EVALUATING FACTOR STRUCTURE AND INSTRUMENT STABILITY THROUGH MEASUREMENT INVARIANCE OF THE DECA, SECOND EDITION

By

KATHRYN LYNN BLACK

Bachelor of Science in Human Development and Family Science
Oklahoma State University
Stillwater, OK
2010

Master of Science in Human Development and Family Science
Oklahoma State University
Stillwater, OK
2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2021

EVALUATING FACTOR STRUCTURE AND INSTRUMENT STABILITY THROUGH MEASUREMENT INVARIANCE OF THE DECA, SECOND EDITION

Dissertation Approved:

Dr. Jam Khojasteh

Dissertation Adviser

Dr. Ki Cole

Dr. Mwavita Mwarumba

Dr. Michael Yough

ACKNOWLEDGEMENTS

I would like to acknowledge my colleagues, friends, and family who have encouraged and supported me throughout the last four years.

To my colleagues and friends, thank you for the continued encouragement, support, and reminders of my personal and professional goals beyond this degree. Your confidence in me and my abilities propelled me to keep going and share my new found knowledge and skill sets, both within my professional career and personal life. I would also like to send my gratitude to CAP Tulsa, to which I have been employed over the last nine years, and who allowed me to use their data to create this dissertation.

To my family, thank you for the ongoing love and support. First, to my husband, Seth, who selflessly gave so much over the last four years so I could both work and attend school. His love and encouragement has meant so much, and I would not have accomplished this degree without him. And to my family, I send lots of gratitude for the continual support, be it meals shared or phone calls to help me navigate the demands of personal life, professional life, and school.

Finally, I would be remiss if I did not thank my dissertation advisor, Dr. Jam Khojasteh, the REMS faculty, and my dissertation committee. I sincerely appreciate Dr. Khojasteh's support throughout this process. He gave me the autonomy to design a study that aligned with my passions and professional goals, and always responded to my many questions in such a punctual fashion. I appreciate his commitment to helping me navigate this degree. To close, I appreciate the support and guidance offered by the REMS faculty at Oklahoma State and my dissertation committee. I thoroughly enjoyed every REMS class throughout these four years, and sincerely appreciate the many opportunities I had to further my knowledge and skill sets.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: KATHRYN LYNN BLACK

Date of Degree: JULY, 2021

Title of Study: EVALUATING FACTOR STRUCTURE AND INSTRUMENT STABILITY THROUGH MEASUREMENT INVARIANCE OF THE DECA, SECOND EDITION

Major Field: EDUCATIONAL PSYCHOLOGY

Abstract: In the process of instrument development, developers follow protocols and conduct analyses to ensure psychometric properties are met. During development, developers determine an assessment format that best aligns to their desired construct, including direct or indirect assessment. Within indirect assessment, participants hold a non-active role in the assessment process, where a separate informant supports the assessment process on behalf of the participant. Within this type of assessment, informant perception, being implicit and explicit biases, as well as informant memory implicate the results of the instrument; and therefore, the psychometric properties. This study sought to understand the factor structure and instrument stability via measurement invariance of the Devereux Early Childhood Assessment (DECA), second edition, within a sample of children enrolled in Head Start. Measurement invariance seeks to measure stability within populations to ensure the instrument is equivalent between groups. Further, achieving invariance within instruments means interpretations in observed change reflect actual change within the latent variable (Millsap, 2010). Exploratory and confirmatory factor analysis identified a four-factor structure. Most of the items within the total protective factors domain aligned to the factor structure described by the DECA. Six of those items cross-loaded; 9 of the 11 items within behavior concerns domain loaded to a factor within the total protective factors domain. Within measurement invariance, gender, race/ethnicity, dual language status, and time variables were examined. No invariance model met configural invariance. To further examine invariance by gender, items within the behavior concerns domain were excluded. Configural and metric invariance were met and partial invariance within strong and strict invariance were met. This notes the need to further examine the items included in the behavior concerns domain. Based on the results of this study, the argument that conducting reliability and validity analyses during instrument developer is insufficient. Developers should conduct more advanced analyses to ensure robustness and appropriateness within the population(s) for which the instrument has been devised. Finally, due to the nature of indirect assessment, these analyses are vitally important due to the potential error that may be imparted into an informant’s judgement via perception and memory.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

In the process of instrument development, developers follow protocols and conduct early analyses to ensure the instrument has adequate psychometric properties, including reliability and validity, to confirm the instrument is appropriate for its intended use (Raykov & Marcoulides, 2011). Upon devising a definition for the construct of interest, developers create an assessment format that most appropriately aligns with the construct under review. This may include direct assessment of the construct, where the participant has active involvement in the data collection process by responding to or completing the instrument, or indirect assessment of the construct, where the participant holds a non-active role within the data collection process, and another individual, or proxy, supports the assessment process. Examples of indirect assessment options include participation in an observation that is conducted prior to completion of the instrument or some other form of indirect assessment like participation in qualitative methods that offer the opportunity to share information and aggregate themes across participants (DePaul

Teaching Commons, 2021). The methods of evaluating soundness and appropriateness for specific populations may be more challenging to conduct within each of these assessment options. In turn, these challenges may warrant additional analyses for ensuring appropriateness for intended use.

Within the realm of indirect assessment, observation is a frequently used format to gain assessment of a construct and may be conducted by a separate informant rather than the participant of interest (DePaul Teaching Commons, 2021). For instance, within the context of education, observers, or informants, regularly enter classrooms to observe the quality of instruction. Upon completion of the observation to which an informant is observing the classroom teacher and her instructional style and quality, that informant then completes the assessment of quality of the instruction on behalf of the teacher. The teacher may or may not have any input, and if input is rendered, it may be coded separately from the informant's assessment. This method of indirect assessment allows an external informant to conduct an observation and rate based on metrics of quality, such as climate of the classroom, interactions between teachers and students, emotional support offered to students, sensitivity of the teacher, modeling of language, feedback quality, behavior management, and concept development (Pianta et al., 2008). Importantly, the outcome variables within this example of an indirect assessment are the teacher's quality of instruction, albeit, the means for collecting the data for analysis is via an informant, not the teacher.

Often, prior to observing, some form of reliability of the informant is gained to ensure the informant possesses the skill set to consistently utilize the instrument and evaluate the construct similarly in comparison to other informants who may be conducting similar observations on the same participant (Raykov & Marcoulides, 2011). There are various forms of reliability, such as internal consistency, test-retest, interrater, and parallel forms (Raykov & Marcoulides, 2011). During the instrument development process, developers seek to ensure the instrument and its items are consistent across each form of reliability. In this instance, interrater reliability is one aspect of reliability that is assessed during the instrument development process. Importantly, sharing and training for informant reliability may vary across instruments, where some instruments offer lengthy, detailed training prior to an informant completing a test or certification of reliability (Pianta et al., 2008), while others offer no specific training, but rather the informant reviews a user's manual prior to administering the instrument (LeBuffe & Naglieri, 2012).

Understanding these differences in gaining informant reliability, also known as interrater reliability, and ensuring holistic psychometric soundness across all forms of reliability and validity are important for indirect assessment because these instruments can pose greater risk for error and skewness if the informant is not appropriately prepared to conduct the observation and rate the same subject similarly in comparison to other informants (Raykov & Marcoulides, 2011). Specifically, Dracobly et al. (2018) assessed reliability and validity of informants within an indirect assessment specific to functional behavior, and found differences between expert ratings and caregiver, or non-expert, ratings. The researchers devised possible reasons related to these differences, namely suggesting the ability to separate environmental influences and the ability to study a variety of situations and not isolated events when assessing behavior. For these reasons, indirect assessments may risk accuracy of scores based on observed behavior.

More specifically, informant biases, both within instances where a formalized process of informant reliability is gained and within instances where informant reliability is informally gained, may implicate the results of the instrument, therefore potentially nullifying the usefulness of the results. These biases pose the risk that an instrument is being inappropriately used for various purposes, such as evaluation or instructional purposes, if the instrument is unable to produce reliable and valid results specifically within the population for which it is being used. For example, a practitioner may implement an instrument and seek to strategize change according to the results or a teacher may implement an instrument and seek to differentiate instruction, when in fact, underlying informant biases may be implicating the results and not necessarily true group differences. Without advanced analytic techniques in the early development and testing phases to ensure the instrument performs similarly across groups and occasions, developers risk not identifying limitations prior to broader use, which is included as a standard for ensuring instrument appropriateness by various subgroups (American Psychological Association, American Educational Research, & National Council on Measurement in Education, 1999).

Within the context of informant biases, biases related to informant perception and informant memory are two examples of potential influences on psychometric soundness. Further, within informant perception, both conscious and unconscious biases offer opportunity for potential error and skewness in the informants' scoring. Conscious biases, also known as explicit bias, is defined as conscious awareness and willingness to identify and share one's preferences, values, and attitudes (Duameyer et al., 2019), and can directly implicate the reliability and validity of the instrument. Unconscious bias, or implicit bias, which is defined as an attitude that exists and is placed on novel objects (Greenwald & Banaji, 1995), can also influence the reliability and validity, but may be imparted into the instrument more subconsciously. Implicit bias includes one's attitudes, prejudices, and stereotypes that emerge without awareness. Finally, within informant bias, memory of the observation may skew perception of the construct being studied.

Based on these examples of informant biases, underlying beliefs or attitudes may directly or indirectly influence how an informant rates a construct, which pose great risk for true psychometric soundness and broad use of the instrument. These limitations may be challenging to identify without use of more advanced analytic techniques, such as differential item functioning and measurement invariance, that study instruments at the item-level and by various subgroup populations to ensure psychometric properties remain adequate.

This study seeks to demonstrate an example of data collected with young children within the realm of indirect assessment and seeks to highlight the challenges instrument developers must tackle. This example utilizes the second edition of a social-emotional instrument, specifically the Devereux Early Childhood Assessment (DECA) (LeBuffe & Naglieri, 2012), that is used within the early childhood field, specifically including preschool classrooms with children ranging from 3-years-old to 5-year-olds. Within this instrument, the informant observes children over an extended period prior to completing the instrument. This informant, specifically a teacher, observes children within the classroom for four weeks. In this example, the outcome variable of the study is the social-emotional skill sets possessed by child participants; however, the means to collect that data for analytical purposes is via a teacher acting as the informant. Indirect means of assessment do not consider the biases and memory possessed by the teacher but seeks to understand specific skill sets children possess without their direct involvement in the assessment process.

Within this example, informant biases are highlighted both within informant perception and informant memory. More specifically, within the conscious level of informant perception, educator philosophical and pedagogical beliefs may directly influence beliefs which may impact informant perception and how the informant rates children along this instrument. A plethora of research has focused on defining the construct of and measuring the components of teacher beliefs. Fives and Buehl (2012) sought to do just that in their literature review of defining and synthesizing the construct of teacher beliefs, as well as the implications those beliefs can have within various facets of the classroom, including the facets of philosophy and pedagogy. From their perspective, defining beliefs from the perspective of the teacher and attaching the many characteristics that can be connected to a definition, are important to framing the use of beliefs within practice (Fives & Buehl, 2012).

Philosophical beliefs, as in beliefs related to one's approach within the classroom, can implicate beliefs in child behavior and child development knowledge. La Paro et al. (2009) defines beliefs as mental representations, to which are subjective, are influenced by views, reasoning, and communication. Further, beliefs guide action, which influences behavior. Within the classroom, philosophical beliefs implicate how teachers structure their classroom, including their stance on offering and engaging children through play-based learning versus more structured learning settings or other types of settings, which ultimately implicates the children's experiences within the classroom. Within the context of pedagogical beliefs, or a teacher's approach to teaching, beliefs related to use of curriculum and developmentally appropriate practices may also implicate their rating of children's skill sets.

Digging deeper, if a teacher believes young children should be able to sit and attend in a whole group situation for an extended period, and if a child is unable to meet that expectation, the teacher may consequently develop certain beliefs related to the child's ability and skill set. Upon time to complete an instrument of the child's skill sets, the informant may recollect those beliefs, which may cause her to over- or under-attend to the true skills the child possesses. These beliefs align to the effects of teacher expectancy theory, as studied by Kuklinski and Weinstein (2001). These researchers found differences when assessing the role teacher expectations play within the achievement of students, and whether expectations and behaviors varied for different students. Their study suggests teacher expectations within the classroom do influence student achievement.

Lastly, regarding these conscious biases, both philosophical and pedagogical beliefs may be implicated by various characteristics of the informant/teacher and child, such as gender, race and ethnicity, and dual language learning status. For instance, as teachers devise specific assumptions related to child development and developmentally appropriate practices (Suk Lee et al., 2006), they may be more open and attentive to the needs and skill sets of certain children. For example, children who have similar temperaments and personalities as the teacher may build a more positive relationship and have more positive interactions due to the aligned personalities. Further, children whose temperament and personality lack alignment with the teacher may be more challenged in creating strong attachments, which may impact the support and guidance offered by the teacher; therefore, potentially implicating the teacher's true knowledge of the skill sets certain children possess. A second example includes relationship development and interaction opportunities between teachers and children who have differing dominant languages. In this instance, lack of communication and interpretation of needs may cause more challenges to devising positive relationships and having positive interactions. Again, potentially implicating the teacher's true knowledge of the skill sets certain children possess.

Unconsciously, informants' subconscious beliefs or stereotypes of certain groups or characteristics of individuals influence perception and biases. These subconscious beliefs can promote or demote diversity, equity, and inclusion within the classroom and broader school system. These unconscious biases may hinder certain children from receiving the instruction and support needed to implicate their schooling experience and longer-term trajectories.

Within informant memory, the social-emotional instrument used in this study requires teachers, i.e., informants, to complete four weeks of observation prior to conducting ratings. Memory related to the children's experiences as a stable metric where the informant recollects the broader observations of skills the child possesses as seen throughout the observation period versus recollection of certain instances of extreme behavior of the child can implicate the informant's true memory of the observation and children's performance on the skills, or construct, under review. Informants may have stronger recollections of the bursts of extreme behavior, interpret those behaviors as more typical behavior possessed by the child, and use that memory when rating skill sets. If this occurs, analysis of the skills being studied may be skewed and possess error, which contributes to the lack of meaningful, comparable data.

Informant perception, both consciously and unconsciously, as well as informant memory are important components to consider within the indirect assessment process. With the potential error and skewness possessed by the indirect process, this study seeks to utilize an example to examine subgroup characteristics, including gender, race and ethnicity, dual language learning status, and time, to provide further validation of this social-emotional instrument. Disaggregating the sample into groups and utilizing advanced techniques will further promote the appropriateness of use of this tool specifically for children in the early childhood setting. Within this example, measurement invariance within the structural equation modeling framework will be utilized to conduct these advanced analyses, and longitudinal measurement invariance will be utilized to conduct the time analysis. Measurement invariance studies a measure's stability within populations, seeking to ensure equivalency between groups so that interpretations in observed change reflect actual change within the latent variable (Millsap, 2010); therefore, further validating this tool will promote continued advocacy for conducting these types of analyses within indirect assessments, and in turn, will help practitioners gain a better understanding of the appropriateness of use of various instruments within their specific context.

CHAPTER II

REVIEW OF LITERATURE

Measurement

One of the many goals of social sciences is to measure a latent, or unobserved, construct that accounts for well-defined behaviors. These behaviors are defined by devising an operational definition of the behavior, also known as a construct, seeking to be studied. Instrument developers then engage in a detailed process to operationalize their definition of the construct, develop instrument specifications that best align to studying the construct under review, devise items that appropriately measure the construct, and evaluate appropriateness of items within various subject trials (Raykov & Marcoulides, 2011). Within this journey to obtaining a well-defined instrument, evaluation of various statistical properties, like reliability and validity, are necessary.

Direct versus Indirect Assessments

Encompassed within studying psychometric properties during instrument development are necessary considerations when assessing a construct directly versus indirectly. More specifically, a direct assessment seeks to gain information related to the construct by engaging with the participant of interest (DePaul Teaching Commons, 2021). Further, the participant plays an active role in the data collection process (DePaul Teaching Commons, 2021). Examples of direct assessment include participant completion of depressive or anxiety symptoms, assessment of motivation, or completion of other standardized instruments. Digging deeper within the early childhood field, when seeking to understand young children's literacy, language, or mathematics skills, an individual, or evaluator, typically sits with a child, read prompts provided by the instrument, and mark whether the child answered correctly or incorrectly. These domains of development are more often tied to discrete skill sets, such as vocabulary knowledge, letter-word identification, number identification, and counting in early childhood, in which this means of assessing skill sets allows for potentially more accurate data collection since the evaluator seeks to remove their own bias or influence, and mark children's responses to pre-specified prompts, and oftentimes, pre-specified correct answer options. Further, discrete skill sets that are studied via direct assessment seek to quantify knowledge or readiness via mastery, such as the total number of lowercase letters identified or identification of correct vocabulary terms when pointing to a picture.

