/PT research designs given that in some circumstances there are differences in
the effect size estimates that result from these two design options. Because the studies did not evaluate measurement invariance over time prior to completing the analysis of the difference of the observed mean scores, there are several threats to validity that were not addressed. As was previously noted, establishing measurement invariance through the level of strong is necessary in order to address the following threats to validity: violated assumptions of statistical tests, inaccurate effect size estimation, inadequate explication of constructs, and intervention sensitive factorial design.
CHAPTER III
METHODOLOGY

Research Questions

This study was designed to evaluate measurement invariance of a communication reticence instrument in TpT/PT and RpT/PT research designs, and the implications of these two research designs on the effect size estimates derived through the analysis of change. Measurement invariance was examined through the evaluation of three hierarchal levels of factorial invariance (i.e., configural, weak, and strong). This study was designed to answer two research questions by altering the traditional pre-post-test design by adding an RpT. The specific research questions were as follows.

RQ 1: What is the extent of measurement invariance in a communication reticence scale?

RQ1.1a Configural Invariance: Is the number of factors and the factor loading pattern of the TpT and RpT measurement occasions equivalent?

RQ1.1b Weak Invariance: Are the factor loadings of the TpT and RpT measurement occasions equivalent?

RQ1.1c Strong Invariance: Are the factor loadings and intercepts of the TpT and RpT measurement occasions equivalent?

RQ1.2a Configural Invariance: Are the number of factors and the factor loading pattern of the TpT and PT measurement occasions equivalent?

RQ1.2b Weak Invariance: Are the factor loadings of the TpT and PT measurement occasions equivalent?

RQ1.2c Strong Invariance: Are the factor loadings and intercepts of the TpT and PT measurement occasions equivalent?
RQ1.3a Configural Invariance: Are the number of factors and the factor loading pattern of the RpT and PT measurement occasions equivalent?

RQ1.3b Weak Invariance: Are the factor loadings of the RpT and PT measurement occasions equivalent?

RQ1.3c Strong Invariance: Are the factor loadings and intercepts of the RpT and PT measurement occasions equivalent?

The second research question is concerned with the estimation of the effect size of the latent and observed means from TpT/PT and RpT/PT research designs and the subsequent implications for interpretation about the magnitude of change. The second research question was specifically as follows:

RQ2: Do the observed mean scores and the latent mean scores yield the same effect sizes in TpT/PT and RpT/PT research designs?

RQ2.1: Is the magnitude of the effect size of the observed mean and the latent mean in a TpT/PT research design is the same?

RQ2.2: Is the magnitude of the effect size of the observed mean and the latent mean in a RpT/PT research design is the same?

Context/Sample

This study implemented a secondary analysis of existing data. A brief description of the original study is provided here to ground the reader’s understanding of the measure under study. The purpose of the original study was to assess whether a college communications course decreased communication reticence. The instrument used in this study was the Reticence Scale (RS), which was developed by Keaten, Kelly, and Finch (1997). In the development of this instrument, Keaten et al. (1997) acknowledged previous literature that suggested that individuals
who are reticent to communicate have skill deficiencies. Hence, they conceptualized the problem of communication reticence in the following way: "the reticence construct assumes that skill deficiencies defining the problem correspond to the rhetorical canons of invention, disposition, style, delivery, and memory" (Keaten et al., 1997, p. 39).

Data for that study were collected in a communications course in a mid-western private university. The RS (Keaten et al., 1997) was administered to several sections of this same communications course in the summer and fall semesters of 2011. Faculty course instructors administered the RS pre-test to students during the first week of each semester and the RS post-test in the last week of the semester, approximately 14 weeks apart. The retrospective pre-test measure was administrated directly after the post-test on a separate form. Separate forms were used for the post-test and retrospective pre-test in order to prevent the change of post-test responses and reduce bias related to effort justification (Taylor et al., 2009).

The description of the course in which the RS was administered was as follows: This course introduces and applies the theories and principles of effective communication to a variety of interpersonal, social, and business situations. Students learn to organize and present clear, logical messages to specific audiences. They develop confidence in public speaking and increase their ability to inform and persuade listeners. They also implement critical thinking and listening skills. Finally, students exhibit the skills and tools necessary to construct, organize, and deliver effective speeches. The learning outcomes were: (1) Evaluate and apply the fundamentals of communication theory, including listening, self-concept and perception, (2) Select the appropriate communication elements necessary to deliver effective oral messages and presentations, (3) Evaluate effective public speaking skills, (4) Construct, using research, a plan
to convey effective messages and presentations, (5) Analyze the audience, (6) Give class presentations that employ delivery methods (including persuasive, informative and ceremonial) appropriate to the purpose and situation, and (7) Apply computerized presentation techniques and presentation technology.

**Instrumentation**

The RS (Keaten et al., 1997) consists of 24 Likert-type items that describe six deficiencies identified as integral to communication reticence: anxiety, knowledge, timing, organization, delivery, and memory. This instrument requires respondents to indicate the extent to which they agree with the statements about their experience communicating. The response scale ranges from 1 (strongly disagree) to 6 (strongly agree). The subscales, items, and scoring can be seen in Table 5.

Table 5

**RS Subscales and Items**

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Item</th>
<th>Subscale Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>Q1. I am nervous when talking.</td>
<td>11+Q1+Q13-Q7-Q19</td>
</tr>
<tr>
<td></td>
<td>Q7. I am relaxed while talking.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q13. I feel tense when talking</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q19. I am comfortable while talking</td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>Q2. I know what to say.</td>
<td>11+Q8+Q20-Q2-Q14</td>
</tr>
<tr>
<td></td>
<td>Q8. I am unaware of what to say.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q14. I know what to discuss.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q20. I am unfamiliar with what to say</td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>Q3. I wait too long to say what I want to say.</td>
<td>11+Q3+Q15-Q9-Q21</td>
</tr>
<tr>
<td></td>
<td>Q9. I say things at the time I want to say them.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q15. I hesitate too long to say what I want to say.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q21. I say things when I want to say them.</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>Q4. I organize my thoughts when talking.</td>
<td>11+Q10+Q22-Q4-Q16</td>
</tr>
<tr>
<td></td>
<td>Q10. My thoughts are disorganized.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q16. I arrange my thoughts when talking.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q22. My thoughts are jumbled.</td>
<td></td>
</tr>
</tbody>
</table>
Table 5—Continued

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Item</th>
<th>Subscale Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>Q23. I fluently say what I want to say.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q6. I remember what I want to say when talking.</td>
<td>11+Q12+Q24-Q6-</td>
</tr>
<tr>
<td></td>
<td>Q12. I forget what I want to say when talking.</td>
<td>Q18</td>
</tr>
<tr>
<td></td>
<td>Q18. I recall what I want to say when talking.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q24. I lose sight of what I want to say when talking.</td>
<td></td>
</tr>
</tbody>
</table>

In order to derive valid inferences from a study, it is critical that the intervention and instrument being used to measure the outcomes are well aligned and that the instrument is psychometrically sound. The Reticence Scale meets these criteria. As stated previously, this instrument was created by Keaton et al. (1997) to measure communication reticence in students in college communication courses. The rationale and intention for the development of the instrument was to develop a self-report measure that could be used both for research and to provide communications teachers more detailed assessment information to help them individualize treatment. Given this, the original intention and specified use of this instrument was well aligned to the intervention for this study (college communications course) and the purpose of the research.

The psychometric properties of the instrument have been well documented. The six dimensions within the scale had the following reliability coefficients: feelings of anxiety (.91); knowledge of communication topics (.90); timing skills (.82); organization of thoughts (.82); delivery skills (.80); and memory skills (.92). The overall reliability of the total scale was .95 (Kelley et al., 1992). Other studies have produced similar reliability estimates (Keaten, Kelly, & Begnal, 1995; Kelly et al., 1994), supporting the reliability of the RS. The factor structure of this instrument was tested in the initial development of the instrument. A second-order confirmatory factor analysis produced an acceptable model fit $X^2(239, N = 267) = 289.181, p = 0.01456$, with a $CFI$ of .975. This second-order model fit allows for the use of the subscores and
total scores.