For other developmental domains, like social-emotional skills that seek to assess ability to manage own feelings, ability to establish and maintain relationships with peers and other adults, cooperation in various social settings, or learning approaches that seeks to assess persistence and flexibility (Teaching Strategies, 2015), challenges arise in assessing young children since these skills are not discrete constructs, as a meta-analysis conducted by La Paro and Pianta (2000) described. When seeking to understand instrument's ability to predict outcomes across an array of studies, these researchers found moderate effect sizes within academic and cognitive domains and small effect sizes within social and behavioral domains. Further, as they reflected on prior research by Meisels (1996; 1999) and others (Tramontana et al., 1988), studies have found social and behavior domains to be poorer indicators for future performance, further raising the concern of construct operationalization and appropriate assessments of readiness.

To articulate an example, a young child, especially those who do not yet have verbal skills or are still developing verbal skills, would be unlikely to respond accurately to a question like, "How often are you happy to see your parent?" Although an evaluator could ask that question to a child, a child would be unlikely to reflect on various prior occasions and decipher their amount of "happiness" during those recent occasions; therefore, researchers should be more cautious when analyzing and interpreting those results, to not risk devising a true understanding of the social and behavior skill sets children possess. With constructs like this, instrument developers more often devise instruments that seek input from another proxy, or informant, as is viewed as developmentally appropriate in early childhood (Bagnato, 2005). This informant is asked to rate participant skill sets on behalf of the participant and is selected due to their qualifications related to the construct under review and/or their relationship with the participant (Saracho, 2017). This informant may engage in an interview or observation process to review the construct and observe behavior prior to rating. Importantly, the type of rating may be varied, with some instruments including yes/no responses such as, "Did the child smile when the caregiver left the classroom." Others may evaluate via a rating or Likert scale, asking questions like, "How often does the child smile when the caregiver leaves the classroom". An example of an instrument that is evaluated on a 5-point Likert scale is the Character Strengths Inventory of Early Childhood (Shoshani, 2019). This instrument is completed by a parent rater and includes an array of strength subscales such as kindness, perspective, curiosity, and creativity.

Questions that gauge frequency or initiate informant judgment may be more challenging to measure due to the risk of bias. Mitigating bias observed within the process of completing instruments is important, as bias can lead to error and can implicate the outcomes and interpretations, as well as the true understanding of skill sets. For example, within the field of education, error can implicate the use of data to devise instructional strategies.

Researchers, specifically within the early childhood field, have sought to understand the shared and unique differences between direct and indirect assessments (Berg-Nelson et al., 2012; Fuhs et al., 2015). Within this field, indirect assessments are often collected by an informant such as via teacher or caregiver ratings. More specifically, Fuhs et al. (2015) examined executive function skills of preschool-aged children. Their study sought to examine the different subskills encompassed within executive functioning, and how those skills correlate based on collection method. The researchers argue gauging children's skill sets via direct and indirect means collects different facets of learning. Direct assessments gauge available processes, whereas indirect assessments gauge usage of available processes (Fuhs et al., 2015). Other researchers have also sought to understand the differences between assessment types, more specifically, honing into the potential for bias. Berg-Nelson et al. (2012) studied informant (dis)agreement between teacher and parent ratings. Specifically, this study sought to understand whether different types of indirect assessment are better determinants of preschoolers' skill sets, particularly within the domain of mental health. The researchers found agreement between parent and teacher ratings; however, parents reported greater child problems and teachers reported lower ratings related to internalizing behaviors, which draws parallels to informant perception and the role of perception within rating.

Other researchers, like Baker et al. (2015), sought to understand teacher perception and the correlation between teacher rating and direct assessments. This research noted the potential role of perceptions in determining appropriate classroom and individual strategies based on developmental progress, and specifically highlighted the importance of accurate data to ensure adequate and accurate progress in developing student's skill sets. The researchers found, via hierarchical linear modeling, teacher perceptions related to various child characteristics explained approximately 40% of variability in direct assessments. Specifically, the researchers noted that teachers either over- or underestimated child performance when comparing to direct assessments. Lastly, the researchers (Baker et al., 2015) shared the impact perception has over time, via longitudinal hierarchical modeling, noting perceptions during the beginning of the school year associated significantly with student outcomes in spring. This research aligns to other work within the field, such as the study conducted by Kilday et al. (2012). These researchers studied children's mathematics skills and the association between teacher judgment via indirect ratings and direct assessment. These researchers found moderate association between the indirect ratings and direct assessments, noting teacher's ability to understand children's skill sets relative to the mean with their inability to accurately rate skill sets. This finding is important, as similar conclusions were drawn by Baker et al. (2015), in which teacher's knowledge of children's specific skill sets can implicate the strategies children receive within the classroom. Those strategies may not directly align to children's needs and therefore may not have lesser impact on their outcomes.

Finally, as researchers have found the strategies children receive within the classroom to be tied to teacher knowledge of child skill sets, a recent study found instructional lessons provided by the teacher was not highly predictive of end-of-year outcomes of Kindergarten students (Patrick et al., 2020). This study sought to understand predictive validity, via hierarchical linear modeling, of specific instructional lessons that were observed throughout the year, and whether those lessons predicted student outcomes. The researchers noted that there was statistical significance between observed instructional lessons and child outcomes; however, the total variability in child outcomes was very small.

Classical Test Theory

Classical test theory, described by Raykov and Marcoulides (2011) as a core method to behavioral measurement, is defined as the true score plus error equals the observed (test) score. The true score is the first of two parts that produce the observed score. This score is the actual score of the individual within the latent construct being evaluated. The second part, error, includes both systematic and random error of measurement. Systematic error is repeatable error when the instrument is conducted under similar conditions, both utilizing the identical instrument and for the same participants (Raykov & Marcoulides, 2011). Examples of systematic error include a child who is hearing-impaired misunderstands the instructions or an instructor mispronounces a word that is unfamiliar to the class. This type of error can lead to misrepresented conclusions, which implicates decision-making efforts. Random error is momentary error that happens by pure chance (Raykov & Marcoulides, 2011). This error also affects a participant's score and brings biases to the observed performance. This performance is implicated by the error, where the performance may be above or below actual levels. Examples of random error include fluctuations in an individual's physiological state due to the temperature in the room or an individual's mood affected by a recent personal situation or hardship. Both types of error implicate the observed score; therefore, implicating the conclusions and interpretations of the data.

With no means to separate true score from error of measurement when devising the observed score, the goal is to minimize error, specifically systematic error, to have the most accurate observed score. Within indirect assessment of skills, potential opportunities arise to implicate bias, including informant perception and memory. Within informant perception, biases, both conscious and unconscious, can be embedded within systematic error, and with no means to understand whether bias is included within error of a particular sample, we also cannot measure the amount of bias; therefore, seeking to minimize bias is prudent for accurate interpretation of the results.

Reliability

Within the journey of obtaining a well-defined instrument, reliability, defined as consistency within measurement (Leary, 2008), is typically amongst the first steps in studying the psychometric properties of an instrument. Reliability includes quantifying consistency within performance of an instrument and embedded within the definition of reliability are various forms of reliability, including: 1) measurement of form equivalence across alternate, or parallel, forms, 2) consistently measuring a construct over time, 3) measurement of the extent to which items consistently measure the construct, and 4) measurement of equivalence between raters or informants. Further, the most common means for estimating reliability is by an examination of the correlation between two instruments; therefore, instruments with higher reliability are less affected by measurement error and the sources that contribute to error.

Validity

Validity, defined as accuracy of measuring an underlying construct (Leary, 2008), is a second necessary psychometric property of an instrument. Validity seeks to measure the variability of scores within the characteristic, or construct, the researcher is seeking to measure. There are three types of validity: face validity, construct validity, and criterion-related validity. Face validity is the degree to which instruments appear to measure the construct of interest (Leary, 2008), and is studied through glancing or reviewing the instrument to determine if it appears to measure the desired construct. Construct validity assesses the relationship between two instruments (Leary, 2008), and is typically studied through correlations with other instruments of the same construct. There is no defined criteria for the size of the correlation, but rather, researchers evaluate the correlation relative to what one may typically find within the desired construct. Finally, criterion-related validity is the degree to which participants are distinguished within specific behavioral criterion of an instrument, both in the current state and future state (Leary, 2008), and is studied via concurrent and predictive validity. The difference between these two types of validity, concurrent and predictive, is the time that has elapsed between administrations (Leary, 2008). Concurrent validity studies the correlation between two instruments completed at the same time, whereas predictive validity studies the correlation between two an instrument completed in current time, and one completed in the future (Leary, 2008).

Considerations of Indirect Assessment

As researchers compare the results of indirect and direct assessment, continued considerations are warranted. Additional attention, specifically towards indirect assessment that incorporate rating skills, should be considered due to potential less objectivity and greater subjectivity that may underline the ratings (Dracobly et al., 2018; Paclawskyj et al., 2008). Informant's perception, memory, and depth of informant administration training to which informants are trained to consistently use the instrument, may implicate informant's judgment of participant skill sets; therefore, adding to the considerations advised for indirect assessments.

Informant Perception. Perception, seen as processes or senses which seek to understand various presented stimuli, contributes to one's decision-making efforts (Zeelenberg et al., 2006). Decisions are a part of everyday life, as individuals must decide which clothing to wear based on weather conditions, breakfast to consume, and speeds at which to operate a car. One's perception is influential in the decisions that are made. Within driving, one must perceive the rate at which traffic is moving and whether it is safe to change from one lane to another. Those perceptions will influence decisions and determine if one will arrive safely at their destination. From there, a new host of decisions must be made.

Beliefs implicate perception, decision making, and bias. Beliefs, as described by Vartuli and Rohs (2009), establish the basis for intentions and actions. Within education, beliefs implicate one's holistic classroom approach and teaching practices, which affect student engagement and outcomes. Further, in early childhood, researchers have identified variation in teacher beliefs. Namely, Di Santo et al. (2017) assessed teacher's pedagogical beliefs via the Teacher Belief Q-Sort. These researchers found that first year, undergraduate early childhood student beliefs related to children include a child-centered approach when teaching children, clear expectations related to behavior management are key to classroom management and showing respect towards children supports child behavior (Di Santo et al., 2017). These findings support continuous reflection of practice and beliefs to ensure skewness within perception and bias does not implicate how children are supported within the classroom.

Further within education, perception implicates interactions between the teacher and student. Within early childhood education, those interactions are key to effectively supporting children's development, and many tools have been devised to assess interactions within early childhood classrooms. Studies have found evidence to support the importance of quality interactions as they relate to various developmental and cognitive outcomes (Burchinal, et al., 2010; De Kruif et al., 2000). Further, embedded within these studies includes the role of teacher sensitivity towards children. Levels of sensitivity is also indicative of children's outcomes, and often work in parallel with interactions. Teacher-child relationships that are sensitive and supportive offer greater opportunity for quality interactions, which leads to better understanding of children's skill sets, just as Burchinal et al. (2008) found. These researchers observed that acquisition of skills was predicted by teacher sensitivity related to offering stimulating interactions and quality of instruction.

As teacher-student interactions and teacher sensitivity impact perception, conscious and unconscious biases are other facets of perception to which can implicate how informants' rate participant skill sets (Greenwald & Banaji, 1995). Legislation has been devised to disavow discrimination of various types, be it in the workplace or within school settings, and dismantling bias is embedded within these laws. Biases, both conscious and unconscious, can percolate through perception (Greenwald & Hamilton Krieger, 2006). Unconscious bias is often subconsciously filtrated through perception and can be indicative of judgments rendered (Greenwald & Banaji, 1995). Studies have been conducted to understand biases and attitudes related to race (Dovidio et al., 1997; Tate & Page, 2018). Both studies specify the influence of attitudes on behavior; specifically identifying, without realization, that individuals and situations are assessed, and judgments can be produced too quickly. Further, these judgments are often made without realization of one's own viewpoints and opinions, lacking awareness of the impact their judgments may have on others (Dovidio et al., 1997; Tate & Page, 2018). These instances are a frequent occurrence, and oftentimes, individuals lack acknowledgement that subconscious beliefs implicate the decisions made and behavior possessed. For individuals who are asked to complete an assessment of another individual, their biases may implicate their decisions, like the conclusions drawn from Quinn (2020). This research assessed racial bias within specified grading scales versus vague grading scales. The author found that White teachers rated, via a vague scale, writing samples of students lower when the student author was suggested to be Black. Additionally, when specified grading scales were utilized, the author did not find evidence related to racial bias; therefore, specific, detailed rating scales versus vague rating scales can be contributors to minimizing bias.

Informant Memory. Memory, one's ability to store new information and retrieve that information in the future, implicates one's accuracy in recollecting situations and behaviors (Loftus, 2003). Every second, individuals are inundated with inputs, yet very few bits of information can be stored. Further, memory is malleable, with inaccurate memories sometimes believed to be 'real' or accurate (Loftus, 2003). Recollecting memories and deciphering their accuracy can be challenging, and when instruments incorporate the use of memory, an additional layer of potential error may be assumed.

Within the classroom, teachers regularly seek to observe and memorize behaviors and situations to be able to recall those behaviors in the future when opportunities arise to plan and individualize instruction. In early childhood, observational methods of formative assessment collection are widely utilized, as this method is widely accepted as developmentally appropriate (Meisels et al., 2010). Within this form of assessment, teachers observe behavior, and oftentimes, hours or days later seek to retrieve that information when planning instructional strategies. Seeking to retrieve information for many students within a classroom can be challenging in and of itself; in addition, retrieving specific child behavior, skill sets, and skill mastery can be challenging.

For some behaviors, like observing children's social-emotional skills, additional features, like noticing both internalizing and externalizing behaviors is imperative. Internalizing behaviors, viewed as inner-directed behaviors (Madigan et al., 2013), are often behaviors that are more challenging to observe. These behaviors often include children who appear isolated and withdrawn, and recognizing those behaviors is more challenging. Oftentimes, relationships between caregiver or teacher and child are the best means for recognizing those behaviors. Within the classroom, children may appear as obedient because they may talk less or play alone; therefore, teachers may misinterpret internalizing behaviors because the child appears to be cooperative.

Opposite, externalizing behaviors include aggression, tantrums, or anger (Sulik et al., 2015), and are a form of behavior that can be viewed by others. Oftentimes, in moments of aggression or anger within young children, adults, including caregivers and teachers, intervene to share strategies to support the development of self-regulation skills. Outside of everyday instructional lessons to support the development of self-regulation skills, teachers oftentimes recognize and memorize externalizing behaviors because these are often bursts of extreme behavior that may harm the child or other children and adults within the classroom. Most likely, in moments of extreme behaviors, teachers intervene; therefore, recognizing and memorizing these behaviors may be easier.

In addition to recognizing certain behaviors, within the classroom, teacher's own emotions and self-regulation skills also intersect in supporting children's social-emotional skills via the relationships they facilitate (Cadima et al., 2016). As an informant of others behavior, recollecting and deciphering the variety of memories is indicative of the ratings this informant will render during data collection. Further, some instruments conduct an observation and immediately ask informants to complete ratings, whereas other instruments may conduct multiple, recurring observations before ratings are completed. These variations are important considerations within indirect assessment, as gauging one's ability to recollect all facets of memories, including the stable behavior, more inward behavior, and extreme behavior, can implicate how one rates behavior. Instrument developers must take this into consideration when devising informant guidelines, as the timeframe of observation can be an added source of error, thus implicating the observed score.