**Intervention Outcome**

The examination of the validity of TpT/PT and models requires an experience or intervention and relative within group measurement occasions over time. The measurement occasions for this study were within group TpT (administered in the first week of the intervention), and PT and RpT (administered simultaneously in the last 2 weeks of the intervention). The initial dataset included 211 cases. While it has been demonstrated that standard errors decrease as \( N \) increases, there is little practical evidence to suggest how large the \( N \) should be in order to obtain sufficiently low standard errors in CFA models. The dataset used was determined to be adequate for the research design.

**Data Preparation**

A missing data analysis was performed. See the presence of missing data for each measurement occasion in Table 6. Usable cases were where there was no more than 4 missing item responses. Overall, there were 211 usable matched cases across the three measurement occasions. In the matched sample, and analysis MCAR and MAR supported the imputation of a small number of item-level responses. Analysis revealed 29 patterns of data and that the 58 missing data elements formed 28 distinct patterns. The end result was 211 matched cases for the 24 items across all three measurement occasions.

Table 6

**Missing Data at Measurement Occasions**

<table>
<thead>
<tr>
<th>Measurement Occasion</th>
<th># of Cases</th>
<th># of Usable Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>TpT</td>
<td>411</td>
<td>364</td>
</tr>
<tr>
<td>RpT</td>
<td>259</td>
<td>226</td>
</tr>
<tr>
<td>PT</td>
<td>266</td>
<td>234</td>
</tr>
<tr>
<td>Total matched cases</td>
<td>213</td>
<td>211</td>
</tr>
</tbody>
</table>
Data Analytics

The principle analytical tool for examining the study research questions involves CFA. CFA models can be evaluated many ways. The absolute fit index used to assess the model fit was the standardized root mean square residual (SRMR). Conceptually, the SRMR is the average discrepancy between the correlations observed in the input matrix and the correlations predicted by the model (Brown, 2015). In actuality, the SRMR is a positive square root average, which is derived from a residual correlation matrix (Brown, 2015). The SRMR values range from 0.0 to 1.0, with 0.0 indicating a perfect model fit. Hence, the smaller the SRMR value, the better the model fit.

Model parsimony was evaluated using the root mean square error of approximation (RMSEA). The RMSEA incorporates a penalty function for poor model parsimony that is not found in other indices such as $X^2$ and SRMR. The RMSEA is a population-based index that relies on the non-central $X^2$ distribution (Brown, 2015). The RMSEA is sensitive to the number of model parameters, but is relatively insensitive to sample size, given that it is a population-based index (Brown, 2015). The upper range of the RMSEA is unbounded; however, it is rare to see a RMSEA value exceed 1.0. An RMSEA value of 0 indicates a perfect model fit, and values very close to 0 indicate a good model fit (Brown, 2015).

Comparative fit indices evaluate the fit of the solution specified by the user in relation to a more restrictive, baseline model (Brown, 2015). Most often, the baseline model is a null or independence model in which the covariances among all input indicators are fixed to zero. No constraints are placed in the indicator variances. Two comparative fit indices were used in this study, the comparative fit index (CFI) and Tucker-Lewis index (TLI). Both of these indices have a range of 0.0 to 1.0, with values moving closer to 1.0 implying a good model fit. However, the
TLI is non-normed, which means the values can fall outside of the 0.0 to 1.0 range.

In addition, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used in tandem with $X^2_{\text{diff}}$ for the comparison of non-nested models (Brown, 2015). These indices are closely related in that they both take into account model fit and model complexity-parsimony. Generally, the models with the lowest AIC and BIC values are deemed to fit the data better than alternative solutions. It should be noted that the AIC and BIC do not provide statistical comparison of models.

Measurement invariance was evaluated by longitudinal confirmatory factor analysis and presented in unstandardized solutions. The presentation of unstandardized solutions, while less common in the literature, is preferred in longitudinal CFA invariance analysis (Brown, 2015). This is because it allows for the analysis of standardized varience-covariance structures and mean structures, thus allowing allows for the estimation of the means of the factors and the intercepts of the indicators (Brown, 2015). It also allows for factor scores to be calculated using factor loadings and factor calculations. Conceptually, a factor score is the score that would have been observed for an individual if it were possible to measure the construct directly (Brown, 2015).

A fixed factor method (Lomax, 2013) of scaling was used in the CFA invariance analysis. For configural invariance in the fixed factor method, the variance of the constructs is fixed to 1.0. For weak invariance when the loadings are equated across time, the scaling constraint at the second measurement occasion is no longer needed (regardless of the scaling method). Constraints at subsequent measurement occasions are no longer needed because the scale for the estimates of the loadings and the variance of the constructs at those measurement occasions are determined by the scaling constraint that is in place at the first measurement occasion. Constraining the factor loadings to be equal across the measurement occasions standardizes the metric of the latent
constructs variances so that they are operationally defined in the same way. At the level of strong when the fixed factor method of scaling is used, the means of the latent constructs are fixed at zero to provide the scale for the mean level information (intercepts and latent means) at both measurement occasions. Because the latent mean at the first measurement occasion is fixed to zero, the difference between the first measurement occasion and the subsequent measurement occasion is the difference from zero.
CHAPTER IV
FINDINGS

The principle aim of this dissertation study was to test if an instrument administered in a TpT/PT design evidenced the same structural meaning as the same instrument administered in an RpT/PT design and to understand the implications on these two research designs on the estimation of the effect sizes derived from analysis of change. If measurement invariance exists in these two designs, then researchers and evaluators have flexibility in the design they employ. Utilizing an RpT/PT design would be more efficient, saving time and effort, and perhaps costing less than employing a TpT/PT design. The RpT/PT design would also have utility when a researcher or evaluator does not have access to the population under study prior to the intervention. This study was designed to evaluate measurement invariance of a communication reticence instrument. Measurement invariance was examined through the evaluation of three hierarchical levels of factorial invariance (i.e., configural, weak, and strong).

Research Questions

This dissertation addresses two research questions. The first research question focuses on invariance between the TpT and RpT measurement occasions within the TpT/PT and RpT/PT designs. The specific research questions were as follows:

RQ 1: What is the extent of measurement invariance in a communication reticence scale?

RQ1.1a Configural Invariance: Is the number of factors and the factor loading pattern of the TpT and RpT measurement occasions equivalent?

RQ1.1b Weak Invariance: Are the factor loadings of the TpT and RpT measurement occasions equivalent?

RQ1.1c Strong Invariance: Are the factor loadings and intercepts of the TpT and RpT
measurement occasions equivalent?

RQ1.2a Configural Invariance: Are the number of factors and the factor loading pattern of the TpT and PT measurement occasions equivalent?

RQ1.2b Weak Invariance: Are the factor loadings of the TpT and PT measurement occasions equivalent?

RQ1.2c Strong Invariance: Are the factor loadings and intercepts of the TpT and PT measurement occasions equivalent?

RQ1.3a Configural Invariance: Are the number of factors and the factor loading pattern of the RpT and PT measurement occasions equivalent?

RQ1.3b Weak Invariance: Are the factor loadings of the RpT and PT measurement occasions equivalent?

RQ1.3c Strong Invariance: Are the factor loadings and intercepts of the RpT and PT measurement occasions equivalent?

The second research question is concerned with the estimation of the effect size of the latent and observed means from TpT/PT and RpT/PT research designs and the subsequent implications for interpretation about the magnitude of change attributed to any intervention effect. The second research question was specifically as follows:

RQ2: Do the observed mean scores and the latent mean scores yield the same effect sizes in TpT/PT and RpT/PT research designs?

RQ2.1: Is the magnitude of the effect size of the observed mean and the latent mean in a TpT/PT research design is the same?