Interrater Reliability. Interrater reliability is an important consideration to instrument development, as this includes the process of ensuring consistency of ratings across raters (Leary, 2008). Instruments with pre-specified protocols, sentence stems, and correct answer options may be less challenging to gain reliability across raters, whereas other instruments that are vague, more ambiguous, or do not offer specific response options may have more challenges. Specifically, within opportunities in which judgment is rendered, instrument developers should seek to ensure enough specificity is offered to minimize the opportunity for bias or misunderstanding of protocols and procedures. Further, some instrument developers devise robust training manuals and certification tests, like the Classroom Assessment Scoring System (Pianta et al., 2008), which includes an intensive, two-day training with a credentialed trainer along with a certification test, while others may share some insights and information to raters without requiring certification. The DECA (LeBuffe & Naglieri, 2012) offers administration guidelines and directions related to completing a record form; however, no formalized training or certification is required prior to use. Five guidelines are included in the user's guide. These guidelines include completion of rating during a time in which the rater is free of distractions, ratings should be based on direct observation of the participant, participant behaviors that have occurred within the last four weeks should only be considered, raters should not compare participant ratings to other participants, and every item should be answered (LeBuffe & Naglieri, 2012). Finally, LeBuffe and Naglieri (2012) share the interrater reliability coefficients between teachers, namely between a teacher and teacher aide, for their testing sample. Fifty-two teachers were included, and the coefficients range from .36 to .77 by domain and subscales. All coefficients are statistically significant at $p < .01$.

Leary (2008) offers four means to decrease measurement error and increase reliability within behavioral measures. Those include devising standardized methods for administration, providing opportunity to clarify various instructions or other questions, carefully training observers or informants, and seeking to minimize opportunity for error within coding of data.

Application

Given the complexities and challenges that arise in conducting indirect assessments within the social sciences field (i.e., measuring a latent construct through an indirect process), further exploration of advanced, modern-day analytics is prudent during the instrument development and evaluation phase. These analytics, including differential item functioning and/or measurement invariance, offer opportunities for instrument developers to ensure the full range of psychometric properties have been conducted and verified. Importantly, the lack of tool validation by disaggregated groups can be challenging for practitioners to identify due to lack of in-depth knowledge and skill sets in accessing, reviewing, and interpreting technical manuals. Without these advanced analytics, instrument users, like practitioners, may conduct instruments without knowledge of appropriateness for their participant group and overarching goals for use of the results, like evaluation, screening purposes, or instruction. Further, they may lack robust knowledge regarding potential pitfalls or shortcomings within an instrument when seeking to interpret and use the data for continuous improvement efforts. Instrument developers hold the responsibility to provide detailed information in a timely fashion so that results produced by these instruments are valid, informative, and actionable.

Devereux Early Childhood Assessment

The DECA consists of 38 Likert scale questions related to two domains: total protective factors and behavior concerns. The first edition of this instrument was published in 1999 (LeBuffe & Naglieri, 1999), and the second edition was published in 2012 (LeBuffe & Naglieri, 2012). There are three versions to the instrument, including teacher, parent, and clinical versions. Informants are asked to rate participants related to the frequency with which specific behaviors are observed. The ratings are along a 5-point Likert scale, which includes the following response options: never, rarely, occasionally, frequently, and very frequently; however, responses within the behavior concerns domain are reverse scored. Finally, there are three forms to the instrument, including an infant, toddler, and preschool form.

The Devereux Early Childhood Assessment for Preschools User's Guide and Technical Manual (LeBuffe & Naglieri, 2012) describes 8 purposes regarding development and use of this assessment. These purposes include: 1) share the child's strengths and areas of need encompassed within the child's protective factors, 2) share the child's strengths and areas of need encompassed within the group's holistic protective factors, 3) promote resilience through strategy implementation within multi-tiered frameworks, 4) categorize children who exhibit problems, specifically emotional problems and behavioral problems, 5) promote collaboration amongst families and professionals, 6) evaluate program effectiveness related to resilience and competence, 7) support meeting Head Start Program Performance Standards, and 8) support research purposes through sharing of a measure that is psychometrically sound for protective factors (LeBuffe & Naglieri, 2012).

Various studies have been conducted with the DECA. These studies have ranged from assessing, via classical test theory and item response theory, the toddler version of this assessment within a sample of children from China (Liang et al., 2019) to comparing agreement of ratings between teachers and parents within an diverse sample of low-income children (Barbu et al., 2012; Crane et al., 2011). Other studies more directly relate to the current sample and research include assessment of test-retest reliability (Carlson & Voris, 2018), assessment of construct validity (Bulotsky-Shearer et al., 2013), assessment of reliability and validity (Lien & Carlson, 2009), and assessment of factor structure, as well as factor invariance by gender (Ogg et al., 2010).

Carlson and Voris (2018) assessed test-retest reliability to understand stability in the long-term of parent ratings within a Head Start sample, more specifically, comparing the ratings between the first and second edition. These researchers found moderate correlations when comparing the one-year ratings, noting the equivalency between the two editions. These statistically significant correlations included both the domains of total protective factors and behavior concerns, as well as the subscales, including initiative, self-regulation, and attachment/relationships, of the parent version of this instrument.

Bulotstky-Shearer et al. (2013) utilized exploratory and confirmatory factor analysis, as well as the Rasch partial credit model, to assess construct validity of the first edition of this instrument. These researchers found consistency within the subscales of the total protective factors domain; however, consistency was not replicated within the behavior concerns domain. Further, these researchers identified two factors within the behavior concerns domain, arguing lack of support for use of this domain within a diverse, low-income sample.

Lien and Carlson (2009) also studied the first edition, utilized data collected from three samples, including a community sample, children enrolled in Head Start, and the sample utilized for standardization, and compared the differences within item loadings of parent ratings. These researchers found reliability coefficients closely aligned between the sample of children enrolled in Head Start and the standardization sample. Further, the researchers compared means and standard deviations across the samples and found that the community sample and sample of Head Start children more closely aligned. Finally, the researchers conducted exploratory factor analysis with the Head Start sample and identified three factors within the total protective factors domain. When comparing the results of the Head Start sample to the standardization sample, the researchers identified the existence of similar factors; however, the items were found to load to different factors.

Finally, Ogg et al. (2010) studied the total protective factors domain within the first edition. Specifically, these researchers assessed the factor structure, as well as sought to understand how the instrument functions by gender, specifically within a sample of children enrolled in Head Start. First, the researchers assessed the intraclass correlation and found that all coefficients were less than .10, and therefore, did not conduct multilevel models since there were minor cluster effects. Next, a two-group confirmatory factor analysis for gender revealed that each item loaded to a single factor within each group and differential item functioning suggested evidence of invariance between gender groups. The results of this study indicated fit to the theoretical model, as well as support that the items within the assessment do function in a similar fashion across genders.

As a host of studies have sought to understand the psychometric properties of this instrument for use within various samples, none of which have specifically incorporated the use of longitudinal measurement invariance, as well as deeper analysis of appropriateness with a diverse, low-income population within the second edition of this assessment. Many studies have assessed various psychometric properties within the first edition, across various samples, and with various informant versions, including teacher report, parent report, and clinician report; however, no studies have specifically assessed invariance across gender, combined race and ethnicity, and dual language learning status, as well as longitudinal invariance across time.

Head Start

To combat poverty, Head Start was initiated by the Lyndon B. Johnson administration to support healthy development in children (Administration for Children & Families, 2021). At that time, the administration believed it must compensate for the conditions young children in poverty were facing, specifically socially and economically (Administration for Children & Families, 2021). Today, more than 1,600 programs are operating across the United States and over one million children are enrolled (Administration for Children & Families, 2020; Administration for Children & Families, 2021). While remaining focused on addressing the health development of young children living in poverty, the program now focuses holistically on development of the whole child, including physical development, social-emotional learning, executive function, language, literacy, and mathematics. Additionally, an abundance of research surrounds the academic success of low-income children (Brooks-Gunn & Duncan, 1997; Duncan et al., 1994), highlighting the value of programs like Head Start. Further, many programs seek to offer opportunities for parents and caregivers to engage in parenting, discipline, and other family-focused classes, and some programs also support adult enrollment in courses to attain technical certificates and degrees.

Not only are children enrolled in Head Start considered low-income by the federal poverty guidelines (U.S. Department of Health and Human Services, 2021), they too, are demographically diverse. According to the Head Start (2021) most recent report, 44% of children enrolled are White, 30% are Black or African American, and 37% identify as Hispanic or Latino. Being that many children across the United States are enrolled in Head Start programs, and the diverse demographics of those children enrolled, instruments that are utilized within those programs must be reliable and valid for that population.

Factor Analysis

Factor analysis seeks to determine the number of factors that account for variation amongst scores by analyzing the correlations between the items (Raykov & Marcoulides, 2011). A factor is a latent trait that is unobserved (Raykov & Marcoulides, 2011), and within the development of an instrument, this latent trait is the construct one is hypothesizing to be measuring. Factor analysis is a multivariate technique that assumes a causal structure is underlying the latent variables, and can be used for varying purposes, including use as a data reduction technique, search for the best factor structure, and using theory to devise a pre-specified factor model to drive examination of the factor structure (Raykov & Marcoulides, 2011). The equation of factor analysis is that of the regression slopes, *X*, and the vectors of τ, λ, ξ, and δ (Khojasteh & Lo, 2015). τ_k is the item intercepts, λ_k is the factor loadings, ξ_k is the score of the latent factor, and δ_k is the "unique factor variances in the *k*th group" (Khojasteh & Lo, 2015, 532).

$$X_k = \tau_k + \lambda_k\xi_k + \delta_k$$

There are numerous steps involved in conducting a factor analysis during instrument development. Once a developer has collected data via subject trials, they then conduct factor analyses to reduce the data and/or assess the structure of the data to ensure that the items fit a factor that the developer deems to be aligned with the construct of interest. Engaging in the steps to devising an instrument may be lengthy, and the developer may continue to make adjustments to the instrument and conduct multiple subject trials as he seeks to ensure the instrument adequate psychometric properties.

Structural Equation Modeling

Structural equation modeling is an approach to test hypothesized models, specifically testing the relationships between latent constructs and observed variables (Khine, 2013). Further, Khine (2013) posits four unique features to structural equation modeling, including use of a confirmatory approach by specifying relationships *a priori*, ability to assess and correct measurement error via error variance parameters, incorporation of both latent and observed variables, and ability to model multivariate relations and ability to estimate effects, both direct and indirect.

Khine (2013) shares four model types within structural equation modeling, including path analysis, confirmatory factor analysis, structural regression, and latent change. In addition, when testing structural equation models, five stages are included. Those stages are 1) model specification, meaning the researcher declares the relationships between the latent variable and observed variables, 2) identification, which is the determination of whether unique values within every free parameter are able to be obtained from the data, 3) estimation, seeking to produce an estimate model-implied covariance matrix that is parallel to the estimated covariance matrix of the sample, 4) evaluation of fit, meaning quality of fit to the hypothesized model, and 5) model modification, re-specifying the hypothesized model if fit is inadequate (Khine, 2013).

Measurement Invariance

As referenced by Millsap (2010), measurement invariance refers to a measure's psychometric property stability within populations and occasions. Further, Millsap (2010) explains that the premise of measurement invariance seeks to assess actual differences within the variable across groups, not differences of psychometric properties within the instrument. Equivalency between groups ensures that interpretations in observed change reflect actual change within the latent variable. The equation of measurement invariance is that of probability, *P*, with the vectors of X, W, and V (Khojasteh & Lo, 2015). X is the measured variable, W is the latent variable, and V is the population indicators which represent group membership (Khojasteh & Lo, 2015).

$$P\ (\mathrm{X}|\mathrm{W},\ \mathrm{V}) = P\ (\mathrm{X}|\mathrm{W})$$

Longitudinal measurement invariance seeks to understand change over time, and with that, observed changes that are due to actual changes in the variable, not change within the psychometric properties (Millsap, 2010). To that end, if a variable is variant, an observed change can be implicated by a change in instrument and its underlying variable, which is undesirable (Millsap, 2010).

Longitudinal measurement invariance has been broadly included within analyses related to social sciences, namely within constructs related to life satisfaction (Esnaola et al., 2019), memory and executive function (Moreira et al., 2018), depressive symptoms (Guo et al., 2017), stress (Suh et al., 2016), and posttraumatic stress disorder (Contractor et al., 2017). These studies have each sought to ensure psychometric soundness, both across populations and across occasions. Importantly, these studies all directly measured the construct.

There are four sequential steps to evaluating measurement invariance, including assessment of configural invariance, metric invariance, scalar invariance, and strict invariance (Millsap, 2007). Each step builds upon the prior step. Configural invariance seeks to study the pattern of factor loadings to ensure equivalency of the factor structure between groups. Next, the metric invariance seeks to study the equality of factor loadings between groups. Scalar invariance seeks to study equality of item intercepts in addition to factor loadings between groups. Finally, strict invariance seeks to study equality of item variances in addition to item intercepts and factor loadings between groups. Through the process of evaluating measurement invariance, the researcher conducts each step and determines whether equivalence holds, or a re-specification of the model is needed to support equivalence or partial invariance. If equivalence or partial invariance cannot be confirmed within a step, the process is discontinued, with the goal to confirm equivalence through the final step of strict invariance.

The Present Study

The present study seeks to understand the factor structure and investigate invariance models of a social-emotional instrument used in preschool classrooms and across Head Start grantees, the DECA, second edition (LeBuffe & Naglieri, 2012). This study seeks to explore and verify the factor structure via exploratory factor analysis and confirmatory factor analysis. Next, this study seeks to evaluate, via measurement invariance, the appropriateness of this assessment for specific subgroup characteristics related to the following variables: gender, combined race and ethnicity, and dual language learning status. In addition, this study will also seek to assess longitudinal measurement invariance by time (fall to spring). Finally, to conclude this study, a guide for practitioner use will be established. This guide is specific to use within indirect assessments, and seeks to provide a methodical approach to reviewing, understanding, and interpreting instrument technical manuals. This will provide clarification regarding appropriateness of use specific to outlined goals devised by practitioners.

CHAPTER III

METHODOLOGY

Background

Two evaluations of a Head Start program were previously conducted via a research-practice partnership between the staff of the Head Start agency and a research team at a local university in Tulsa, Oklahoma. The first evaluation was conducted during the 2014-2015 and 2015-2016 school years. The second evaluation was conducted during the 2018-2019 school year. Both evaluations sought to understand the outcomes of children enrolled in the Head Start program and results of the evaluation were used to promote outcome-related continuous improvement efforts. This Head Start program operates 10 schools across the county and serves approximately 2,300 children annually through center-based and home-based services.

Within both evaluations, student data related to various domains of development, including social-emotional learning, expressive language, receptive language, literacy, mathematics, and executive function were collected in fall and spring. Finally, reliable observers conducted classroom observations in the winter to understand the quality of early childhood classrooms, as well as to understand the experiences of children and teachers within these classrooms.

To determine the student sample, the team first reviewed the projected total number of classrooms, then randomly selected Preschool 3 and Pre-Kindergarten classrooms, ensuring every school was represented. Next, within those selected classrooms, the team randomly selected 6 to 8 student participants. Further, selection of classrooms and participants occurred at the beginning of the school year; therefore, participants with data in fall and spring were enrolled throughout the duration of the school year. More specifically, the assessment utilized within this analysis was conducted approximately 6-8 weeks after the beginning of the school year, and repeated, approximately 6 months after the pre-assessment.

Participants

Participants of this study include three-year-old and four-year-old students enrolled in a Midwestern, urban Head Start program. Participants were randomly selected within the classrooms that had been selected to be included in the study.

A total of 785 unique student participants were included in this study, utilizing an archival dataset. These participants had data in both fall and spring of the school year in which they participated in the study. Further, across the three years of these studies, 65 participants had data as both three-year-olds and four-year-olds; therefore, a total of 1,700 observations, which includes 850 observations in fall and 850 observations in spring, were collected and included in this analysis.

Of the 785 unique participants, 54.3% were male and 45.7% were female. Utilizing a combined race and ethnicity variable, which aligns to the national Head Start evaluation known as the Family and Child Experiences Survey (Administration for Children & Families, ND), 43.2% were Hispanic/Latino, 27.8% were African American, Non-Hispanic, and 13.6% were White, Non-Hispanic. Finally, based on results of the English oral language proficiency screener, 63.6% of participants were coded English Monolingual, 23.2% of participants were coded as bilingual, and 13.2% of participants were coded as dual language learners. More information related to this screener and the determination of oral language proficiency status is provided in the instruments section. Finally, detailed descriptive results are shared in Table 1.

Table 1

Descriptive results

	Frequency	Percent	Cumulative Percent
Total Observations	1700		
Total Participants	785		
Observations by Study Year			
2014-2015	235	27.6	27.6
2015-2016	237	27.9	55.5
2018-2019	378	44.5	100.0
Number of Years Participant Included in Study			
1 year	720	91.7	91.7
2 years	65	8.3	100.0
Age of Participants by Study Year Observation			
3-years-old	438	51.5	51.5
4-years-old	412	48.5	100.0
Participant Gender			
Male	426	54.3	54.3
Female	359	45.7	100.0
Participant Race + Ethnicity			
African American, Non-Hispanic	218	27.8	27.8
American Indian or Alaska Native	21	2.7	30.4
Asian or Pacific Islander	30	3.8	34.3
Hispanic/Latino	339	43.2	77.5
Multi-Racial/Biracial, Non-Hispanic	46	5.9	83.3

	Frequency	Percent	Cumulative Percent
Other	24	3.1	86.4
White, Non-Hispanic	107	13.6	100.0
Participant English proficiency			
Dual Language Learner	104	13.2	13.2
Bilingual	182	23.2	36.4
English Monolingual	499	63.6	100.0
Informant- Teacher Position			
Lead Teacher	1525	89.7	89.7
Assistant Teacher	155	9.1	9.1
Other/Unknown	20	1.2	100.0

Teacher Informants

In addition to the student participants, 89.7% of the teacher informants were lead teachers in the program. Additionally, 9.1% of the teacher informants were assistant teachers, and 1.2% were considered other/unknown.

Instruments

Devereux Early Childhood Assessment

The DECA, second edition, consists of 38 Likert scale questions related to two domains: total protective factors and behavior concerns (LeBuffe & Naglieri, 2012). This study analyzes the preschool form, by which the teacher is the informant. Informants are asked to rate participants related to the frequency with which specific behaviors are observed and are rates along a 5-point Likert scale.

Embedded within the total protective factors domain are three subscales. Those subscales include initiative, self-regulation, and attachment/relationships. Devereux Early Childhood Assessment for Preschoolers (ND) defines initiative as a child's skill to meet their own needs by utilizing independent thinking and action. The researchers also define self-regulation as a child’s skill to healthily express feelings and control behaviors (Devereux Early Childhood Assessment for Preschoolers, ND). Finally, attachment/relationships is defined as a child’s skill to encourage and sustain mutual connections that are positive with children and adults (Devereux Early Childhood Assessment for Preschoolers, ND).

All questions begin with the sentence stem, “During the past 4 weeks, how often did the child…”. One full question, which includes the sentence stem and statement ending, states, “During the past 4 weeks, how often did the child try different ways to solve a problem?” This is an example within the initiative subscale. Other questions (statement endings) related to the initiative subscale include show an interest in learning new things and make decisions for himself/herself. Within the self-regulation subscale, example questions include listen to or respect others and control his/her anger. Within the attachment/relationship subscale, sample questions include show affection to familiar adults and show a preference for a certain adult, teacher, or parent. Finally, within the behavior concerns subscale, example questions include seem sad or unemotional at a happy occasion and seem uninterested in other children or adults.

Nine questions are included within each subscale of the total protective factors domain. Additionally, 11 questions are included within the behavior concerns domain. Figure 1 provides an overview of the factor structure of this instrument and is provided in the appendix.

Finally, this instrument is completed by an informant who conducts observations of the participant throughout a four-week period prior to completing the instrument. The instrument utilized within this study is specific to observations and ratings conducted by the participant's teacher, and therefore, can be categorized as an indirect assessment.

Pre-IPT Oral Test

The Pre-IPT oral test, fourth edition, is an English oral language screener, for children ages three-years-old to five-years-old (Ballad & Tighe, 2010). To begin the screener, participants are asked to answer specific questions from short stories and pictures that are provided. The evaluator points to various parts of the picture and asks the participant questions such as, "This is Sarah's father and this is her ___." The child is asked to respond to these statements, and the evaluator codes for correct and incorrect responses.

At the end of each level, the evaluator calculates the total errors, or incorrect responses. The number of total errors determines whether the screener is either concluded or continued at the end of each level. Based on progress made within the screener and the total number of errors, the final score level is either deemed level A, B, C, D, or E. A score level of A is associated with the oral proficiency designation of beginner, B is designated as early intermediate, C is designated as intermediate, D is designated as early advanced, and E is designated as advanced. This screener includes five levels, beginning with level A

Within this study, participants whose home language, as determined by language provided at enrollment, was not English received the screener in fall of the study year.

Participants whose English oral language score level upon completion of the screener was coded as level A were designated as “dual language learner” within this study. Participants whose English oral language was coded as level B, C, D, or E were designated as “bilingual” within this study. Finally, participants whose home language was English were coded as “English monolingual”.

The Present Study

Research Goals

1. Explore the structure of the data via exploratory factor analysis.
2. Evaluate the structure of the data via confirmatory factor analysis.
3. Investigate measurement invariance of fall data across the following groups:
    a. Gender, including Male and Female
    b. Combined race and ethnicity, including White/Non-Hispanic, Hispanic, and African American/Non-Hispanic
    c. Dual language status, including Dual language learner, Bilingual, English monolingual
4. Investigate longitudinal measurement invariance by time
5. Establish a standard protocol for indirect assessments.

Analytic Plan

To answer the research questions described in this study, exploration and confirmation of the data structure via exploratory factor analysis and confirmatory factor analysis were conducted first. The full dataset of 1,700 observations was randomly split into two datasets to answer research questions one and two. In addition, SPSS software

(IBM Corp, 2016) was utilized to conduct the exploratory factor analysis, and R software (R Core Team, 2017) with the ‘lavaan’ package (Rosseel, 2012) was utilized to conduct the confirmatory factor analysis.

Evaluating the appropriateness of the exploratory factor analysis included the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy, Bartlett’s Test of Sphericity, and the determinant of the correlation matrix. Further, four strategies were deployed to determine factor retention. Those strategies include studying eigenvalues that are >1.0, determining the percent of variance extracted, review of the scree plot, and a parallel analysis.

Evaluating the fit of the confirmatory factor analysis included the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). Additionally, the Akaike (AIC) index was studied to compare the two models, specifically seeking the smaller index to indicate the better fitting model. To evaluate the local fit of the confirmatory factor analysis models, the normalized residuals was evaluated. Specifically, the normalized residual matrix was analyzed to study residual covariance estimates with other items that are greater than |2|. Items with residual covariance estimates greater than |2| were reviewed for local fit, as items greater than |2| indicates significant misfit.

Next, to answer the third research question, the dataset investigated measurement invariance by the following characteristics: gender, combined race and ethnicity, and dual language learner status. This analysis utilized data collected in fall of study year to establish invariance by each characteristic. More specifically, within the realm of gender, both male and female variables were included. Within combined race and ethnicity, White/Non-Hispanic, Hispanic, and African American/Non-Hispanic variables were included due to the larger group size. Within the dual language learner status variable, dual language learner, bilingual, and English monolingual were included. The grouping variable of time, fall and spring, was utilized to answer the fourth research question investigating longitudinal measurement invariance. RStudio software, version 3.6.3 (RStudio, 2020), with the ‘lavaan’ package (Rosseel, 2012) was utilized to conduct the measurement invariance and longitudinal measurement invariance analyses.

To conduct each invariance analysis, the following tests assessed model fit. First the chi-square difference ($\Delta\chi^2$) test evaluated reductions of statistical significance within model fit. Next, the delta goodness-of-fit indices (ΔGOF) was reviewed as descriptive indicators to support the chi-square difference test. These indices include the change in comparative fit index (ΔCFI) of less than .002, change in root mean square error of approximation (ΔRMSEA) of less than .01, and the change in standard root mean square residual (ΔSRMR) of less than .02 (Khojasteh & Lo, 2015).

Finally, a practitioner guide for reviewing social science instrument was established for practitioner use. This protocol sought to support practitioners in reviewing and interpreting assessment technical manuals.

Practitioner guide.

The practitioner guide offers practitioners with a quick, step-by-step tutorial for reviewing and interpreting technical manuals. Since practitioners may or may not have specific knowledge and skill sets to understand whether an instrument is appropriate for their sample and use, the goal of this protocol seeks to offer guidance to interpretation of technical manuals and ensure relevance and appropriateness for practitioner participant groups. More information related to the practitioner guide, as well as the guide itself, is

Variables

Table 1 provides a description of each variable included within research question 3, investigating measurement invariance, as well as a few additional variables pertinent to the data cleaning process.

Table 2

Variable descriptions

Variable	Description
Participant Age	Age calculated as of September 1st of study year
Primary Language	At application, home language of the participant
Ethnicity	At application, ethnicity of the participant
Gender	At application, gender of the participant
Race	At application, race of the participant
Combined Race + Ethnicity	Based on a combination of application race and ethnicity variables
Dual Language Status	Based on English oral proficiency

CHAPTER IV

FINDINGS

To answer the research questions described in this study, exploration and confirmation of the data structure via exploratory factor analysis and confirmatory factor analysis were conducted first. Prior to beginning the exploratory factor analysis, questions that were intended to measure behavior concerns were reverse coded so that a rating of 1 was reversed to 5, a rating of 2 was reversed to 4, a rating of 4 was reversed to 2, and a rating of 5 reversed to 1 so that for all responses a higher number corresponded to higher frequency of behavior and a lower number corresponded to lower frequency of behavior. Next, the full dataset of 1,700 completed observations was randomly split by generating a numerical value, ranging from .0007 to .9997, for each sample observation, then determining a threshold value, which was found to be .4825, that would evenly split the full sample into two datasets, and finally, splitting the datasets into two by selecting those observations whose numerical value was equal to or below the threshold for use within the exploratory factor analysis and those observations whose numerical value was above the exploratory factor analysis and those observations whose numerical value was above the threshold for use within the confirmatory factor analysis. Each split dataset included 850 sample observations. Finally, exploratory and confirmatory factor analyses were conducted.

Next, the model deemed to fit best was utilized to conduct measurement invariance analyses by gender, race and ethnicity, and dual language status with fall observation data. The full dataset with fall observations was utilized as this aligned to other relevant studies conducted with the DECA (Bulotstky-Shearer et al., 2013; Lien & Carlson, 2009; Ogg et al., 2010). Finally, longitudinal measurement invariance was conducted utilizing the time variable, comparing fall and spring and utilizing the full dataset of observations.

Descriptive Statistics

The descriptive statistics of the DECA are provided in Table 3. The highest mean was 3.73 within item 35, "During the past 4 weeks, how often did the child touch children or adults in a way that you thought was inappropriate?" This question was reverse coded; therefore, this average equated to informants rating participants as rarely or occasionally exhibiting this behavior. The lowest mean was 2.02 within item 30, "During the past 4 weeks, how often did the child get easily distracted?" This average equated to informants rating participants as rarely exhibiting this behavior. The standard deviations ranged from 0.697 for item 35 to 1.20 for item 6.

Table 3
Descriptive Results

During the past 4 weeks, how often did the child. . .	*M*	*SD*
1: act in a way that made adults smile or show interest in him/her?	3.16	0.803
2: listen to or respect others?	2.80	0.911
3: control his/her anger?	2.77	1.007
4: seem sad or unemotional at a happy occasion?*	2.99	0.950
5: show confidence in his/her abilities?	2.78	0.947
6: have a temper tantrum?*	2.90	1.200
7: keep trying when unsuccessful (show persistence)?	2.57	0.858
8: seem uninterested in other children or adults? *	3.11	0.969
9: use obscene gestures or offensive language? *	3.60	0.836
10: try different ways to solve a problem?	2.34	0.857
11: seem happy or excited to see his/her parent or guardian?	3.46	0.722
12: destroy or damage property?*	3.43	0.949
13: try or ask to try new things or activities?	2.61	0.924
14: show affection for familiar adults?	3.19	0.807
15: start or organize play with other children?	2.63	1.002
16: show patience?	2.52	0.968
17: ask adults to play with or read to him/her?	2.67	1.006
18: have a short attention span?*	2.21	1.150
19: share with other children??	2.60	0.869
20: handle frustration well?	2.38	1.023
21: fight with other children?*	2.59	1.060
22: become upset or cry easily?*	2.38	1.086
23: show an interest in learning new things?	2.86	0.839
24: trust familiar adults and believe what they say?	3.13	0.800
25: accept another choice when his/her first choice was not available?	2.67	0.911
26: seek help from children/adults when necessary?	3.01	0.745
27: hurt others with actions or words?*	2.91	1.085
28: cooperate with others?	2.74	0.820
29: calm himself/herself down/	2.56	0.901
30: get easily distracted?*	2.02	1.127
31: make decisions for himself/herself?	2.94	0.760
32: appear happy when playing with others?	3.23	0.819
33: choose to do a task that was hard for him/her?	2.35	0.907
34: look forward to activities at home or school?	3.19	0.823
35: touch children or adults in a way that you thought was inappropriate?*	3.73	0.697
36: show a preference for a certain adult, teacher, or parent?	2.28	1.090
37: play well with others?	2.91	0.930
38: remember important information?	2.74	1.056

Note. All items were on a 5-point Likert-type response scale where a lower number corresponded to lower frequency of behavior and a high number corresponds to higher frequency of behavior.

A correlation matrix was analyzed to ensure moderate to high correlations among the items. The correlation matrix of the DECA is provided in Table 4 in the appendix. Most items had moderate correlations with one another; however, many items had correlations of <.20. Namely, item 36 had very low correlations with most other items. Based on the correlation matrix, conducting an EFA is appropriate.

Exploratory Factor Analysis

An exploratory factor analysis (EFA) of the DECA was conducted using SPSS, version 24, software (IBM Corp, 2016). EFA utilizing both principal axis factoring and maximum likelihood were conducted, with the results being substantively similar. Specifically, the results were similar within the number of eigenvalues >1.0, cumulative variance extracted, the scree plot, and factor loadings alignment with factors. For brevity, only maximum likelihood results are presented. The KMO Measure of Sampling Adequacy was .946, which indicates the data are good for structure detection. The Bartlett's Test of Sphericity was .000, which indicates that EFA is useful in detecting the structure. Finally, the determinant of the correlation matrix was 1.041E-11, which is within appropriate range; therefore, it is appropriate to conduct an EFA. An initial EFA, via maximum likelihood, was conducted to study the eigenvalues, amount of shared variance, and scree plots.

Initially, four eigenvalues were >1.0, with cumulative variance equating to 53.99%. The scree plot supported four eigenvalues were >1.0. Finally, a parallel analysis was conducted to before determining the final number of factors to extract. The parallel analysis determined that three factors should be extracted. Based on review, four factors were extracted for further analysis due to theoretical alignment with the four subscales of the DECA.

An oblique, via promax, rotation was conducted to develop a simple structure. After review of the factor loadings, the oblique rotation was determined to provide the best fit. The factor loadings and communalities based on a four-factor analysis via the promax rotation is provided in Table 5. Additionally, the rotation sums of square loadings for each factor is included.

Table 5

Factor Loadings & Communalities via Promax Rotation

	Factor					
	1	2	3	4	Communality	Factor Specified
1			0.63		0.55	Attachment/Relationships
2	0.60				0.66	Self-Regulation
3	0.77				0.69	Self-Regulation
4			0.31		0.22	Attachment/Relationships
5		0.60			0.47	Initiative
6	0.91				0.68	Self-Regulation
7		0.66			0.57	Initiative
8			0.45		0.30	Attachment/Relationships
9	0.52				0.31	Self-Regulation
10		0.74			0.60	Initiative
11			0.39		0.20	Attachment/Relationships
12	0.56		0.36		0.57	Self-Regulation
13		0.82			0.58	Initiative
14		0.31	0.46		0.37	Attachment/Relationships
15		0.53			0.52	Initiative

	1	2	3	4	Communality	Factor Specified
16	0.72				0.69	Self-Regulation
17		0.56			0.40	Initiative
18				0.72	0.80	Behavior Concerns
19	0.56				0.65	Self-Regulation
20	0.87				0.77	Self-Regulation
21	0.73				0.59	Self-Regulation
22	0.66				0.45	Self-Regulation
23		0.79			0.67	Initiative
24			0.60		0.63	Attachment/Relationships
25	0.71				0.68	Self-Regulation
26		0.40			0.39	Initiative
27	0.76	-0.33			0.56	Self-Regulation
28	0.63				0.62	Self-Regulation
29	0.70				0.55	Self-Regulation
30				0.75	0.76	Behavior Concerns
31		0.65			0.45	Initiative
32			0.72		0.65	Attachment/Relationships
33		0.73			0.55	Initiative
34		0.58			0.43	Attachment/Relationships
35			0.53		0.33	Attachment/Relationships
36	-0.38		0.47		0.19	Attachment/Relationships
37	0.46		0.57		0.75	Self-Regulation*
38		0.49	0.38		0.56	Initiative
Rotation Sums of Square Loadings	12.02	10.62	8.68	4.00		

Note. Asterick represents factor specified for model 2; item was specified to factor "attachment/relationships" within model 1

Of the 38 items, 6 items cross-loaded to two factors. Those items include 12, 14, 27, 36, 37, and 38. All other factors loaded to a single factor. Fifteen items load to the first factor. Further, utilizing the subscales described by the DECA, the first factor may be described as self-regulation. The proportion of variance extracted for factor one was 12.02, which was the highest of the four factors. Thirteen items loaded to the second factor, which may be described as initiative. The proportion of variance extracted for the second factor was 10.62. Twelve factors loaded to the third factor, which may be described as attachment/relationships. The proportion of variance extracted for the third factor was 8.68. Finally, two items loaded to the fourth factor, which may be described as behavior concerns. The proportion of variance extracted for the fourth factor was 4.00.