RQ2.2: Is the magnitude of the effect size of the observed mean and the latent mean in a RpT/PT research design is the same?
Psychometric Properties of the RS

Prior to evaluating the study research questions and hypotheses, the psychometric properties of the RS were evaluated. The RS is a six-factor instrument (Kelly et al., 1992) with six subscales (anxiety, knowledge, timing, organization, delivery, and memory). Historically, researchers who have used the RS have used both subscale and total scale scores. Given the limited sample size in this study, however, an item-level analysis of the RS items was not practical or statistically justified. Therefore, items were summed (unit weights) into composite variables reflecting the six factors of the RS, as described in Chapter III. The range of the possible RS scores was 6 to 126. The range of the possible scores for each of the six subscales was 1 to 21. Lower scores indicate less reticence, or higher levels of confidence in communication skills. The total sample size for the data analysis was 211.

RS CFA

CFA of the RS was examined to verify the hypothesized single-factor structure within each measurement occasion (TpT, PT, RpT). Robust maximum likelihood estimation was used because the underlying multivariate normality, assumed by a ML estimator, was not evident in the sample data (see Table 7). The CFA modeled a single factor structure, which is depicted in Figure 2.

Table 7

Multivariate Normality Tests ($N = 211$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Test</th>
<th>Kurtosis</th>
<th>Test Statistic Value</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Traditional Pre-Test (TpT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.987</td>
<td>0.047</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.983</td>
<td>0.010</td>
</tr>
<tr>
<td>Timing</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.981</td>
<td>0.006</td>
</tr>
<tr>
<td>Organization</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.990</td>
<td>0.167</td>
</tr>
</tbody>
</table>
Table 7—Continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Test</th>
<th>Kurtosis Test</th>
<th>Test Statistic Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Traditional Pre-Test (TpT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.985</td>
<td>0.027</td>
</tr>
<tr>
<td>Memory</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.985</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>Mardia Skewness</td>
<td>2.878</td>
<td>103.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>Mardia Kurtosis</td>
<td>54.310</td>
<td>4.692</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>Post-Test (PT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.98</td>
<td>0.005</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.978</td>
<td>0.002</td>
</tr>
<tr>
<td>Timing</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.961</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Organization</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.984</td>
<td>0.020</td>
</tr>
<tr>
<td>Delivery</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.984</td>
<td>0.014</td>
</tr>
<tr>
<td>Memory</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.979</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Mardia Skewness</td>
<td>4.752</td>
<td>170.16</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>Retrospective Pre-Test (RpT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.983</td>
<td>0.014</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.977</td>
<td>0.002</td>
</tr>
<tr>
<td>Timing</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.982</td>
<td>0.008</td>
</tr>
<tr>
<td>Organization</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.988</td>
<td>0.077</td>
</tr>
<tr>
<td>Delivery</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.986</td>
<td>0.031</td>
</tr>
<tr>
<td>Memory</td>
<td>Shapiro-Wilk</td>
<td>.</td>
<td>0.982</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Mardia Skewness</td>
<td>11.0698</td>
<td>396.424</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>Mardia Kurtosis</td>
<td>78.6644</td>
<td>22.731</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Figures 3 through 5 present the estimated models of the CFA results, and Table 8 summarizes the overall model fit. Overall, an adequate model fit was observed for the RS administered as a TpT and RpT. The model fit for the PT was not as good overall, but well within conventional limits (see Table 8).

Table 8

*Baseline CFA Model Fit (N = 211)*

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>(X^2)</th>
<th>p</th>
<th>RMSEA</th>
<th>CFI</th>
<th>TLI</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TpT</td>
<td>9</td>
<td>13.146</td>
<td>0.156</td>
<td>.047</td>
<td>.993</td>
<td>.988</td>
<td>6160.055</td>
<td>6220.389</td>
</tr>
<tr>
<td>RpT</td>
<td>9</td>
<td>12.617</td>
<td>0.181</td>
<td>.044</td>
<td>.993</td>
<td>.989</td>
<td>6128.255</td>
<td>6188.588</td>
</tr>
<tr>
<td>PT</td>
<td>9</td>
<td>17.746</td>
<td>0.038</td>
<td>.068</td>
<td>.986</td>
<td>.977</td>
<td>5949.595</td>
<td>6009.929</td>
</tr>
</tbody>
</table>
Figure 3. TpT estimated model (unstandardized solution).

Figure 4. RpT estimated model (unstandardized solution).
Given that the overall model fit for the TpT, RpT, and PT assessments generally followed a single factor model, coefficient alpha was calculated for each of the three measurement occasions. Coefficient Alpha was .91, .93, and .95 at the TpT, RpT, and PT, respectively. The Average Variance Extracted was .62, .75, and .68 for the TpT, RpT, and PT, respectively. The Composite Reliability was .91, .95, and .93 for the TpT, RpT, and PT, respectively.

**Descriptive Statistics**

The descriptive statistics for the scale and subscales for the TpT, PT, and RpT subscales can be seen in Table 9. It should be noted that higher scores on the RS indicate greater problems...
with social interaction (Kelley et al., 1992), and lower scores are associated with less reticence (higher levels of confidence) in communication skills. The mean RS scale scores were 58.0 (18.236) for the TpT, 59.9 (22.729) for the RpT, and 48.8 (18.555) for the PT. The mean subscale scores ranges were as follows: 8.6 to 11.4 for the TpT, 9.1 to 11.74 for the RpT, and 7.4 to 9.6 for the PT.

Table 9

*Descriptive Statistics for Reticence Scale and Subscales (N = 211)*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traditional Pre-Test (TpT)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>11.360</td>
<td>4.863</td>
<td>-0.092</td>
<td>-0.601</td>
</tr>
<tr>
<td>Knowledge</td>
<td>8.559</td>
<td>3.129</td>
<td>0.058</td>
<td>0.376</td>
</tr>
<tr>
<td>Timing</td>
<td>9.716</td>
<td>3.384</td>
<td>-0.048</td>
<td>-0.131</td>
</tr>
<tr>
<td>Organization</td>
<td>8.758</td>
<td>3.485</td>
<td>0.186</td>
<td>-0.340</td>
</tr>
<tr>
<td>Delivery</td>
<td>10.100</td>
<td>3.601</td>
<td>0.158</td>
<td>-0.307</td>
</tr>
<tr>
<td>Memory</td>
<td>9.460</td>
<td>3.653</td>
<td>0.331</td>
<td>0.203</td>
</tr>
<tr>
<td>RS</td>
<td>57.953</td>
<td>18.236</td>
<td>0.127</td>
<td>0.037</td>
</tr>
<tr>
<td><strong>Retrospective Pre-Test (RpT)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>11.735</td>
<td>5.163</td>
<td>-0.078</td>
<td>-0.731</td>
</tr>
<tr>
<td>Knowledge</td>
<td>9.209</td>
<td>3.828</td>
<td>0.222</td>
<td>0.314</td>
</tr>
<tr>
<td>Timing</td>
<td>10.118</td>
<td>4.071</td>
<td>0.075</td>
<td>-0.298</td>
</tr>
<tr>
<td>Organization</td>
<td>9.137</td>
<td>4.111</td>
<td>0.158</td>
<td>-0.139</td>
</tr>
<tr>
<td>Delivery</td>
<td>10.009</td>
<td>4.207</td>
<td>-0.020</td>
<td>-0.422</td>
</tr>
<tr>
<td>Memory</td>
<td>9.720</td>
<td>4.200</td>
<td>0.288</td>
<td>-0.130</td>
</tr>
<tr>
<td>RS</td>
<td>59.929</td>
<td>22.729</td>
<td>0.017</td>
<td>-0.317</td>
</tr>
<tr>
<td><strong>Post-Test (PT)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>9.550</td>
<td>4.746</td>
<td>0.319</td>
<td>-0.298</td>
</tr>
<tr>
<td>Knowledge</td>
<td>6.919</td>
<td>3.100</td>
<td>0.600</td>
<td>1.264</td>
</tr>
<tr>
<td>Timing</td>
<td>8.313</td>
<td>3.569</td>
<td>0.250</td>
<td>0.205</td>
</tr>
<tr>
<td>Organization</td>
<td>7.403</td>
<td>3.329</td>
<td>0.329</td>
<td>0.055</td>
</tr>
<tr>
<td>Delivery</td>
<td>8.521</td>
<td>3.690</td>
<td>0.220</td>
<td>-0.340</td>
</tr>
<tr>
<td>Memory</td>
<td>8.104</td>
<td>3.349</td>
<td>0.391</td>
<td>0.295</td>
</tr>
<tr>
<td>RS</td>
<td>48.810</td>
<td>18.555</td>
<td>0.325</td>
<td>0.379</td>
</tr>
</tbody>
</table>
Mean Scale Score Differences