Finally, review of the correlation across factors, due to use of the promax rotation, indicates a range in correlations between factors. Factors one and two have the highest correlation of 0.542, whereas factors three and four have the lowest correlation of 0.259.

Review of the communalities shows varying levels of shared variance, ranging from 0.19 to 0.80. The highest shared variance was item 18, and the lowest shared variance was item 36. Overall, the majority of communalities display moderate to high levels of shared variance.

Review of the reproduced correlation matrix shows that most items fit well within the model. The largest correlation from the reproduced correlation matrix was 0.78 between items 18 and 30. Additionally, 23 items correlations produced residuals >.05. This equated to 3.3% of the residual correlations. The greatest residual correlation was 0.17 between items 4 and 8. Overall, most residuals were very low.

Twenty-six of the 38 factor loadings of the exploratory factor analysis aligned with the subscales described by the DECA. Of the 11 items within the behavior concerns domain, only two items loaded to that factor. Of the 9 items from the behavior concerns domain that loaded to a separate factor, 6 of those items loaded to the self-regulation factor, zero loaded to initiative, and three loaded to attachment/relationships.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA), of the DECA was conducted using R, version 3.4.0 (R Core Team, 2017), software and the lavaan package (LAtent VAriable ANalaysis) (Rosseel, 2012). Due to the cross-loading of items within the EFA, two higher-order models were analyzed and compared. To devise the two models for comparison, the cross-loaded items were initially reviewed. Within five of the 6 items that cross-loaded, the higher loading of each item aligned with the model described by the subscales of the DECA; therefore, the higher item loading from the EFA was included in the CFA models. For item 37, the higher loading did not align with the subscale described by the DECA; therefore, two models were tested and compared that assessed better fit based on item 37. In model 1, item 37 was included in attachment/relationships factor. In model 2, item 37 was included in the self-regulation factor. Finally, it is important to note that 11 items loaded to factors that did not align with the subscales described by the DECA, and for this analysis, the factor to which the items loaded was used for analytical purposes.

To begin assessing model fit, the CFI, TLI, RMSEA, and SRMR criteria were compared. The global fit indices are included in Table 6.

Table 6
Global Fit Indices

	Model 1	Model 2	Criteria for good fit
Comparative Fit Index (CFI)	0.807	0.808	> 0.95
Tucker-Lewis Index (TLI)	0.794	0.795	> 0.95
Root Mean Square Error of Approximation (RMSEA)	0.084	0.084	< 0.05
Standardized Root Mean Square Residual (SRMR)	0.081	0.079	<0.08

	Model 1	Model 2	Criteria for good fit
Akaike (AIC)	69152.47	69144.32	

According to the global fit indices, the two models were similar in comparison, with model 2 fitting slightly better within CFA, TLI, and SRMR.

Within local fit, since model 2 global fit indices fit slightly better, the normalized residual matrix within model 2 was studied. Estimates greater than |2| indicate significant misfit. Across all 38 items, no items had at least one residual covariate estimates greater than |2| with another item. Further, the variance accounted for by each item is provided in Table 7. The highest R^2 value is 0.825 within item 30. The lowest R^2 value is 0.084 within item 36. Additionally, 32 items have an R^2 >.30, and 6 items have an R^2 <.30.

Table 7
Variance Accounted for (R^2)

	R^2
1: act in a way that made adults smile or show interest in him/her?	0.546
2: listen to or respect others?	0.630
3: control his/her anger?	0.625
4: seem sad or unemotional at a happy occasion?	0.261
5: show confidence in his/her abilities?	0.557
6: have a temper tantrum?	0.524
7: keep trying when unsuccessful (show persistence)?	0.604
8: seem uninterested in other children or adults?	0.426
9: use obscene gestures or offensive language?	0.139
10: try different ways to solve a problem?	0.669
11: seem happy or excited to see his/her parent or guardian?	0.248
12: destroy or damage property?	0.408
13: try or ask to try new things or activities?	0.651
14: show affection for familiar adults?	0.432
15: start or organize play with other children?	0.505

	R^2
16: show patience?	0.633
17: ask adults to play with or read to him/her?	0.365
18: have a short attention span?	0.744
19: share with other children??	0.600
20: handle frustration well?	0.619
21: fight with other children?	0.474
22: become upset or cry easily?	0.397
23: show an interest in learning new things?	0.618
24: trust familiar adults and believe what they say?	0.576
25: accept another choice when his/her first choice was not available?	0.573
26: seek help from children/adults when necessary?	0.242
27: hurt others with actions or words?	0.423
28: cooperate with others?	0.593
29: calm himself/herself down/	0.443
30: get easily distracted?	0.825
31: make decisions for himself/herself?	0.431
32: appear happy when playing with others?	0.599
33: choose to do a task that was hard for him/her?	0.606
34: look forward to activities at home or school?	0.387
35: touch children or adults in a way that you thought was inappropriate?	0.129
36: show a preference for a certain adult, teacher, or parent?	0.084
37: play well with others?	0.557
38: remember important information?	0.542

Overall, the two models were similar in comparison. Because model 2 fit slightly better, model 2 was selected for use within the measurement invariance models. The standardized factor loadings for model 2 are provided in Table 8. The factor structure for model 2 is displayed in Figure 2 in the appendix.

Table 8
Standardized Factor Loadings

	Factor Loading
1: act in a way that made adults smile or show interest in him/her?	0.738

	Factor Loading
2: listen to or respect others?	0.798
3: control his/her anger?	0.790
4: seem sad or unemotional at a happy occasion?	0.505
5: show confidence in his/her abilities?	0.750
6: have a temper tantrum?	0.723
7: keep trying when unsuccessful (show persistence)?	0.773
8: seem uninterested in other children or adults?	0.646
9: use obscene gestures or offensive language?	0.378
10: try different ways to solve a problem?	0.813
11: seem happy or excited to see his/her parent or guardian?	0.498
12: destroy or damage property?	0.643
13: try or ask to try new things or activities?	0.807
14: show affection for familiar adults?	0.658
15: start or organize play with other children?	0.713
16: show patience?	0.797
17: ask adults to play with or read to him/her?	0.608
18: have a short attention span?	0.907
19: share with other children??	0.774
20: handle frustration well?	0.783
21: fight with other children?	0.692
22: become upset or cry easily?	0.627
23: show an interest in learning new things?	0.786
24: trust familiar adults and believe what they say?	0.761
25: accept another choice when his/her first choice was not available?	0.753
26: seek help from children/adults when necessary?	0.492
27: hurt others with actions or words?	0.655
28: cooperate with others?	0.77
29: calm himself/herself down/	0.661
30: get easily distracted?	0.864
31: make decisions for himself/herself?	0.655
32: appear happy when playing with others?	0.779
33: choose to do a task that was hard for him/her?	0.777
34: look forward to activities at home or school?	0.624
35: touch children or adults in a way that you thought was inappropriate?	0.352
36: show a preference for a certain adult, teacher, or parent?	0.289
37: play well with others?	0.746

	Factor Loading
38: remember important information?	0.741

Measurement Invariance

In accordance to procedures Vandenburg and Lance (2000) recommend, the first model conducted evaluated the factor structure of all sample participants. Upon completion of the first model, group factor structure (i.e., gender, race and ethnicity, and dual language status invariance analysis) was then assessed. Further, across all of the invariance analyses, factor loadings were constrained to be equal during the evaluation of metric invariance and the referent indicator within the invariance analyses included the loading of the first factor.

For the purpose of clarification, metric refers to constraining of loadings, scalar refers to constraining of intercepts, and strict refers to constraining of residuals (Vandenburg & Lance, 2000). All analyses were conducted in RStudio, version 3.6.3 (RStudio, 2020), using the lavaan (Rosseel, 2012) and semTools (Useful Tools for Structural Equation Modeling) (Jorgensen et al., 2021) packages.

Descriptive Statistics

Descriptive statistics, including means and standard deviations, by item and invariance models are presented in Table 9 in the appendix.

Gender

Configural invariance of gender did not fit the data well, CFI = 0.737, SRMR = 0.100, and RMSEA = 0.097; therefore, configural invariance was not met, $\Delta\chi^2$ (661) = 853.64, $p < .001$. Since configural invariance was not met, no further analyses of this model were conducted.

Combined Race and Ethnicity

Configural invariance of race and ethnicity also did not fit the data well, CFI = 0.766, SRMR = 0.09, and RMSEA = 0.099; therefore, configural invariance was not met, $\Delta\chi^2$ (1322) = 1703.2, $p < .001$. Again, since configural invariance was not met, no further analyses of this model were conducted.

Dual Language Status

Configural invariance of dual language status also did not fit the data well, CFI = 0.737, SRMR = 0.099, and RMSEA = 0.099; therefore, configural invariance was not met, $\Delta\chi^2$ (1322) = 1707.3, $p < .001$. Again, since configural invariance was not met, no further analyses of this model were conducted.

Longitudinal Measurement Invariance

As shared above, longitudinal measurement invariance via time (fall and spring) followed the same recommended model evaluation criteria. Configural invariance of time did not fit the data well, CFI = 0.799, SRMR = 0.084, and RMSEA = 0.086; therefore, configural invariance was not met, $\Delta\chi^2$ (661) = 970.7, $p < .001$. Again, since configural invariance was not met, no further analyses of this model were conducted.

Table 10

Invariance Analysis by Gender, Dual Language Status, Race + Ethnicity, and Time for the DECA

	χ^2	*df*	$\Delta\chi^2$	Δdf	SRMR	CFI	RMSEA (90% CI)
Configural (by gender)	6595.2	1322	853.64***	661	0.100	0.737	0.097 [.095, .099]
Configural (by race and ethnicity)	6069.0	1983	1703.2***	1322	0.090	0.766	0.093 [.090, .095]
Configural (by dual language status)	7448.9	1983	1707.3***	1322	0.099	0.737	0.099 [.096, .101]
Configural (by time)	9682.5	1322	970.7***	661	0.084	0.799	0.086 [.085, .088]

Note. DECA = Devereux Early Childhood Assessment; SRMR = standard root mean square residual; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval.

***p < .001

CHAPTER V

DISCUSSION

Biases posed by informant perception and memory can implicate the usefulness of instruments, specifically those collected indirectly. This study sought to highlight the challenges developers of social science instruments must tackle, while also serving amongst the first to conduct advanced psychometric testing, specifically conducting measurement invariance and longitudinal measurement invariance, of the DECA. Further, this study is the first to provide a deeper evaluation of appropriateness of use within a diverse, low-income population, specifically within the second edition of this instrument.

Methodologically, this study included a random sample of participants and conducted detailed steps, including exploratory factor analysis and confirmatory factor analysis, before analyzing the measurement invariance and longitudinal measurement invariance models. Further, the variables included in the invariance models were more robust, specifically, a more cohesive variable was derived through the combination of both independent race and ethnicity variables. Further, dual language status did not simply include the participant's native language, but rather was the result of an English oral language proficiency screener.

The results of this study indicate all items do not load to the same subscales as described by the DECA, namely, only two items within the behavior concerns domain loaded to that factor. Second, the CFA models did not fit the data well; therefore, to begin the measurement invariance analyses, a poorer fitting model was utilized. Third, of the four invariance models conducted, no model met configural invariance.

The results of this study share parallels with the results described by Bulotstky-Shearer et al. (2013), Lien and Carlson (2009), and Ogg et al. (2010). All three of these studies also included samples of children enrolled in Head Start; however, these studies all studied the first edition of this instrument. Further, Bulotstky-Shearer et al. (2013) and Ogg et al. (2010) studied the teacher version, whereas Lien and Carlson (2009) studied the parent version.

All three studies also assessed the factor structure via confirmatory factor analysis. Further, all three studies found marginal to adequate fit within the total protective factors domain and excluded the behavior concerns domain from the total protective factors model. Bulotstky-Shearer et al. (2013) and Lien and Carlson (2009) both identified item misfit within the total protective factors domain where items load to different factors than those aligned to the theoretical model. In addition, Bulotstky-Shearer et al. (2013) also conducted a separate confirmatory factor analysis with the behavior concerns domain. Based on the results of their models, they suggest an alternate factor structure for children of Head Start and a two-factor structure for the behavior concerns domain, which distinguished between different types of behavior, including internalizing and externalizing behavior.

Finally, once Ogg et al. (2010) released error covariances to 10 items, baseline fit improved, and ultimately, the researchers found the theoretical model that was devised by the authors of the DECA to align, and the items functioned similarly between boys and girls.

Practitioner Guide

A practitioner guide that offers practitioners a quick, step-by-step tutorial for reviewing and interpreting technical manuals supports practitioner's ability to independently review and learn about an instrument, as well as the ability to make informed decisions related to relevance and appropriateness of use within their participant group. As articulated in this study, understanding, recognizing, and addressing informant biases via perception and memory are important considerations, namely to developers, but also to practitioners. During the training and implementation phases of data collection, practitioners must be aware of the various skills and biases related to the informant and assessment format that may implicate the rating, scoring, and use of the results. Further, as described, instruments that are collected indirectly may have varied considerations during the review process, and without specific knowledge of considerations, practitioners may not gain the information needed to make the most informed decision. To that end, this study proposes a guide that includes tips and insights and is detailed to offer a deeper understanding of the various facets related to instrument development, yet is short enough that practitioners can review and independently complete it.

A five domain, 8-question guide is shared below in Figure 3. The five domains include 1. Defining your purpose and goals, 2. Defining your sample and instrument standardization sample, 3. Assessing psychometric properties, 4. Instrument training, administration, and interpretation, and 5. Other insights and insider tips. Beyond specific questions, insights, and tips related to review of technical manuals, this guide also offers introductory questions related to the purpose and goals for which a practitioner may use an instrument, so as to ensure alignment between instrument purposes and practitioner goals.

Figure 3
Practitioner Guide for Selecting Social Science Instruments

Practitioner Guide for Selecting Social Science Instruments

Purpose of guide: This guide is intended to support practitioners in making informed decisions regarding social science instrument use via support and guidance related to interpreting and understanding technical manuals.

This guide offers tips, insights, and questions practitioners should answer to best support decision making prior to use of an instrument with their participant group.

Defining Your Purpose and Goals

1. Prior to researching, reviewing, and selecting an instrument for use, define your purpose. Do you plan to use the instrument for an evaluative purpose, instructional purpose, or screener purpose? The validity of an instrument depends upon alignment between the intended purposes and your defined purpose.
    - Insider tip: Instruments each have a purpose. Ensure the instrument details its intended purpose and how the outcomes derived from conducting the instrument are intended to be used.
    - Insight: Some instruments are summative and offer a single score, while other instruments are formative and are used for the purpose of collecting multiple rounds of data to inform instruction. When deriving your purpose, review the breadth of scores that are available for analytical and other purposes.

2. Define your goal(s). Do you plan to use the instrument one time, multiple times for comparison, year after year? How will the results of the instrument be used?
    - Insider tip: Devising a logic model with short-term (changes in learning), mid-term (changes in behavior), and long-term (changes in condition) goals may help identify an appropriate instrument that will support your goals.