Although the principle research questions of the study focus on the RS total score, descriptive analysis of the RS composite score scales is presented below. Typical of pre-test-post-test designs, paired t-tests were conducted to test for differences among the mean scale scores across the measurement occasions. To control for an inflated type-I error rate, a family-adjusted type-I error rate was set as .05/7 = .007. Statistically significant differences were found for the overall RS mean scale scores and each of the subscales between the TpT and PT and the RpT and PT measurement occasions, but not for the TpT and RpT occasions (see Table 10). The mean difference between the TpT and PT overall RS scale was -9.142 (15.605), indicating that the magnitude of the decrease in reticence from the beginning to the end of the semester revealed a statistically significant treatment effect. A statistically significant mean difference was also found between the RpT and PT overall RS scale ($M = -11.119, SD = 16.067$).

Measurement Invariance

Longitudinal measurement invariance was evaluated by incrementally estimating progressively more stringent measurement models, as previously described. Figure 6 depicts a longitudinal CFA model. In this figure, the circles with $M_1$ and $M_2$ represent the two distinct measurement occasions. The squares with X1 through X6 represent the observed measures at each measurement occasion. The arrows pointing from the circles to the squares are the factor loadings. The arrows pointing into the squares identify the variance (understood as measurement error in this model) of the observed measure. In configural invariance, all parameter estimates (where there are arrows) are freely estimated. Weak invariance constrains the factor to indicator loadings to be equal over time and strong invariance constrains both the factor to indicator loadings and the indicator intercepts to be equal over time.
Table 10

*t-Test Analysis of RS Scale Scores (N = 211)*

<table>
<thead>
<tr>
<th>Variable</th>
<th>M diff</th>
<th>SD</th>
<th>t</th>
<th>p</th>
<th>95% CL Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TpT/PT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>-1.810</td>
<td>4.510</td>
<td>-5.83</td>
<td>&lt;.001</td>
<td>-2.422 -1.198</td>
</tr>
<tr>
<td>Knowledge</td>
<td>-1.640</td>
<td>3.122</td>
<td>-7.63</td>
<td>&lt;.001</td>
<td>-2.064 -1.216</td>
</tr>
<tr>
<td>Timing</td>
<td>-1.403</td>
<td>3.525</td>
<td>7.000</td>
<td>&lt;.001</td>
<td>-1.881 -0.925</td>
</tr>
<tr>
<td>Organization</td>
<td>-1.356</td>
<td>3.086</td>
<td>-6.38</td>
<td>&lt;.001</td>
<td>-1.774 -0.937</td>
</tr>
<tr>
<td>Delivery</td>
<td>-1.578</td>
<td>3.544</td>
<td>-6.47</td>
<td>&lt;.001</td>
<td>-2.059 -1.097</td>
</tr>
<tr>
<td>Memory</td>
<td>-1.356</td>
<td>3.406</td>
<td>-5.78</td>
<td>&lt;.001</td>
<td>-1.818 -0.893</td>
</tr>
<tr>
<td><strong>RpT/PT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>-2.185</td>
<td>3.889</td>
<td>-8.16</td>
<td>&lt;.001</td>
<td>-2.713 -1.657</td>
</tr>
<tr>
<td>Knowledge</td>
<td>-2.289</td>
<td>3.333</td>
<td>-9.98</td>
<td>&lt;.001</td>
<td>-2.742 -1.837</td>
</tr>
<tr>
<td>Timing</td>
<td>-1.806</td>
<td>3.457</td>
<td>-7.59</td>
<td>&lt;.001</td>
<td>-2.275 -1.337</td>
</tr>
<tr>
<td>Organization</td>
<td>-1.735</td>
<td>3.347</td>
<td>-7.53</td>
<td>&lt;.001</td>
<td>-2.189 -1.280</td>
</tr>
<tr>
<td>Delivery</td>
<td>-1.488</td>
<td>3.333</td>
<td>-6.49</td>
<td>&lt;.001</td>
<td>-1.941 -1.036</td>
</tr>
<tr>
<td>Memory</td>
<td>-1.616</td>
<td>3.189</td>
<td>-7.36</td>
<td>&lt;.001</td>
<td>-2.049 -1.183</td>
</tr>
<tr>
<td>RS</td>
<td>-11.119</td>
<td>16.067</td>
<td>-10.05</td>
<td>&lt;.001</td>
<td>-13.299 -8.938</td>
</tr>
<tr>
<td><strong>TpT/RpT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.374</td>
<td>4.705</td>
<td>1.16</td>
<td>0.249</td>
<td>-0.264 1.013</td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.649</td>
<td>3.872</td>
<td>2.44</td>
<td>0.016</td>
<td>0.124 1.175</td>
</tr>
<tr>
<td>Timing</td>
<td>0.403</td>
<td>3.639</td>
<td>1.61</td>
<td>0.109</td>
<td>-0.091 0.897</td>
</tr>
<tr>
<td>Organization</td>
<td>0.379</td>
<td>3.378</td>
<td>1.63</td>
<td>0.105</td>
<td>-0.079 0.838</td>
</tr>
<tr>
<td>Delivery</td>
<td>-0.090</td>
<td>3.799</td>
<td>-0.34</td>
<td>0.731</td>
<td>-0.606 0.426</td>
</tr>
<tr>
<td>Memory</td>
<td>0.261</td>
<td>3.535</td>
<td>1.07</td>
<td>0.285</td>
<td>-0.219 0.740</td>
</tr>
<tr>
<td>RS</td>
<td>1.976</td>
<td>17.829</td>
<td>1.61</td>
<td>0.109</td>
<td>-0.443 4.396</td>
</tr>
</tbody>
</table>
Figure 6. Longitudinal CFA model.

Research Question 1

The first research question was: RQ 1: What is the extent of measurement invariance in a communication reticence scale?

RQ1.1a Configural Invariance: The number of factors and the factor loading pattern of the TpT and RpT measurement occasions is equivalent.

RQ1.1b Weak Invariance: The factor loadings of the TpT and RpT measurement occasions are equivalent.

RQ1.1c Strong Invariance: The factor loadings and intercepts of the TpT and RpT measurement occasions are equivalent.
A longitudinal confirmatory factor analysis was conducted in order to test for configural, weak, and strong invariance under the hypotheses that the single factor model of the RS was stable over time and not interacting with the intervention. If the instructional intervention did not interact with the RS, longitudinal measurement invariance should hold at the TpT and RpT measurement occasions. This hypothesis was tested in pairwise invariance models and results are presented in Table 11.

Table 11

<table>
<thead>
<tr>
<th>RpT/TpT Measurement Invariance Models for the RS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fit Statistic</strong></td>
</tr>
<tr>
<td>Chi-Square Test of Model Fit</td>
</tr>
<tr>
<td>df = 47</td>
</tr>
<tr>
<td>Scaling Correction Factor 1.1708 for MLR</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>BIC</td>
</tr>
<tr>
<td>Adj BIC</td>
</tr>
<tr>
<td>RMSEA</td>
</tr>
<tr>
<td>90 Percent C.I.</td>
</tr>
<tr>
<td>Pr(RMSEA) &lt;= .05</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>CFI</td>
</tr>
<tr>
<td>TLI</td>
</tr>
<tr>
<td>SRMR</td>
</tr>
</tbody>
</table>

Not surprisingly, there is conflicting evidence for configural, weak, and strong invariance for the RpT/TpT measurement model (see Table 11). The overall goodness of fit measures (TLI and CFI) are well above the threshold levels (.95) indicative of a good model fit; however, chi-square tests show evidence of some possibility for fit improvement, e.g., statistically significant
values, particularly in the strong invariance model. The RMSEA values in all three models are below 0.05 suggesting minimally a moderate to good data-model fit. Examination of AIG, BIC and Adj BIC, where lower values reflect preferred models, suggests an adequate model fit for each of the three models. Table 12 presents Satorra-Bentler (SB) scaled chi-square tests comparing configural to weak and weak to strong invariance. Since each consecutive level of invariance places added restrictions on the measurement model, this test evaluates the impact of these restrictions (Brown, 2015). A statistically significant SB chi-square thus reflects a statistically significant decrement in data-model fit, or that the restriction of the more conservative level of invariance resulted in a poorer model-data fit (Brown, 2015).

Table 12

<table>
<thead>
<tr>
<th>Measurement Model</th>
<th>Satorra-Bentler Scaled Chi Square</th>
<th>df</th>
<th>p</th>
<th>Delta(CFI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural-Weak</td>
<td>5.949</td>
<td>5</td>
<td>0.311</td>
<td>-.001</td>
</tr>
<tr>
<td>Weak-Strong</td>
<td>10.620</td>
<td>5</td>
<td>0.059</td>
<td>-.003</td>
</tr>
</tbody>
</table>

The computed SB $X^2$ for the configural-weak comparison was not statistically significant, indicating that weak invariance exists between TpT and RpT. The SB $X^2$ for testing strong invariance between TpT/RpT also did not reach the .05 level, indicating that strong invariance also holds between TpT and RpT measurement occasions. According to Cheung and Rensvold (2002) Delta(CFI) > .01 between consecutive models in invariance testing is indicating a substantial deterioration in model fit (p. 251). The Delta(CFI) for the configural to weak, and weak to strong invariance models did not meet this threshold.

The next set of analyses examined the TpT/PT research design. The research questions were as follows:
RQ1.2a Configural Invariance: The number of factors and the factor loading pattern of the TpT and PT measurement occasions is equivalent.

RQ1.2b Weak Invariance: The factor loadings of the TpT and PT measurement occasions are equivalent.

RQ1.2c Strong Invariance: The factor loadings and intercepts of the TpT and PT measurement occasions are equivalent.

The findings for measurement invariance for the TpT/PT are presented in Table 13.

Table 13

<table>
<thead>
<tr>
<th>TP/PT Measurement Invariance Models for the RS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Fit Statistic</strong></td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Chi-Square Test of Model Fit</td>
</tr>
<tr>
<td>Scaling Correction Factor 1.1636 for MLR</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>BIC</td>
</tr>
<tr>
<td>Adj BIC</td>
</tr>
<tr>
<td>RMSEA</td>
</tr>
<tr>
<td>CFI/TLI</td>
</tr>
<tr>
<td>SRMR</td>
</tr>
</tbody>
</table>

As was the case in the previous models, there is conflicting evidence for configural, weak, and strong invariance for the TpT/PT measurement model (see Table 13). The overall
goodness of fit measures (TLI and CFI) are well above the threshold levels (.95) indicative of a good model fit; however, chi-square tests show evidence of some possibility for fit improvement, e.g., statistically significant values, in all three invariance models. The RMSEA values in all three models are below 0.05, suggesting minimally a moderate to good data-model fit. Examination of AIG, BIC, and Adj BIC, where lower values reflect preferred models, suggests an adequate model fit for each of the three models.

The Satorra-Bentler scaled chi-square (SB) was used to estimate the difference between the configural-weak invariance models as well as the weak-strong invariance models for the TpT/PT research design. The computed $X^2$ for the paired measurement occasions was less than the critical value of 11.07 at the .05 level, indicating that the configural and weak models were not statistically significantly different from each other. The SB $X^2$ for TpT/PT was below the critical value of 11.07 at the .05 level, indicating that there was not a statistically significant difference between the weak and strong models and the Delta(CFI) was less than .01; thus, strong invariance holds over the measurement occasions.

Table 14

*Satorra-Bentler Scaled Chi-Square Difference Tests and Delta (CFI) for TpT/PT*

<table>
<thead>
<tr>
<th>Measurement Model</th>
<th>Configural-Weak</th>
<th>Weak-Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SB</td>
<td>df</td>
</tr>
<tr>
<td>TpT/PT</td>
<td>6.544</td>
<td>5</td>
</tr>
</tbody>
</table>

The next set of analyses examined the RpT/PT research design. The research questions were as follows:

RQ1.3a Configural Invariance: The number of factors and the factor loading pattern of the RpT and PT measurement occasions is equivalent.
RQ1.3b Weak Invariance: The factor loadings of the RpT and PT measurement occasions are equivalent.

RQ1.3c Strong Invariance: The factor loadings and intercepts of the RpT and PT measurement occasions are equivalent.

The findings for invariance between RpT/PT are presented in Table 15.

Table 15

*RpT/PT Measurement Invariance Models for the RS*

<table>
<thead>
<tr>
<th>Model Fit Statistic</th>
<th>Configural Invariance</th>
<th>Weak Invariance</th>
<th>Strong Invariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square Test of Model Fit</td>
<td>70.064, p = 0.0162 df = 47</td>
<td>75.425, p = 0.0185 df = 52</td>
<td>99.910, p = 0.0004 df = 57</td>
</tr>
<tr>
<td>Scaling Correction Factor 1.2904 for MLR</td>
<td></td>
<td>Scaling Correction Factor 1.2707 for MLR</td>
<td>Scaling Correction Factor 1.2427 for MLR</td>
</tr>
<tr>
<td>AIC</td>
<td>11772.511</td>
<td>11767.948</td>
<td>11786.258</td>
</tr>
<tr>
<td>BIC</td>
<td>11916.641</td>
<td>11895.319</td>
<td>11896.869</td>
</tr>
<tr>
<td>Adj BIC</td>
<td>11780.390</td>
<td>11774.912</td>
<td>11792.305</td>
</tr>
<tr>
<td>RMSEA</td>
<td>Estimate 0.048 90 Percent C.I. 0.021 0.071 Pr(RMSEA) &lt;= .05 0.526</td>
<td>Estimate 0.046 90 Percent C.I. 0.020 0.068 Probability RMSEA &lt;= .05 0.587</td>
<td>Estimate 0.060 90 Percent C.I. 0.040 0.079 Probability RMSEA &lt;= .05 0.196</td>
</tr>
<tr>
<td>CFI/TLI</td>
<td>CFI 0.986 TLI 0.980</td>
<td>CFI 0.986 TLI 0.982</td>
<td>CFI 0.974 TLI 0.970</td>
</tr>
<tr>
<td>SRMR</td>
<td>Value 0.032</td>
<td>Value 0.036</td>
<td>Value 0.046</td>
</tr>
</tbody>
</table>

Similar to the first measurement occasion, there is conflicting evidence for all three levels of invariance, especially at the level of strong (see Table 15). Overall goodness of fit measures (TLI and CFI) are above the threshold levels (.95), indicative of a good model fit; however, chi-square tests show evidence of some possibility for fit improvement, e.g., statistically significant values, especially at the level of strong. Lastly, RMSEA values in all three models are below
0.05 at the configural and weak levels, suggesting minimally a moderate to good data-model fit, but was .06, just above the threshold at the level of strong invariance. The AIC, BIC, and Adj BIC values were reasonable.

The Satorra-Bentler scaled chi-square (SB) was used to estimate the difference between the configural-weak invariance models as well as the weak-strong invariance models for the RpT/PT research design. For the weak-strong comparison in the RpT/PT design, the computed $X^2$ was above the critical value of 11.07 at the .05 level, indicating that there was a statistically significant difference between the weak and strong models, and the Delta(CFI) was just slightly above the -.01 threshold; thus, suggesting that strong invariance does not hold between RpT and PT.