Defining Your Sample & Instrument Standardization Sample

1. Define relevant participant characteristics, such as age, race, ethnicity, language, gender, income status, of your projected participant group.

2. Using the above participant characteristics, review the instrument's technical manual. Does the population for which the instrument was standardized align with the characteristics defined in your sample? If there is misalignment, this could lead to misinterpretation of scores.
    - Insider tip: If the two groups do not align, the instrument developer may have more insights into other studies for which the participant sample aligns. Ask the instrument developer for the study information.

3. Does the technical manual describe the process by which the developers standardized (i.e., normed) the instrument? Review the process in detail, as misalignment within your participant group may affect the reliability of your scores.

4. Does the technical manual describe the process by which the developers selected items and further organized the items into scales (if multiple constructs are included in the instrument)? These descriptions will support your interpretation of scores.
    - Insight: A construct is a concept that has been operationally defined.

Assessing Psychometric Properties

5. Does the technical manual explain the research approach to collecting and assessing reliability? What types of reliability were assessed and what results were shared?
    - Insight: Reliability seeks to quantify the consistency of the measure and its items. There are different types of reliability, including internal consistency (measures stability across items), test-retest reliability (measures stability over time), interrater reliability (measures stability across raters), and parallel forms (measures stability across forms).
    - Insider tip: Does the manual state that adequate reliability (alpha coefficients of ≥.70) was gained during the testing phase?

6. Does the technical manual explain the research approach to collecting and assessing validity? What types of validity were assessed and what results were shared?

- Insight: Validity seeks to understand how well the instrument measures the underlying construct(s) it intends to measure. There are different types of validity, including content-related (measures the degree to which an instrument adequately represents a performance domains or construct), criterion-related (measures the degree to which the results of an instrument relate or predict a variable), and construct-related (measures the degree to which an instrument measures the construct it is supposed to measure).
- Insider tip: Does the manual state the literature review process for which the constructs were devised? Were other instruments of similar construct(s) included in the standardization sample? Did the instrument correlate well with these instrument(s)?

Instrument Training, Administration, and Interpretation

7. Within your participant group, who will be administering the instrument? Do they need training? DO they have the necessary expertise/relationship with the participants? Have they been adequately informed about the process and purpose of administration?

8. Does the instrument developer/vendor include (and/or require) training prior to administering the instrument?
   - Insider tip: Check for any requirements related to interrater reliability and costs associated with any training. Additionally, some instruments require in-person training with a trainer. Include travel costs associated with in-person training.

9. Does the technical manual clearly articulate appropriate interpretation of scores?
   - Insider tip: Communicating scores to various audiences can be challenging. Their level of expertise and knowledge may vary; therefore, review the technical manual or connect with the instrument developer/vendor to ensure your interpretation of the scores aligns with the instrument.

Other Insights and Insider Tips

- During the review process of an instrument, schedule a call with the vendor to discuss the instrument’s purpose, intended use, costs, and customer service availability in detail. Further, some instruments include access to an online platform.
- If an online platform is available: reports may be available for download. Ensure CSV files are available for exporting, if you seek to analyze the data yourself or in collaboration with an analytical expert. Further, discuss batch uploading options for participant information with the instrument developer/vendor.
- Depending on the purpose of using the instrument, you may want to collect feedback from stakeholders after administration and review of

	the results. This feedback may inform modifications and/or improvements you may make to your purpose in future time.

Limitations

During the exploratory factor analysis, four strategies were deployed to determine factor retention. Three of those strategies identified a four-factor model, whereas the parallel analysis identified a three-factor model. The four-factor model was initially selected, as this aligned to the model described by the DECA that included the subscales of attachment/relationships, self-regulation, initiative, and behavior concerns. However, with the four-factor model, the majority of items within the behavior concerns subscale loaded to other factors and not to a factor that could be characterized as behavior concerns. Only two of the 11 items loaded to a factor characterized as behavior concerns; therefore, this poses a limitation to this study. Due to the three-factor model identified through the parallel analysis, additional models were tested and reviewed to better understand the implications posed by the items of behavior concerns loading to other factors. Two additional models were reviewed, including a three-factor model that included the items from behavior concerns and a three-factor model that excluded the items from behavior concerns.

The three-factor model that included items from behavior concerns had a similar fit to the four-factor model, namely due to the fact that 9 of the 11 items had already loaded to one of the three factors; therefore, this model sought to understand to which factors the remaining two items of behavior concerns loaded. The three-factor model that excluded items from behavior concerns was a much better fitting model. Further, when that model was utilized similarly, including use of fall data, within the measurement invariance analysis by gender, configural and metric invariance was met and partial invariance was met within strong and strict invariance. A table describing this analysis is provided in Table 11.

Table 11

Invariance Analysis by Gender for the DECA, excluding behavior concerns subscale

	χ^2	*df*	$\Delta\chi^2$	Δ*df*	SRMR	CFI	RMSEA (90% CI)
Configural	2818.2	592	331.8	296	0.081	0.833	0.094 [.091, .098]
Weak	2848.5	617	2848.5	30	0.084	0.832	0.092 [.089, .096]
Strong	2899.3	639	50.735***	22	0.084	0.830	0.091 [.088, .095]
Strong, release item 14	2886.0	638	37.428*	21	0.084	0.831	0.091 [.088, .094]
Strong, release item 17	2874.5	637	25.9	20	0.084	0.832	0.091 [.088, .094]
Strict	2973.3	663	98.811***	26	0.085	0.827	0.091 [.087, .094]
Strict, release item 15	2958.3	662	83.832***	25	0.085	0.828	0.090 [.087, .094]
Strict, release item 37	2947.4	661	72.934***	24	0.085	0.828	0.090 [.087, .094]
Strict, release item 2	2937.9	660	63.476***	23	0.085	0.829	0.090 [.087, .093]
Strict, release item 24	2931.0	659	56.52***	22	0.085	0.829	0.090 [.087, .093]
Strict, release item 25	2923.5	658	49.01***	21	0.085	0.830	0.090 [.087, .093]
Strict, release item 31	2916.1	657	41.609**	20	0.085	0.830	0.090 [.087, .093]
Strict, release item 32	2909.8	656	35.34*	19	0.085	0.831	0.090 [.087, .093]
Strict, release item 11	2904.5	655	29.992*	18	0.084	0.831	0.090 [.087, .093]
Strict, release item 20	2898.9	654	24.5	17	0.084	0.831	0.090 [.087, .093]

Note. DECA = Devereux Early Childhood Assessment; SRMR = standard root mean square residual; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval.

***p < .001, ** < .01, *< .05

This additional exploration of best fitting models signifies the potential issues items from the behavior concerns subscale are causing within both the factor analysis models and measurement invariance models. A review of the technical manual of the DECA, second edition, outlined the process by which the instrument developers tested the items. Namely, factor analysis of items within the total protective factors domain were analyzed, and the instrument developers compared the factor analysis results between the parent raters and teacher raters. The factor analysis concluded a three-factor model which included the subscales of attachment/ relationships, self-regulation, and initiative. Further, the technical manual described a separate process for items selected within the behavior concerns subscale. Namely, those items were not included in the factor analysis models, and reliability and correlations of item totals were analyzed to determine the items retained in that subscale (LeBuffe & Naglieri, 2012).

With this information in mind, the lack of inclusion of items within the behavior concerns domain within instrument development poses a limitation. Within this study, the items within that domain load to subscales within the total protective factors domain and appear to prohibit a better fitting model. Further analysis with the items within the behavior concerns subscale should be conducted to ensure appropriate interpretability and use of the results from the behavior concerns subscale. Without those analyses, interpreting the results from the behavior concerns subscale should be limited.

A second limitation of this study is exclusion of participants younger than age three. As shared, this instrument has infant and toddler forms available. The infant form includes 33 items related to the attachment/relationships and initiatives subscales, and is appropriate for children ages one month to 18 months. The toddler form includes 36 items related to all subscales of the total protective factors domain, and is appropriate for children ages 18 months to 36 months. Exploration of the factor structure and measurement invariance should be conducted to ensure appropriateness within those age groups, namely as observing and rating children's skill sets may be more challenging the younger the child due to lack of communication skills and the variation in developmental progressions within young children. An informant who rates these skills might need a keener eye and more training to ensure the observational and rating processes align with the instrument goals.

Further, in support of diversity, equity, and inclusion efforts, and as related to the second limitation, an example related to informant perception includes both an understanding of children's language acquisition style for dual language learners, as well as the ability to differentiate between internalizing and externalizing behaviors. As Espinosa (2015) describes, children learning a second language either engage in a simultaneous or sequential language acquisition style. A simultaneous style is one in which the child travels through fundamental, language development milestones as their monolingual peers. Whereas the sequential style occurs when a child follows a varied progression. Within the sequential style, dual language learners engage in a four-stage process for second language development. Those stages include home language use, nonverbal/observational period, engagement in telegraphic and formulaic speech, and finally productive language (Espinosa, 2015, 53). Specifically, within the second stage, the young child enters a period in which he becomes nonverbal, and more observational. Further, during this time, the child may not communicate with other peers or teachers as he engages in observation of his surroundings. A teacher may perceive the situation in a variety of means, perhaps perceiving the lack of communication as withdrawal or anxiety, forms of internalizing behavior, or as a lack of skill set the child possesses. These various perceptions may implicate the teacher's understanding of the true skill sets the child possesses; therefore, potentially implicating judgments being rendered within an assessment of that child. For this purpose, education leaders and teachers should be cognizant of initiatives related to diversity, equity, and inclusion; seeking to ensure robust understanding of language acquisition styles for children learning a second language. Lack of knowledge may implicate a teacher's perception of a child's skill sets; therefore, implicating the true score derived from an assessment and the use of those results.

Future Opportunities

As described in the limitations, further exploration of the items embedded within the behavior concerns domain is warranted; therefore, future research should include a deeper review and analysis of the items within this domain. Specifically, a one-factor model of behavior concerns may be insufficient, as also shared by Bulotstky-Shearer et al. (2013). Researchers should review the items currently embedded within this domain, as well as research related to externalizing and internalizing behaviors. Perhaps a two-factor model that separates externalizing and internalizing behaviors may be warranted. Further, if this is the case, additional training and guidance may be necessary, as internalizing behaviors may be more challenging to observe and may be perceived within the subscales of total protective factors, such as initiatives and attachment/relationships.

A second opportunity for future research includes deeper exploration of the role teachers' beliefs and expectations play within indirect assessment. Teachers facilitate the learning process of their students, and the beliefs and expectations underlying those learning experiences implicates the experience students receive in the classroom and the outcomes they achieve. More specifically, as argued in this study, future research should seek to incorporate instruments related to beliefs and expectations, including assessment of cultural and other student characteristic biases in expectations, and the association to various metrics of student achievement, specifically those instruments conducted indirectly. Indirect assessment opportunities are valuable and add insight into specific study goals; however, disentangling active and passive participants and informant qualities is important for collecting reliable and valid data. For the purposes of this instrument, the outcomes of children's social-emotional skill sets were analyzed and used for individualization opportunities; however, use of the results may be compromised if teacher's biases are unaccounted.

Finally, a third opportunity for future research includes conducting advanced methods, such as multilevel modeling, within indirect assessment. More specifically, assessing various characteristics of the informant, including qualifications, should be considered and explored, as these analyses can offer a deeper understanding of the variance accounted for within each level of the data.

Conclusions

Practitioners, like early childhood educators, seek robust instruments for varying purposes, such as evaluation, intervention, and creating individualized opportunities within instruction. Early childhood teachers play a direct role in facilitating the learning experience for their students, and as a strengths-based instrument, the DECA seeks to provide practitioners valuable information to form meaningful experiences. Further, being mindful of the demographic characteristics of children enrolled in Head Start, and the holistic needs of these students and their families, early intervention that seeks to address social-emotional learning skill sets is vital to promoting long-term success (Lee, 2008; McCoy et al., 2017).

To provide meaningful experiences that promote student achievement, researchers should explore teacher philosophical and pedagogical beliefs, as variation across these foundational beliefs are implicative of the experience children are offered within the classroom. To promote equitable, inclusive, and culturally responsive experiences, selecting the most appropriate instrument that clearly articulates student strengths and areas of need is important. School leaders can also support the triangulation of data collection, data review, and strategy implementation to ensure teacher beliefs and skill sets positively support the learning experiences for each student.

Based on the results of this study, the argument that conducting reliability and validity analyses during instrument development is insufficient; therefore, instrument developers should conduct more advanced analyses, like measurement invariance and differential item functioning, to ensure robustness and appropriateness within the population(s) for which the instrument has been devised. Moreover, it is vitally important these advanced analyses are conducted within instruments of indirect assessment due to the greater risk of potential error that may be imparted into informant's judgement via perception and memory.

REFERENCES

Administration for Children & Families. (2020). *Head start services*. https://www.acf.hhs.gov/ohs/about/head-start

Administration for Children & Families. (2021). *Head start history*. https://www.acf.hhs.gov/ohs/about/history-head-start

Administration for Children & Families. (ND). *Head start family and child experiences survey (FACES), 1997-2022*. https://www.acf.hhs.gov/opre/project/head-start-family-and-child-experiences-survey-faces-1997-2022

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

Bagnato, S. J. (2005). The authentic alternative for assessment in early intervention: An emerging evidence-based practice. *Journal of Early Intervention, 28*(1), 17-22. https://doi.org/10.1177/105381510502800102

Baker, C. N., Tichovolsky, M. H., Kupermidt, J. B., Voegler-Lee, M. E., & Arnold, D. H. (2015). Teacher (mis)perception of preschoolers' academic skills: Predictors and associations with longitudinal outcomes. *Journal of Educational Psychology, 107*(3), 805-820. https://doi.org/10.1037/edu0000008

Ballad, W. S., & Tighe, P. L. (2010). *Pre-idea proficiency test – oral.* Ballad & Tighe, Educational IDEAS.

Barbu, O. C., Levine-Donnerstein, D., Marx, R. W., & Yaden Jr., D. B. (2012). Reliability and validity of the devereux early childhood assessment (DECA) as a function of parent and teacher ratings. *Journal of Psychoeducational Assessment, 31*(5), 469-481. https://doi.org/10.1177/0734282912467758

Berg-Nielson, T. S., Solheim, E., Belsky, J., & Wichstrom, L. (2012). Preschoolers' psychosocial problems: In the eyes of the beholder? Adding teacher characteristics as determinants of discrepant parent-teacher reports. *Child Psychiatry & Human Development*, *43*(3), 393-413. https://doi.org/10.1007/s10578-001-0271-0

Brooks-Gunn, J., & Duncan, G.T. (1997). The effects of poverty on children. *Poverty and Children*, 7(2), 55-71. https://doi.org/10.2307/1602387

Bulotsky-Shearer, R. J., Fernandez, V. A., & Rainelli, S. (2013). The validity of the devereux early childhood assessment for culturally and linguistically diverse head start children. *Early Childhood Research Quarterly, 28*(4), 794-807. https://doi.org/10.1016/j.ecresq.2013.07.009

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science, 12*(3), 140-153. https://doi.org/10.1080/10888690802199418

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25*(2), 166-176. https://doi.org/10.1016/j.ecresq.2009.10.004

Cadima, J. Verschueren, K., Leal, T., & Guedes, C. (2016). Classroom interactions, dyadic teacher-child relationships, and self-regulation in social disadvantages young children. *Journal of Abnormal Child Psychology, 44*(1), 7-17. https://doi.org/10.1007/s10802-015-0060-5

Carlson, J. S., & Voris, D. S. T. (2018). One-year stability of the devereux early childhood assessment for preschoolers, second edition. *Journal of Psychoeducational Assessment, 36*(8), 829-834. https://doi.org/10.1177/0734282917710890

Contractor, A. A., Bolton, E., Gallagher, M. W., Rhodes, C., Nash, W. P., & Litz, B. (2017). Longitudinal measurement invariance of posttraumatic stress disorder in deployed marines. *Journal of Traumatic Stress, 30*(3), 259-269. https://doi.org/10.1002/jts.22181

Crane, J., Mincic, M. S., & Winsler, A. (2011). Parent-teacher agreement and reliability on the devereux early childhood assessment (DECA) in English and Spanish for ethnically diverse children living in poverty. *Early Education & Development, 22*(3), 520-547. https://doi.org/10.1080/10409289.2011.565722

De Kruif, R. E. L., McWilliam, R. A., Ridley, S. M., & Wakely, M. B. (2000). Classification of teachers' interaction behaviors in early childhood classrooms, *Early Childhood Research Quarterly, 15*(2), 247-268. https://doi.org/10.1016/S0885-2006(00)00051-X

DePaul Teaching Commons. (2021). Direct versus indirect assessment of student learning. https://resources.depaul.edu/teaching-commons/teaching-guides/feedback-grading/Pages/direct-assessment.aspx

Devereux early childhood assessment for preschoolers. (ND). Center for Resilient Children. www.CenterForResilientChildren.org

Di Santo, A., Timmons, K., & Lenis, A. (2017). Preservice early childhood educators' pedagogical beliefs. *Journal of Early Childhood Teacher Education. 38*(3), 223-241. https://doi.org/10.1080/10901027.2017.1347588

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33*(5), 510-540. https://doi.org/10.1006/jesp.1997.1331

Dracobly, J. D., Dozier, C. L., Briggs, A. M., & Juanico, J. F. (2018). Reliability and validity of indirect assessment outcomes: Experts versus caregivers. *Learning and Motivation 62*, 77-90. https://doi.org/10.1016/j.lmot.2017.02.007

Duameyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology, 84,* 103812-103821. https://doi.org/10.1016/j.jesp.2019.04.010

Duncan, G.J., Brooks-Gunn, J., & Klebanov, P.K. (1995). Economic deprivation and early childhood development. *Child Development*, *65*(2), 296-318. https://doi.org/10.1111/j.1467-8624.1994.tb00752.x

Esnaola, I., Benito, M., Antonio-Agirre, I., Axpe, I., & Lorenzo, M. (2019). Longitudinal measurement invariance of the satisfaction with life scale in adolescence. *Quality of Life Research, 28*(10), 2831-2837. https://doi.org/10.1007/s11136-019-02224-7

Espinosa, L. M. (2015). *Getting it right for young children from diverse backgrounds: Applying research to improve practice with a focus on dual language learners*. Pearson Education.