Table 16

*Satorra-Bentler Scaled Chi-Square Difference Tests and Delta(CFI) for RpT/PT*

<table>
<thead>
<tr>
<th>Measurement Model</th>
<th>Configural-Weak</th>
<th>Weak-Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SB</td>
<td>df</td>
</tr>
<tr>
<td>RpT/PT</td>
<td>5.004</td>
<td>5</td>
</tr>
</tbody>
</table>

When weak invariance exists, the factor loadings are more generalizable than they would be if they were not equated across time (Little, 2013). Because the factor loadings are equal at both measurement occasions, the scale for interpreting the variance and covariance estimates is the same (Little, 2013). Hence, when the factor loadings of the two measurement occasions are found to be the same, it means that the operationalization of the construct (through the assessment items) is the same at both measurement occasions. This means that each of the assessment items contributes to the latent construct at both measurement occasions. Weak invariance was upheld for both the TpT/PT and RpT/PT research designs.

When testing for strong invariance, the fixed factor method of scaling focuses on the
observed means and the estimated intercepts of the indicators (Little, 2013). All corresponding intercepts are constrained to be equal across measurement occasions. For strong invariance, the means of the latent constructs are fixed to 0 to provide the scale for the mean-level information (intercepts and latent means) at both measurement occasions (Little, 2013). Once the intercepts of the indicators are constrained to be equal across measurement occasions, the means of the latent constructs can be freely estimated and the mean of the second measurement occasion will become the mean difference from the first measurement occasion. Thus, when strong measurement invariance exists, it allows for valid inferences to be derived about the magnitude of the mean difference between measurement occasions (effect size) because the scale for the mean-level information is the same at both measurement occasions. Strong measurement invariance was found for the TpT/PT, but not the RpT/PT research design.

Because strong measurement invariance was not found for the RpT/PT design, partial measurement invariance was evaluated. Partial invariance occurs when one or more of the loadings or intercepts in the measurement invariance models cannot be constrained to be equal across time (Little, 2013). Because the measurement invariance in this study only occurred in the RpT/PT strong invariance model, the partial invariance analysis was conducted for that design. Six partial invariance models were examined, each relaxing an individual equality constraint for the intercept of a different indicator. Findings from these analyses can be seen in Table 17.
Table 17

Weak to Strong Partial Invariance for the Unconstrained RpT/PT Model

<table>
<thead>
<tr>
<th>Unconstrained Factor Loading Intercept</th>
<th>Chi-Square Test of Model Fit</th>
<th>Scaling Correction Factor for MLR</th>
<th>df</th>
<th>p</th>
<th>AIC BIC Adj BIC</th>
<th>RMSEA</th>
<th>CFI TLI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>99.418</td>
<td>1.2448</td>
<td>56</td>
<td>0.0003</td>
<td>11787.858 11901.821 11794.088</td>
<td>0.061</td>
<td>0.974 0.969</td>
<td>0.046</td>
</tr>
<tr>
<td>Knowledge</td>
<td>83.010</td>
<td>1.2523</td>
<td>56</td>
<td>0.0110</td>
<td>11767.708 11881.671 11773.938</td>
<td>0.048</td>
<td>0.984 0.981</td>
<td>0.037</td>
</tr>
<tr>
<td>Timing</td>
<td>99.406</td>
<td>1.2470</td>
<td>56</td>
<td>0.0003</td>
<td>11788.061 11902.024 11794.291</td>
<td>0.061</td>
<td>0.974 0.969</td>
<td>0.046</td>
</tr>
<tr>
<td>Organization</td>
<td>99.351</td>
<td>1.2486</td>
<td>56</td>
<td>0.0003</td>
<td>11788.152 11902.115 11794.382</td>
<td>0.061</td>
<td>0.974 0.969</td>
<td>0.046</td>
</tr>
<tr>
<td>Delivery</td>
<td>91.342</td>
<td>1.2523</td>
<td>56</td>
<td>0.0020</td>
<td>11778.489 11892.452 11784.719</td>
<td>0.055</td>
<td>0.979 0.975</td>
<td>0.045</td>
</tr>
<tr>
<td>Memory</td>
<td>97.500</td>
<td>1.2466</td>
<td>56</td>
<td>0.0005</td>
<td>11785.652 11899.615 11791.882</td>
<td>0.059</td>
<td>0.975 0.970</td>
<td>0.045</td>
</tr>
</tbody>
</table>

All of the RpT/PT partial invariance models resulted in model fit statistics that were very similar to the unconstrained RpT/PT model. Overall goodness of fit measures (TLI and CFI) are above the threshold levels (.95), indicative of a good model fit, and the chi-square tests continued to show evidence of some possibility for fit improvement, e.g., statistically significant values. The AIG, BIC, and Adj BIC values were reasonable; however, the RMSEA values suggested some difference among the models. The models that relaxed the equality constraint for the anxiety, timing, delivery, and organization reflected RMSEA values of .06 or above, while the model that relaxed the restraint for knowledge was at 0.05. The chi-square model fit test was used to evaluate the improvement in model fit of the partial invariance model relative to the strict
invariance model. An SB $X^2$ test was conducted to evaluate the extent of the difference between the full RpT/PT model and the RpT/PT model that eliminated the constraint for knowledge. A statistically significant difference was observed, SB $X^2 = 28.495$ (df =1), $p < 0.0001$, and the Delta(CFI) was .01. This would suggest that the model that relaxes the constraint for the knowledge indicator provides a statistically significant improvement over the constrained model, and is the closest to achieving strong invariance. Relaxing the equality constraint for the other indicators variables did not improve model data fit and strong invariance was not achieved. Further, investigation of the estimates for the intercepts in the constrained and unconstrained RpT/PT models revealed differences. The estimate for the knowledge intercept in the standardized solution of the constrained RpT/PT model was 1.800. The estimate for the knowledge intercept in the standardized solution of the RpT/PT model in which the knowledge indictor was not constrained was 2.033.

The Satorra-Bentler scaled chi-square (SB) was used to estimate the difference between the full weak model and the partial strong model (without the constraint for knowledge in the RpT/PT research design. A statistically significant difference was not observed, SB $X^2 = 8.006$ (df =4), $p = 0.091$, and the Delta(CFI) was .002. The computed $X^2$ was below the critical value of 9.49 at the .05 level, indicating that there was not a statistically significant difference between the full weak and partial strong models, and the Delta(CFI) was below the -.01 threshold, suggesting that strong invariance does hold.

**Research Question 2**

The second research question asked: Do the observed mean scores and the latent mean scores yield the same effect sizes in TpT/PT and RpT/PT research designs? Specifically, this
was concerned with the estimation of the effect size of the latent and observed means in TpT/PT and RpT/PT research designs and the subsequent interpretation about the magnitude of change based on the intervention: RQ2.1 The magnitude of the effect size of the observed mean and the latent mean in a TpT/PT research design are the same.

Research question 1.2c revealed that measurement invariance to the level of strong was upheld for the TpT/PT model. Thus, using a dependent $t$-test to evaluate the gain score for the course intervention effects is a plausible option. The intervention effects estimated from the raw scores reveal a mean change in RS scores of \(-9.142\), a statistically significant decrease in communication reticence. Using latent scores one gets to the same hypothesis test conclusion, but the standardized effect size was larger. A direct comparison of the intervention effect as estimated by raw and latent scores suggests that raw scores underestimated the intervention effect.

Dependent testing usually yields a higher power than independent testing due to the interconnection between the two measurement occasions. In order to account for this the interconnection between the two measurement occasions, the correlation between the measurement occasions has to be accounted for (Bornenstein et al., 2009; Dunlap et al., 1996).

According to Borenstein et al. (2009), for dependent-samples (e.g., pre-post), the formula for Cohen’s $d$ is:

$$d = \frac{\bar{Y}_{diff}}{S_{within}} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{within}}$$

where

$$S_{within} = \frac{S_{diff}}{\sqrt{2(1 - r)}}$$

and $r$ is the correlation between pairs of observations, with variance
\[ V_d = \left( \frac{1}{n} + \frac{d^2}{2n} \right) 2(1 - r) \]

and

\[ SE_d = \sqrt{V_d} \]

Intervention effects for raw and latent trait scores are presented in Table 18. Also in Table 18 is the change from pre-test to post-test standardized, e.g., Cohen’s \( d \) statistic for both raw and latent scores.

Table 18

<table>
<thead>
<tr>
<th>Variable</th>
<th>( M ) diff</th>
<th>( SD ) diff</th>
<th>( t )</th>
<th>( p )</th>
<th>95% CL Mean</th>
<th>( r )</th>
<th>( p )</th>
<th>Cohen’s ( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent score</td>
<td>-.506</td>
<td>.741</td>
<td>-9.910</td>
<td>&lt;.001</td>
<td>-.606</td>
<td>-.405</td>
<td>.697</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

As shown in Table 18, the Cohen \( d \) for the TpT/PT was .497 based on raw scores and was .531 for latent trait scores, or .034 \( d \)-units lower. Direct statistical comparison of the Cohen’s \( d \) statistics via \( z \)-test reveals, however, there was no statistically significant different in the raw or latent means (\( p = .2839 \)). This lack of difference can also be observed by the substantial overlap in 95% confident intervals around \( d \) (see Table 19).

Table 19

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cohen’s ( d )</th>
<th>( Z )-Value</th>
<th>( SE )</th>
<th>( p )</th>
<th>95% CL Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw score</td>
<td>.497</td>
<td>8.025</td>
<td>.062</td>
<td>&lt;.0001</td>
<td>.376</td>
</tr>
<tr>
<td>Latent score</td>
<td>.531</td>
<td>9.277</td>
<td>.057</td>
<td>&lt;.0001</td>
<td>.419</td>
</tr>
</tbody>
</table>
Research question 2.2 states that the magnitude of the effect size of the observed mean and the latent mean in an RpT/PT research design are the same. Because of the lack of measurement invariance in the RpT/PT at the level of strong (RQ1.3c), using a dependent t-test to calculate mean differences in the RpT/PT would not be deemed appropriate; however, dependent t-tests were calculated similar to research question 2.1 for illustration purposes. Intervention effects for raw and latent trait scores are presented in Table 20. Also in Table 20 is the change from pre-test to post-test standardized, e.g., Cohen’s d statistic for both raw and latent scores. The intervention effects estimated from the raw scores reveal a mean change in RS scores of -11.119, a statistically significant decrease in communication reticence. Similar to the TpT/PT research model, using latent scores one gets to the same hypothesis test conclusion, but the standardized effect size was larger.

Table 20

<table>
<thead>
<tr>
<th>Variable</th>
<th>M diff</th>
<th>SD diff</th>
<th>t</th>
<th>p</th>
<th>95% CL Mean</th>
<th>r</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw score</td>
<td>-11.119</td>
<td>16.067</td>
<td>-10.05</td>
<td>&lt;.001</td>
<td>-13.299</td>
<td>.715</td>
<td>&lt;.0001</td>
<td>.523</td>
</tr>
<tr>
<td>Latent score</td>
<td>-.604</td>
<td>.788</td>
<td>-11.125</td>
<td>&lt;.001</td>
<td>-.711</td>
<td>-.497</td>
<td>.759</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The Cohen’s d for the observed mean in the RpT/PT was .523, whereas the mean difference for the latent trait was .532, a difference of .009. Clearly the two effect sizes (raw score and latent score) are very close, z-test of the mean difference confirms there was no statistically significant difference between them (p = .4326).

Table 21

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cohen’s d</th>
<th>Z-Value</th>
<th>SE</th>
<th>p</th>
<th>95% CL Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw score</td>
<td>.523</td>
<td>9.435</td>
<td>.055</td>
<td>&lt;.001</td>
<td>.414</td>
</tr>
<tr>
<td>Latent score</td>
<td>.532</td>
<td>10.410</td>
<td>.051</td>
<td>&lt;.001</td>
<td>.431</td>
</tr>
</tbody>
</table>
These findings demonstrate that in the context of this study, both the TpT/PT and the RpT/PT research designs, the effect of the intervention is underestimated when compared to the latent mean difference, but the underestimation is not statistically significantly different in either research design. Because the TpT/PT measurement occasions were found to be invariant to the level of strong and the raw and latent mean scores are not statistically significantly different, effect size estimates and subsequent inferences about the magnitude of the change in this research design may be seen as valid. In the RpT/PT research design, although there is not a statistically significant difference between the raw and latent scores, because invariance was only established to the level of weak, questions remain about the validity of the magnitude of the scores. Hence, it is impossible to derive any inference about the magnitude of the change in this model.
CHAPTER V
CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

Conclusions

The principle aim of this dissertation study was to test if an instrument administered in a TpT/PT design evidenced the same structural meaning as the same instrument administered in a RpT/PT design, and the extent to which observed scores obtained in the context of these designs yield effect size estimates that are consistent with the latent means. The first research question and set of hypotheses was concerned with the measurement invariance of the TpT and RpT measurements. The TpT and RpT measurement occasions were found to be invariant to the level of strong indicating that the underlying structure of these two measurement occasions is the same. Specifically, this means that the number of factors and the factor loading pattern and the intercepts are equivalent at both measurement occasions.

The second set of hypotheses were concerned with the measurement invariance in the TpT/PT research design. The TpT and PT measurement occasions were found to be invariant to the level of strong. As with the previous model, this indicates that the underlying structure of these two measurement occasions, including the number of factors, the factor loading pattern, and intercepts are equivalent at both measurement occasions. This indicates that in the context of this study, alpha change (Brown, 2015), which is the true score change in the construct, can be estimated because it has been demonstrated that the conceptual domain is constant and structural (measurement) properties of the measurement occasions is the same.

The third set of hypotheses related to the first research question evaluated measurement invariance across the RpT and PT measurement occasions. The RpT and PT measurement occasions were found to be invariant to the level of weak, providing support for research
question 1.2a and research question 1.2b, but not research question 1.2c. Because research question 1.2c was not upheld, this means that, in the context of this study, the RpT/PT research design cannot be used to evaluate alpha change. Instead, in this design, beta change (Brown, 2015) occurs. Beta change occurs when the construct remains constant, but the measurement properties of the indicators of the construct are inconsistent across the measurement occasions. This means that in this study, the RpT and PT measurement occasions were not on the same measurement scale; hence, the mean difference between the RpT and PT measurement occasions cannot be interpreted as the change in the latent construct. This represents a change in the mean expression and a change is the structure of the construct, which are now confounded. Because of this, partial invariance was explored to unpack the structural components of the model that changed over time. This analysis indicated that when the constraint for equality of intercepts for the knowledge indicator was relaxed, there was a statistically significant improvement in the model fit and invariance held at the level of strong. This suggests that RpT/PT models that include the measurement of knowledge may differ than measurement of other attributes. Specifically, it suggests that the scalar properties for the knowledge rating may be different at the RpT and PT measurement occasions. The difference in the standardized estimates for the knowledge indicator in the constrained and unconstrained RpT/PT models also supports this conclusion.

The second research question and hypotheses was concerned with the magnitude of the effect size estimate and the extent to which the observed means or the mean of latent trait yield similar effect size estimates. The first hypothesis related to this research question was that the magnitude of the effect size estimate of the observed mean and the latent mean in a TpT/PT research design are the same. A comparison of the Cohen’s $d$ for the paired $t$-test and the latent
mean difference revealed that in the TpT/PT research design, the effect size estimate was not the same, and that the observed score mean difference underestimated the effect size estimate by .034. An evaluation of the second hypothesis revealed that the magnitude of the effect size estimate of the observed mean and the latent mean in a RpT/PT research design are the same, yielded a similar finding. The observed score mean difference underestimated the effect by a lesser amount (.009).