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, & T. Uradan (Eds.), *APA Educational Psychology Handbook: Vol. 2. Individual Differences and Cultural and Contextual Factors* (pp. 471-499). American Psychological Association. https://doi.org/10.1037/13274-019

Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2015). Prekindergarten children's executive functioning skills and achievement gains: The utility of direct assessment and teacher ratings. *Journal of Educational Psychology, 107*(1), 207-221. https://doi.org/10.1037/a0037366

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., & Hamilton Krieger, L. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945-967. https://doi.org/10.2307/20439056

Guo, B., Kaylor-Hughes, C., Garland, A., Nixon, N., Sweeney, T., Simpson, S., Dalgleish, T., Ramana, R., Yang, M., & Morriss, R. (2017). Factor structure and longitudinal measurement invariance of PHQ-9 for specialist mental health care patients with persistent major depressive disorder: Exploratory structural equation modelling. *Journal of Affective Disorders, 219*, 1-8. https://doi.org/10.1016/j/jad/2017.05.020

Head Start (2021). *Head start program facts: Fiscal year 2019*. https://eclkc.ohs.acf.hhs.gov/about-us/article/head-start-program-facts-fiscal-year-2019

IBM Corp. (2016). IBM SPSS Statistics for Windows (Version 24.0). IBM Corp. https://www-01.ibm.com/support/docview.wss?uid=swg21476197

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling. R package version 0.5-4. Retrieved from https://CRAN.R-project.org/package=semTools

Khine, M. S. (2013). *Application of structural equation modeling in educational research and practice*. Sense Publishers.

Khojasteh, J., & Lo, W-J. (2015). Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in a bifactor model. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 531-541. https://doi.org/10.1080/10705511.2014.937791

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment, 30*(2), 148 –159. https://doi.org/10.1177/0734282911412722

Kuklinksi, M. R., & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development*, *72*(5), 1554-1578. https://doi.org/10.1111/1467-8624.00365

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research, 70*(4), 443-484. https://doi.org/10.3102/00346543070004443

La Paro, K. M., Siepak, K., & Scott-Little, C. (2009). Assessing beliefs of preservice early childhood education teachers using q-sort methodology. *Journal of Early*

Childhood Teacher Education, 30(1), 22-36. https://doi.org/10.1080/10901020802667805

Leary, M. R. (2008). *Introduction to behavioral research methods*. Pearson Education.

LeBuffe, P. A., & Naglieri, J. A. (1999). *Devereux Early Childhood Assessment: User's guide.* Kaplan Early Learning Company.

LeBuffe, P. A., & Naglieri, J. A. (2012). *Devereux Early Childhood Assessments for Preschool, second edition* (Technical manual). Kaplan Early Learning Company.

Lee, K. (2008). The effects of children's head start enrollment age on their short- and long-term developmental outcomes. *The Social Science Review, 82*(4), 663-702. https://doi.org/10.1086/597018

Liang, S H., Chou, J., Wu, Y., Lee, C., Kelsen, B. A., & Lee, Y. (2019). Validity and reliability study of the Chinese (traditional) version of the devereux early childhood assessment for toddlers (DECA-T). *Neuropsychiatric Disease and Treatment, 15*, 3375-3385. https://doi.org/10.2147/NDT.S218943

Lien, M. T., & Carlson, J. S. (2009). Psychometric properties of the devereux early childhood assessment in a head start sample. *Journal of Psychoeducational Assessment, 27*(5), 386-396. https://doi.org/10.1177/0734282909331754

Loftus, E. (2003). Our changeable memories: legal and practical implication. *Neuroscience, 4*(3), 231-234. https://doi.org/10.1038/nrn1054

Madigan, S., Atkinson, L., Laurin, K., & Benoit, D. (2013). Attachment and internalizing behavior in early childhood: A meta-analysis. *Developmental Psychology, 49*(4), 672-689. https://doi.org/10.1037/a0028793

McCoy, D. C., Yoshikawa, H., Zoil-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., Yang, R., Koeep, A., & Shonkoff, J. P. (2017). Impacts of early childhood education on medium- and long-term educational outcomes. *Educational Researcher, 46*(8), 474-487. https://doi.org/10.3102/0013189X17737739

Meisels, S. J. (1996). Performance in context: Assessing children's achievement at the outset of school. In A. J. Sameroff & M. M. Haith (Eds.), *The five to seven year shift: The age of reason and responsibility* (pp. 410-431). University of Chicago Press.

Meisels, S. J. (1999). Assessing readiness. In R. C. Pianta & M. Cox (Eds.), *The transition to kindergarten: Research, policy, training, and practice*. Paul H. Brookes.

Meisels, S. J., Wen, X., & Beachy-Quick, K. (2010). Authentic assessment for infants and -toddlers: Exploring the reliability and validity of the ounce scale. *Applied Developmental Science, 14*(2), 55-71. https://doi.org/10.1080/10888691003697911

Millsap, R. E., (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473. https://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*(1), 5-9. https://doi.org/10.1111/j.1750-8606.2009.00109.x

Moreira, P. S., Santos, N., Castanho, T., Amorim, L., Portugal-Nunes, C., Sousa, N., & Costa, P. (2018). Longitudinal measurement invariance of memory performance and executive functioning in healthy aging. *PLoS ONE, 13*(9), e0204012-e0204012. https://doi.org/10.1371/journal.pone.0204012

Ogg, J. A., Brinkman, T. M., Dedrick, R. F., & Carlson, J. S. (2010). Factor structure and invariance across gender of the devereux early childhood assessment protective factor scale. *School Psychology Quarterly, 25*(2), 107-118. https://doi.org/10.1037/a0020251

Paclawskyj, T. R., Matson, J. L., Rush, K. S., Smalls, Y., & Vollmer, T. R. (2008). Assessment of the convergent validity of the questions about behavioral function scale with analogue functional analysis and the motivation assessment scale. *Journal of Intellectual Disability Research, 45*(6), 484-494. https://doi.org/10.1046/j.1365-2788.2001.00364.x

Patrick, H., Mantzicopoulos, P., & French, B. F. (2020). The predictive validity of classroom observations: Do teachers' framework for teaching scores predict kindergarteners' achievement and motivation?, *American Educational Research Journal, 57*(5), 2021-2058. https://doi.org/10.3102/002831219891409

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS) manual, pre-K*. Brookes.

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis, 42*(3), 375-392. https://doi.org/10.3102/0162373720932188

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

R Core Team. (2017). R: A language and environment for statistical computing [R Foundation for Statistical Computing]. Vienna, Austria. http://www.R-project.org/

RStudio Team. (2020). RStudio: Integrated Development for R. RStduio, PBC. http://www.rstudio.com

Rosseel, Y. (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, 48(2), 1–36. https://www.jstatsoft.org/v48/i02/.

Saracho, O. N. (2017). Writing and publishing qualitative studies in early childhood education. *Early Childhood Education Journal*, *45*(1), 15-26. https://doi.org/10.1007/s10643-016-0794-x

Shoshani, A. (2019). *Character Strengths Inventory for Early Childhood* [Database record]. PsycTESTS. https://doi.org/10.1037/t76710-000

Suh, H., Rice, K. G., Choi, C-C., van Nuenen, M., Zhang, Y., Morero, Y., & Anderson, A. (2016). Measuring acculturative stress with the SAFE: Evidence for longitudinal measurement invariance and associations with life satisfaction.

Personality and Individual Differences, 89, 217-222. https://doi.org/10.1016/j.paid.2015.10.002

Sulik, M. J., Blair, C., Mills-Koonce, R., Berry, D., & Greenberg, M. (2015). Early parenting and the development of externalizing behavior problems: Longitudinal mediation through children's executive function. *Child Development, 86*(5), 1588-1603. https://doi.org/10.1111/cdev.12386

Suk Lee, Y., Baik, J., & Charlesworth, R. (2006). Differential effects of kindergarten teacher's beliefs about developmentally appropriate practice on their use of scaffolding following inservice training. *Teaching and Teacher Education, 22(*7), 935-945. https://doi.org/10.1016/j.tate.2006.04.041

Tate, S., & Page, D. (2018). Whiteliness and institutional racism: Hiding behind (un)conscious bias. *Ethics and Education, 13*(1), 141-155. https://doi.org/10.1080/17449642.2018.1428718

Teaching Strategies. (2015). *GOLD objectives for development and learning, birth through third grade*. Teaching Strategies, LLC.

Tramontana, M. G., Hooper, S. R., & Selzer, S. C. (1988). Research on the preschool prediction of later academic achievement: A review. *Developmental Review, 8*(2), 89-146. https://doi.org/10.1016/0273-2297(88)90001-9

U.S. Department of Health and Human Services. (2021). *U.S. federal poverty guidelines used to determine financial eligibility for certain federal programs*. https://aspe.hhs.gov/poverty-guidelines

Vandenburg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-69. https://doi.org/10.1177/109442810031022

Vartuli, S., & Rohs, J. (2009). Early childhood prospective teacher pedagogical belief shifts over time. *Journal of Early Childhood Teacher Education, 30*(4), 310-327. https://doi.org/10.1080/10901020903320262

Zeelenberg, R. Wagenmakers, E-J., & Rotteveel, M. (2006). The impact of emotion on perception. *Psychological Science*, *17*(4), 289-291. https://doi.org/10.1111/j.1467-9280.2006.01700.x

APPENDICES

Figure 1

Devereux early childhood assessment factor structure aligned to user manual

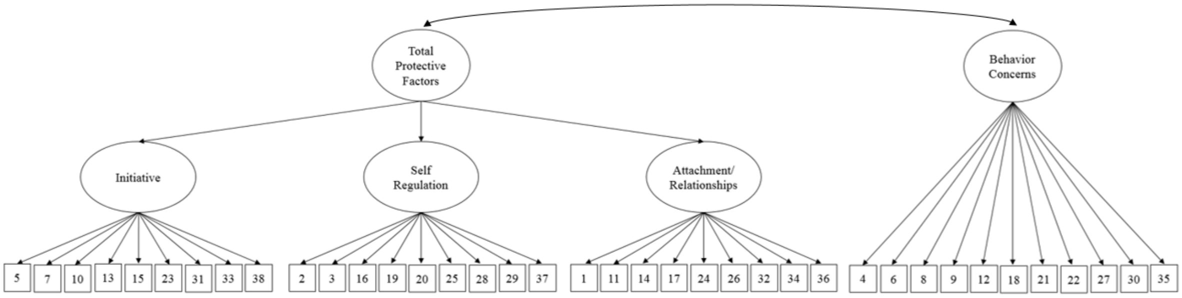

Figure 2

Devereux early childhood assessment factor structure alignment to study

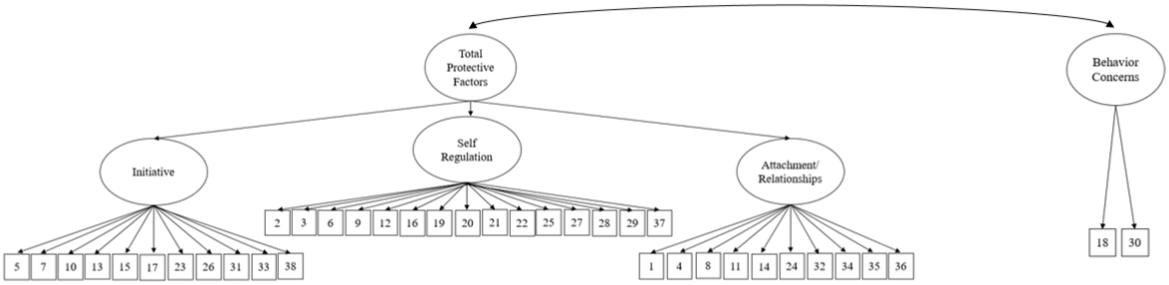

Table 4

Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	.476**	1											
3	.472**	.653**	1										
4	.339**	.286**	.336**	1									
5	.432**	.366**	.350**	.339**	1								
6	.244**	.595**	.658**	.311**	.230**	1							
7	.368**	.486**	.432**	.281**	.564**	.387**	1						
8	.373**	.287**	.305**	.476**	.340**	.218**	.276**	1					
9	.206**	.448**	.415**	.178**	.150**	.445**	.224**	.098**	1				
10	.308**	.434**	.381**	.235**	.521**	.368**	.663**	.239**	.216**	1			
11	.390**	.283**	.280**	.202**	.217**	.169**	.227**	.188**	.268**	.193**	1		
12	.391**	.551**	.583**	.271**	.263**	.559**	.311**	.318**	.521**	.256**	.290**	1	
13	.360**	.343**	.219**	.185**	.497**	.202**	.538**	.279**	.075*	.558**	.234**	.198**	1
14	.513**	.323**	.300**	.287**	.304**	.203**	.308**	.275**	.184**	.290**	.413**	.230**	.358**
15	.461**	.407**	.413**	.333**	.513**	.321**	.465**	.460**	.129**	.492**	.224**	.322**	.515**
16	.330**	.692**	.590**	.260**	.324**	.639**	.521**	.250**	.353**	.495**	.201**	.468**	.350**
17	.408**	.385**	.221**	.192**	.383**	.209**	.395**	.246**	.145**	.381**	.208**	.211**	.499**
18	.221**	.538**	.331**	.225**	.304**	.460**	.418**	.189**	.313**	.407**	.146**	.387**	.340**
19	.467**	.644**	.635**	.320**	.375**	.519**	.451**	.386**	.313**	.445**	.228**	.557**	.333**
20	.324**	.623**	.675**	.299**	.303**	.716**	.493**	.244**	.360**	.507**	.220**	.486**	.320**
21	.323**	.596**	.618**	.243**	.177**	.555**	.285**	.225**	.431**	.234**	.198**	.555**	.108**
22	.320**	.475**	.558**	.395**	.310**	.602**	.397**	.278**	.252**	.373**	.137**	.436**	.188**
23	.417**	.473**	.349**	.220**	.552**	.344**	.596**	.306**	.175**	.604**	.248**	.264**	.671**
24	.587**	.501**	.522**	.338**	.465**	.336**	.438**	.411**	.248**	.390**	.390**	.490**	.370**
25	.364**	.634**	.617**	.304**	.365**	.651**	.526**	.299**	.357**	.484**	.200**	.511**	.356**
26	.432**	.424**	.340**	.248**	.315**	.328**	.413**	.293**	.233**	.377**	.282**	.346**	.380**
27	.259**	.569**	.536**	.179**	.117**	.540**	.259**	.138**	.526**	.212**	.214**	.553**	.087*
28	.376**	.666**	.578**	.253**	.311**	.547**	.466**	.289**	.372**	.456**	.254**	.478**	.368**
29	.280**	.506**	.529**	.275**	.306**	.591**	.451**	.194**	.291**	.451**	.201**	.383**	.274**
30	.272**	.503**	.349**	.232**	.297**	.398**	.347**	.221**	.288**	.353**	.109**	.397**	.301**
31	.331**	.383**	.287**	.190**	.425**	.270**	.487**	.238**	.090**	.509**	.199**	.199**	.490**
32	.564**	.381**	.497**	.418**	.449**	.263**	.343**	.468**	.158**	.308**	.324**	.456**	.292**
33	.355**	.373**	.318**	.211**	.514**	.282**	.591**	.228**	.149**	.631**	.155**	.246**	.535**
34	.384**	.351**	.273**	.298**	.400**	.233**	.418**	.295**	.149**	.414**	.373**	.216**	.493**
35	.340**	.300**	.337**	.198**	.191**	.191**	.168**	.230**	.329**	.112**	.217**	.460**	0.057
36	.212**	0.041	0.018	.105**	.100**	-.134**	0.010	.107**	-.071*	-0.001	.113**	0.031	.101**
37	.564**	.622**	.664**	.363**	.389**	.460**	.418**	.466**	.327**	.364**	.326**	.586**	.298**
38	.483**	.424**	.391**	.255**	.513**	.238**	.501**	.305**	.167**	.499**	.249**	.352**	.472**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4