Discussion

These findings have implications for the validity of inferences that may be derived from research designs that utilize TpT/PT or RpT/PT measurement occasions, and considerations for planning such studies. In research, drawing valid inferences from findings is dependent on proposed use and interpretation, as well as the assumptions of the research design, measurement, and analysis techniques used to derive the test scores. Threats to validity related to inferences about constructs are concerned with the match between the study operations and the constructs used to describe the operations. There may be issues with the research design, or the measurement strategies, or both. The findings from this study reveal that it is possible for the factor structure in research designs that utilize TpT/PT or RpT/PT measurements to be equivalent; however, the measurement of the factor is only equivalent (to the level of strong invariance) across measurement occasions in the TpT/PT design. This indicates that, at least in this study, there is not a shift in the understanding of the construct at the TpT, RpT, or PT measurement occasions; however, the structure of measurement is equivalent in the TpT/PT but not the RpT/PT. The partial invariance analysis indicated that it was the measurement of the knowledge indicator that was led to the invariance at the level of strong in the RpT/PT design.

These findings add to the existing literature about measuring self-reported change by
examining threats to validity associated with measurement invariance in TpT/PT and RpT/PT research designs, and the implications for the theory of response shift. Response shift theory or response shift effect (Bray & Howard, 1984) suggests that participants experience a change in the understanding of the meaning of the construct being measured, suggesting that threat to statistical conclusion validity of heterogeneity of units, and treatment-sensitive factorial structure and instrumentation are concerns in TpT/PT research designs. The assumption related to heterogeneity of units is that at the TpT measurement occasion subjects are more heterogeneous than at RpT measurement occasion, which occurs after they have experienced a common intervention. This assumption is based on the notion that prior to an intervention, the variability of the vocabulary, knowledge, understanding, behaviors, or skills associated with the intervention is likely to be greater between and among participants than after the intervention. It is theorized that this is because the common experience of participation in the intervention itself leads to less variability within the participants. Hence, the TpT/PT results in reduced covariation between measurement occasions. Similarly, the threat to treatment-sensitive factorial structure suggests that the intervention itself may lead to a change in the factor structure at the PT measurement occasion, leading to measurement invariance at the level of weak. Finally, the instrumentation threat to validity is concerned with the possibility that the nature of the measure may change over time, which could be confused with a treatment effect. This study suggests that these threats to validity may not be primary concerns in TpT/PT or RpT/PT research designs given lack of invariance to the level of weak in both designs.

The finding in this study that the RpT/PT was not invariant at the level of strong may indicate that RpT/PT research designs are more susceptible to the threat of violated assumptions of statistical tests, e.g., assumption of measurement invariance, which suggests that if statistical
test assumptions are violated, the effect size of the intervention may be over or underestimated. Although it was confirmed that the factor structure and operationalization of the factors was upheld (weak measurement invariance) for the RpT/PT research design, the lack of invariance at the level of strong means that the estimation of the magnitude of the mean difference between the two measurement occasions is not appropriate. This indicates that the change measured in the RpT/PT design is a combination of a change in the structural meaning of how the instrument measures communication reticence (RS) plus a possible change in the mean (amount) of RS in the sample. Given this, tests of mean difference to estimate change due to an intervention in the RpT/PT research design may not allow for appropriate inferences about the magnitude of the change. However, this study adds to the evolution of the theory of response shift by specifically demonstrating that the self-reported attribute of knowledge may function differently than other self-reported changes related to attributes such as behavior or memory in the RpT/PT design. Because this difference occurred at strong invariance, this may indicate that subjects use the rating scale slightly differently in the RpT and PT measurement occasions.

The threats to validity that the second research question and associated hypotheses are primarily concerned with is statistical conclusion validity, incorrectly concluding the strength of the covariation between two measurement occasions leading to the over or underestimate of the effect size of the intervention; specifically the inaccurate effect size estimation. This study demonstrates that even in TpT/PT research designs, there may be a difference between the estimate of the effect size of the observed mean difference and the latent trait mean difference. This has implications for the inferences that may be able to be drawn from findings of TpT/PT and RpT/PT research designs. Minimally, the violated assumptions of statistical tests and
inaccurate effect size estimation threats to validity, and the impact of these threats on inaccurate effect size estimation, should be considered in the research design and the desired inferences.

The overarching concern of the theory of response shift and of this study is the validity of the inferences about change, and the relationship between research design and the validity of inferences. In this study, the fundamental assumptions of response shift that there is a change in construct from pre-test to post-test which results in the over or underestimation of the intervention effect, was not observed. There was not a statistically significant over or underestimation of the construct at pre-test (RpT or TpT). However, it is important to note that the initial analysis of mean score also did not suggest that response shift had occurred because a reasonable change was observed. This study would suggest that in this context, the TpT/PT and RpT/PT research designs resulted in the explication of the same construct at pre-test and post-test. In this context, both RpT and TpT measurements could be used to derive valid inferences about the constructs being measured. However, given that measurement invariance in the RpT/PT was only found to the level of weak, only the TpT/PT research design could be used to derive valid inferences about the magnitude of the scores. The partial invariance model revealed that this may only be true when including the measurement of the knowledge indicator.

**Limitations**

Understanding how research design impacts the validity of inferences that may be drawn from findings in these designs is critical because the experimenter must use logic, design, and measurement to assess the potential to do so. The findings from this study suggest that at least in this context, the theory of response shift as defined as a change in the understanding of the construct is not sufficient to explain the difference in the magnitude of the scores yielded from RpT and TpT measurement occasions. Because weak invariance was found in the TpT/PT and
RpT research designs, it is evident that the understanding of the construct remained the same at each measurement occasion. However, because strong invariance was found in the TpT/PT but not the RpT/PT for all indicators, there remains a question about why this might be the case. Further research that explores why measurement of different types of indicators, for example, knowledge versus behaviors, would be useful.

**Recommendations**

To understand the implications of the difference in the magnitude of the RpT and TpT measurement occasions, future research incorporating additional variables that may be deemed useful based on prior research in partial invariance analysis such as assessment of prior knowledge or exposure to the content (Manthei, 1997) could be useful. Future research should be replicated with an instrument that measures attitudes, knowledge, and behaviors to determine if one is more susceptible to response shift than the others. Ideally, further research would use similar research designs including utilizing a tool with sound, previously documented psychometric properties, similar timing between measurement occasions, and the RpT measurement occasion delivered directly after the PT measurement on a separate form.

More generally, future research might focus on understanding how various research designs affect interpretation of effect size estimates. Knowing that a shift in the understanding of the meaning of the construct may not be the reason for the variation in the magnitude of the TpT and RpT measurement occasions, an emphasis could be placed specifically on understanding the reasons for the differences. Knowing this would help researchers understand what inferences are appropriate to derive from various analyses, and perhaps begin to understand the extent to which there is commonality among the difference in the magnitude of differences.

In addition to recommendations for further research, this study also highlights the need
for practitioners engaged in the estimation of effect size based on TpT, PT, or RpT measurement models to incorporate measurement invariance evaluation into their regular practice. It is important to certify measure invariance in all studies, not only those where a little or no effect is noted. The finding that measurement invariance exists is the only way to give credence to the inference of change and then to the magnitude of the change. Hence, establishing measurement invariance when assessing change is the only way to derive valid inferences about the impact of an intervention.

Finally, there is still relatively little research that identifies and demonstrates processes or procedures for evaluating model fit in measurement invariance models, particularly when the model fit indices may provide nuanced information that may discriminate between what may be determined to be a “good” versus a “moderate” model fit. This is also true for partial measurement invariance models. Continued research into understanding how to proceed with partial invariance given various scenarios encountered in the examination of model fit in CFA would be helpful, particularly if trying to meet the expectation for incorporating measurement invariance into common practice when assessing change.
REFERENCES