Correlations

	14	15	16	17	18	19	20	21	22	23	24	25	26
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14	1												
15	.354**	1											
16	.273**	.435**	1										
17	.487**	.475**	.368**	1									
18	.171**	.369**	.535**	.303**	1								
19	.363**	.523**	.657**	.353**	.424**	1							
20	.269**	.411**	.726**	.262**	.476**	.647**	1						
21	.218**	.245**	.532**	.197**	.368**	.530**	.528**	1					
22	.167**	.347**	.495**	.140**	.361**	.485**	.617**	.461**	1				
23	.398**	.534**	.478**	.530**	.415**	.477**	.446**	.258**	.288**	1			
24	.527**	.458**	.387**	.361**	.256**	.555**	.413**	.385**	.372**	.491**	1		
25	.293**	.445**	.682**	.329**	.466**	.621**	.713**	.498**	.538**	.498**	.510**	1	
26	.439**	.426**	.415**	.458**	.314**	.459**	.419**	.269**	.311**	.472**	.515**	.497**	1
27	.194**	.139**	.509**	.140**	.386**	.470**	.501**	.706**	.390**	.216**	.295**	.467**	.223**
28	.327**	.466**	.658**	.365**	.485**	.662**	.656**	.522**	.442**	.494**	.458**	.664**	.461**
29	.246**	.370**	.589**	.255**	.384**	.502**	.669**	.387**	.530**	.395**	.358**	.602**	.391**
30	.132**	.365**	.472**	.281**	.782**	.388**	.421**	.394**	.342**	.379**	.282**	.415**	.266**
31	.246**	.478**	.373**	.390**	.399**	.383**	.384**	.166**	.290**	.533**	.360**	.414**	.422**
32	.456**	.486**	.330**	.301**	.190**	.521**	.348**	.348**	.362**	.393**	.636**	.389**	.399**
33	.297**	.479**	.410**	.408**	.380**	.421**	.424**	.216**	.336**	.625**	.401**	.441**	.394**
34	.428**	.454**	.338**	.508**	.295**	.351**	.330**	.138**	.171**	.548**	.415**	.331**	.407**
35	.162**	.164**	.187**	.081*	.157**	.337**	.181**	.379**	.235**	.093**	.337**	.210**	.143**
36	.296**	.092**	-0.053	.208**	0.011	0.061	-.078*	0.027	-0.028	.079*	.241**	-0.020	.165**
37	.380**	.536**	.551**	.326**	.321**	.709**	.545**	.552**	.477**	.412**	.631**	.560**	.417**
38	.366**	.489**	.392**	.430**	.431**	.451**	.345**	.261**	.303**	.568**	.527**	.390**	.398**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4

Correlations

	27	28	29	30	31	32	33	34	35	36	37	38
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27	1											
28	.522**	1										
29	.407**	.572**	1									
30	.377**	.397**	.338**	1								
31	.119**	.469**	.420**	.326**	1							
32	.255**	.426**	.334**	.222**	.324**	1						
33	.160**	.416**	.366**	.352**	.499**	.339**	1					
34	.125**	.338**	.300**	.249**	.409**	.397**	.441**	1				
35	.370**	.219**	.174**	.224**	.068*	.361**	.111**	.122**	1			
36	-0.032	-0.004	-.127**	.069*	.091**	.213**	.093**	.148**	.128**	1		
37	.490**	.629**	.463**	.331**	.342**	.693**	.356**	.360**	.438**	.119**	1	
38	.191**	.371**	.249**	.435**	.435**	.500**	.528**	.441**	.255**	.227**	.502**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 9

Item Means (Standard Deviations) by Gender, Dual Language Status, Race + Ethnicity, and Time

	Gender		Dual Language Status		
	Male	Female	Dual Language Learner	Bilingual	English Monolingual
1	3.04 (0.84)	3.27 (0.79)	3.01 (0.85)	3.22 (0.87)	3.15 (0.79)
2	2.62 (0.94)	3.04 (0.81)	2.72 (0.99)	3.07 (0.84)	2.74 (0.90)
3	2.63 (1.03)	3.04 (0.90)	2.92 (0.99)	3.11 (0.94)	2.68 (0.99)
4	2.9 (0.97)	3.06 (0.91)	2.92 (0.92)	3.08 (0.87)	2.94 (0.97)
5	2.63 (0.99)	2.91 (0.92)	2.37 (1.01)	2.89 (0.85)	2.80 (0.98)
6	2.76 (1.19)	3.14 (1.05)	3.07 (1.06)	3.35 (0.90)	2.75 (1.20)
7	2.42 (0.88)	2.68 (0.82)	2.36 (0.87)	2.73 (0.81)	2.50 (0.87)
8	3.00 (0.98)	3.19 (0.93)	2.83 (1.04)	3.20 (0.89)	3.10 (0.96)
9	3.48 (0.92)	3.82 (0.52)	3.69 (0.72)	3.70 (0.68)	3.60 (0.82)
10	2.24 (0.88)	2.49 (0.83)	2.20 (0.87)	2.52 (0.80)	2.32 (0.88)
11	3.41 (0.74)	3.55 (0.66)	3.54 (0.66)	3.53 (0.65)	3.43 (0.74)
12	3.29 (0.99)	3.68 (0.74)	3.51 (0.89)	3.61 (0.79)	3.40 (0.94)
13	2.47 (0.98)	2.75 (0.89)	2.27 (1.01)	2.80 (0.80)	2.59 (0.97)
14	3.04 (0.85)	3.33 (0.74)	3.00 (0.89)	3.26 (0.82)	3.18 (0.79)
15	2.44 (1.06)	2.81 (0.92)	2.41 (1.03)	2.87 (0.87)	2.55 (1.05)
16	2.37 (0.97)	2.76 (0.86)	2.54 (0.98)	2.84 (0.83)	2.43 (0.94)
17	2.44 (1.05)	2.87 (0.90)	2.20 (1.11)	2.70 (0.90)	2.71 (1.00)
18	1.95 (1.16)	2.51 (1.05)	2.00 (1.23)	2.47 (1.08)	2.16 (1.13)
19	2.47 (0.89)	2.75 (0.81)	2.51 (0.90)	2.84 (0.84)	2.52 (0.85)
20	2.22 (1.04)	2.63 (0.91)	2.51 (0.98)	2.78 (0.90)	2.23 (0.99)
21	2.40 (1.09)	2.89 (0.92)	2.70 (1.13)	2.81 (0.96)	2.55 (1.04)
22	2.33 (1.11)	2.46 (1.04)	2.55 (1.05)	2.70 (1.03)	2.23 (1.08)
23	2.71 (0.91)	3.02 (0.76)	2.59 (0.94)	3.03 (0.75)	2.85 (0.86)
24	3.06 (0.80)	3.24 (0.74)	3.02 (0.80)	3.27 (0.74)	3.12 (0.78)
25	2.54 (0.94)	2.88 (0.80)	2.65 (0.97)	3.00 (0.76)	2.59 (0.89)
26	2.94 (0.77)	3.08 (0.72)	2.90 (0.84)	3.15 (0.66)	2.97 (0.76)
27	2.77 (1.15)	3.14 (0.90)	3.08 (1.12)	3.09 (0.90)	2.85 (1.09)
28	2.60 (0.83)	2.95 (0.75)	2.69 (0.84)	3.00 (0.71)	2.69 (0.82)
29	2.50 (0.92)	2.70 (0.84)	2.62 (0.93)	2.87 (0.82)	2.48 (0.88)
30	1.81 (1.11)	2.28 (1.05)	1.84 (1.19)	2.32 (1.05)	1.95 (1.08)
31	2.83 (0.80)	3.06 (0.71)	2.75 (0.83)	3.07 (0.71)	2.92 (0.77)
32	3.14 (0.86)	3.33 (0.75)	3.09 (0.90)	3.33 (0.81)	3.21 (0.80)
33	2.20 (0.93)	2.46 (0.86)	2.17 (0.96)	2.48 (0.84)	2.29 (0.91)
34	3.06 (0.90)	3.30 (0.77)	2.95 (0.90)	3.32 (0.71)	3.17 (0.88)
35	3.65 (0.78)	3.83 (0.57)	3.82 (0.62)	3.73 (0.70)	3.72 (0.71)
36	2.24 (1.09)	2.35 (1.06)	2.23 (1.09)	2.16 (1.14)	2.35 (1.04)
37	2.77 (0.96)	3.08 (0.82)	2.88 (0.93)	3.11 (0.88)	2.85 (0.91)
38	2.56 (1.08)	2.87 (1.00)	2.36 (1.06)	2.79 (1.01)	2.74 (1.07)
Factor ρ	0.95	0.94	0.95	0.95	0.95

Note . Factor ρ: estimate of the reliability of latent factor

Table 9

Item Means (Standard Deviations) by Gender, Dual Language Status, Race + Ethnicity, and Time

	Race + Ethnicity			Time	
	African American, Non-Hispanic	Hispanic/Latino	White, Non-Hispanic	Fall	Spring
1	3.12 (0.81)	3.21 (0.79)	3.15 (0.78)	3.04 (0.88)	3.25 (0.74)
2	2.76 (0.93)	2.91 (0.89)	2.69 (0.95)	2.74 (0.91)	2.89 (0.91)
3	2.66 (1.03)	3.01 (0.93)	2.69 (1.00)	2.74 (1.05)	2.89 (0.93)
4	2.90 (0.96)	3.06 (0.90)	2.97 (0.98)	2.91 (0.96)	3.03 (0.93)
5	2.78 (1.02)	2.81 (0.92)	2.67 (0.98)	2.62 (1.02)	2.90 (0.89)
6	2.70 (1.23)	3.18 (1.01)	2.63 (1.24)	2.95 (1.12)	2.92 (1.16)
7	2.49 (0.89)	2.61 (0.86)	2.44 (0.86)	2.46 (0.87)	2.62 (0.86)
8	3.14 (0.95)	3.14 (0.93)	3.00 (1.01)	3.01 (1.00)	3.17 (0.91)
9	3.57 (0.88)	3.67 (0.74)	3.63 (0.75)	3.71 (0.72)	3.56 (0.82)
10	2.29 (0.91)	2.41 (0.83)	2.27 (0.89)	2.25 (0.85)	2.46 (0.86)
11	3.36 (0.80)	3.54 (0.64)	3.53 (0.68)	3.46 (0.72)	3.49 (0.69)
12	3.41 (0.95)	3.59 (0.78)	3.38 (0.96)	3.44 (0.94)	3.50 (0.86)
13	2.53 (0.99)	2.65 (0.90)	2.54 (0.98)	2.47 (0.95)	2.72 (0.93)
14	3.18 (0.80)	3.18 (0.82)	3.22 (0.82)	3.08 (0.87)	3.28 (0.74)
15	2.58 (1.05)	2.72 (0.95)	2.49 (1.06)	2.49 (1.02)	2.72 (1.00)
16	2.39 (0.95)	2.70 (0.90)	2.39 (0.99)	2.48 (0.95)	2.61 (0.92)
17	2.72 (1.02)	2.59 (1.00)	2.66 (1.00)	2.49 (1.03)	2.78 (0.96)
18	2.18 (1.14)	2.27 (1.13)	2.05 (1.21)	2.20 (1.13)	2.22 (1.15)
19	2.58 (0.84)	2.73 (0.81)	2.46 (0.92)	2.51 (0.89)	2.69 (0.82)
20	2.20 (1.02)	2.61 (0.93)	2.22 (1.03)	2.36 (0.99)	2.44 (0.99)
21	2.57 (1.08)	2.72 (1.01)	2.57 (1.07)	2.64 (1.07)	2.62 (1.01)
22	2.16 (1.06)	2.61 (1.02)	2.18 (1.16)	2.36 (1.11)	2.41 (1.04)
23	2.83 (0.86)	2.89 (0.84)	2.75 (0.89)	2.78 (0.88)	2.92 (0.83)
24	3.12 (0.80)	3.22 (0.70)	3.11 (0.77)	3.03 (0.86)	3.25 (0.67)
25	2.61 (0.91)	2.82 (0.84)	2.57 (0.97)	2.64 (0.89)	2.76 (0.89)
26	2.98 (0.78)	3.07 (0.71)	2.97 (0.81)	2.95 (0.76)	3.05 (0.74)
27	2.88 (1.13)	3.01 (1.01)	2.77 (1.11)	2.99 (1.04)	2.88 (1.07)
28	2.73 (0.82)	2.84 (0.79)	2.67 (0.83)	2.74 (0.80)	2.78 (0.82)
29	2.43 (0.91)	2.72 (0.84)	2.48 (0.95)	2.55 (0.90)	2.64 (0.87)
30	1.96 (1.10)	2.11 (1.10)	1.97 (1.15)	1.98 (1.08)	2.07 (1.12)
31	2.89 (0.80)	2.97 (0.76)	2.92 (0.76)	2.87 (0.78)	3.00 (0.75)
32	3.27 (0.79)	3.27 (0.77)	3.19 (0.77)	3.13 (0.93)	3.32 (0.68)
33	2.28 (0.91)	2.38 (0.91)	2.20 (0.90)	2.24 (0.90)	2.40 (0.91)
34	3.11 (0.91)	3.23 (0.78)	3.20 (0.86)	3.10 (0.88)	3.25 (0.81)
35	3.72 (0.73)	3.79 (0.59)	3.68 (0.77)	3.69 (0.77)	3.78 (0.61)
36	2.36 (1.06)	2.25 (1.10)	2.44 (1.06)	2.17 (1.14)	2.41 (1.00)
37	2.87 (0.92)	3.03 (0.83)	2.80 (0.92)	2.82 (0.99)	3.01 (0.81)
38	2.68 (1.14)	2.74 (1.01)	2.69 (1.00)	2.55 (1.09)	2.86 (1.00)
Factor ρ	0.95	0.95	0.95	0.95	0.96

Note . Factor ρ: estimate of the reliability of latent factor

VITA

Kathryn Lynn Black

Candidate for the Degree of

Doctor of Philosophy

Thesis: EVALUATING FACTOR STRUCTURE AND INSTRUMENT STABILITY THROUGH MEASUREMENT INVARIANCE OF THE DECA, SECOND EDITION

Major Field: Educational Psychology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Educational Psychology with a concentration in Research, Evaluation, Measurement and Statistics at Oklahoma State University, Stillwater, Oklahoma in July, 2021.

Completed the requirements for the Master of Science in Human Development and Family Science at Oklahoma State University, Stillwater, Oklahoma in 2012.

Completed the requirements for the Bachelor of Science in Human Development and Family Science at Oklahoma State University, Stillwater, Oklahoma in 2010.

Experience:

Associate Director, Research and Innovation and ECP Projects, CAP Tulsa Fellowship, Strategic Data Project

Professional Memberships: