THE UNITY AND DIVERSITY OF NEUROPSYCHOLOGICAL TASKS OF EXECUTIVE FUNCTIONING: CONSTRUCT AND ECOLOGICAL VALIDITY OF COMMON ASSESSMENT MEASURES

by

SAMANTHA E. JOHN

M.A., University of Colorado Colorado Springs, 2013

B.A., The University of Texas at Austin, 2008

A dissertation submitted to the Graduate Faculty of the University of Colorado Colorado Springs in partial fulfillment of the requirements for the degree of Doctor of Philosophy Department of Psychology 2016
This dissertation for the Doctor of Philosophy degree by

Samantha E. John

has been approved for the

Department of Psychology

By

Brandon Gavett

Molly Maxfield

Kelli Klebe

Lori James

Lee Ashendorf

4/8/2016
Date
Research from cognitive neuroscience has provided evidence that the subcomponents of executive functioning (EF) are strongly related concepts that can take the form of either distinctly independent constructs representing diversity or overlapping constructs with a shared underlying contributor, representing unity. However, this body of research relies upon computerized tests of EF that can distill the domain of executive functioning into some of its smallest and most simplistic subcomponents. While informative to the field of neuropsychology, these studies are only minimally generalizable to clinical neuropsychological settings, and they are limited in their ability to inform treatment recommendations with patient populations. The present study utilized confirmatory factor analysis to evaluate the construct validity of several commonly administered clinical measures of EF. Four hypothesized models were tested, none of which achieved adequate model fit. The best fitting model was a one-factor model with a single global factor representing the domain of EF, $(\chi^2 = 96.04, \text{df} = 27, p < .001; \text{CFI} = .84; \text{TLI} = .79; \text{RMSEA} = .12, 90\% \text{CI} [.09, .14], p < .001)$. Further exploratory models provided evidence for both unity as well as diversity within the domain; however, all of the results suggest that total performance scores are poor measures of subcomponent functioning that should not be used for subsequent predictions about ecological functioning. Alternative approaches to measurement are discussed, as well as the clinical implications of the present findings.
Keywords: executive functioning, construct validity, ecological validity, older adults
DEDICATION

This dissertation is dedicated to my family members that gave me the self-discipline and ambition necessary for getting through the marathon-length training of clinical psychology and the intellectual curiosity necessary for the life-long learning of a career in aging. My parents are the first in a very long line of psychology educators who taught me that life really does suck for everyone, so I should probably get over myself and offer compassion to others. They taught me that delayed gratification is worth the wait and the work. And they taught me - through their constant modeling - that some priorities, no matter how challenging, really are worthy of unfaltering commitment. To Christy, my first psychotherapist, for always being willing to listen, understand, and then cook me a meal. To Papa, for always taking me seriously and for showing me that life passions can, and should, continue into old age. To Momo, for aging so well and so gracefully, and for reiterating all these 95 years that women, and your granddaughters in particular, should pursue whatever the hell they want and not take shit from anyone along the way.
ACKNOWLEDGEMENTS

The graduate career and the dissertation, as the capstone project of that journey, are opportunities to assert one’s independence, to work autonomously toward a self-determined research career. The last six years, and this project in particular, have instead been more of a joint effort. For all my attempts at independence during my graduate career, it has been the contributions and collaborations that have more significantly shaped my future path. For each of these, I am extremely grateful. Our tiny program has given me such focused attention, criticism, and opportunity for growth. Many joint efforts strung together have provided me with the confidence to work independently and the wisdom to know better – that consultation, validation, and the tweaking of ideas will always lead to more successful outcomes than a solo venture.

From interview day to the last day of internship, Brenna Renn has been my academic partner in crime and the coauthor of many joint efforts. She provided me with the support and ambition I needed to get through this journey. Her compassion, intelligence, and big picture ideas have made her my forever colleague. Thanks for taking care of me, now let’s start doing this thing for real.

If I have seen further, it is by standing on the shoulders of Molly and Brandon. I could not have asked for better mentorship giants. These two have guided me through each career challenge and provided me with models I hope to emulate. They are supportive of their students’ ideas and goals and receptive to feedback, always willing to
teach and to prioritize education. Molly, thank you for your endless accountability and your focus on realistic goals and clear-cut expectations. You have taught me to reign in my expectations and get things done. Brandon, thank you for exaggerating my intelligence and holding me to my own unreasonably high standards. You have taught me to experiment with things I do not understand and to strive for new skills and knowledge. I admire you both immensely, and I am continually flattered by your willingness to consider me a colleague.

Thank you to my committee members, Kelli, Lori, and Lee, for your thoughtful contributions and challenges to this dissertation. I hope I can take each of your strengths on this project into my future work. Finally, a big THANK YOU to my fellow researchers on the project for sharing their time and energy with me. Jess, Jen, and Thomas, I could not have completed this without you.
# TABLE OF CONTENTS

CHAPTER

1. INTRODUCTION ........................................................................................................1
   a. Executive Functioning .......................................................................................2
   b. Measuring Executive Functioning ....................................................................5
   c. Clinical Judgment and Use of Assessment .....................................................11
d. The Present Study .........................................................................................13

2. METHOD ...............................................................................................................22
   a. Participants .....................................................................................................22
   b. Measures .........................................................................................................23
c. Procedure .........................................................................................................36
d. Task Scoring and Data Analysis .........................................................................37

3. RESULTS .............................................................................................................40
   a. Participant Characteristics ...............................................................................40
   b. Latent Variable Relationships .........................................................................40
c. Measurement Models .......................................................................................42
d. Models of Prediction .........................................................................................55

4. DISCUSSION .......................................................................................................58
   a. Implications .....................................................................................................64
   b. Limitations .....................................................................................................66
TABLES

TABLE

1. Means, Standard Deviations, and Percentages for Demographic Variables .................................................................41

2. Descriptive Statistics of the Measured Variables Used in the Hypothesized Measurement Models .................................................43

3. Correlation Matrix of Neuropsychological Tests’-Total Performance Scores ..............................................................................44

4. Factor Loadings for the Three-Factor and One-Factor Exploratory Factor Analyses ................................................................53
FIGURES

FIGURE

1. Hypothesized structural model with three correlated latent factors based on the original Miyake model ..............................................................21

2. Hypothesized three-factor measurement model with three correlated latent factors based on the original Miyake model ........................................45

3. Hypothesized bi-factor measurement model with three uncorrelated latent factors based on the Miyake modification model .................................46

4. Hypothesized two-factor measurement model with correlated latent factors based on the Hull model ........................................................................47

5. Hypothesized one-factor measurement model representing a unifying global factor for executive functioning ...........................................................48

6. Two-factor CFA solution based on the Hull model, with standardized estimates and latent factor variances set to 1.00 .............................................50

7. One-factor CFA solution with a global executive functioning factor depicted with standardized estimates ........................................................................52

8. Scree plot from the exploratory factor analysis using Principal Axis Factoring and Oblimin rotation ........................................................................54
CHAPTER 1
INTRODUCTION

Recent research from the cognitive sciences (Miyake, et al., 2000; Miyake & Friedman, 2012; Vaughan & Giovanello, 2010) has provided evidence that the subcomponents of executive functioning (EF) are separate, though strongly related constructs. That is, distinct subcomponents of EF can be measured as independent latent factors contributing variance to performance on given EF tasks, representing diversity, and these factors are strongly correlated to one another, representing unity. Alternatively, there is evidence that unity in the measurement of EF may also be represented as a single underlying factor contributing to performance on all tasks of EF and that diversity in the measurement of EF may best be represented as additional factors of EF subcomponents added hierarchically to the shared factor (Miyake & Freidman, 2012). This relationship is evident in children (Rose, Feldman, & Jankoswki, 2011) as well as young (Miyake et al., 2000) and older adults (Vaughan & Giovanello, 2010). The developmental stability (Friedman, Miyake, Robinson, & Hewitt, 2011; Mischel et al., 2011; Moffit et al., 2011) of this “unity and diversity” pattern (Miyake et al., 2000, p. 72) serves to predict functional as well as self-regulatory behaviors (Hall, Fong, Epp, & Elias, 2008; Klauer, Schmitz, Teige-Mocigemba, & Voss, 2010). Utilizing confirmatory factor analysis (CFA) as well as structural equation modeling (SEM) procedures, this body of research indicates a pattern of shared and non-overlapping statistical variance among
computerized tests of EF (Friedman et al., 2011; Miyake et al., 2000; Miyake & Friedman, 2012). While informative to the broad field of neuropsychology, these computerized tests, the limited number of subcomponents evaluated, and the settings in which tasks are administered to participants bear only minimal resemblance to the wide array of neuropsychological measures most commonly administered during a clinical neuropsychological evaluation. As such, there is little evidence that these studies can directly inform referral questions within a clinical setting. Moreover, the majority of research on the structure and function of EF subcomponents has occurred in a university-derived, cognitively healthy, younger adult population. In contrast, clinicians studying the executive functions are likely to encounter those whose cognitive abilities are declining or impaired, such as older adults or those with neurological injury. There is a present need to extend these findings from cognitive neuroscience into the neuropsychological literature and to utilize the results for clinically and ecologically significant interventions and recommendations. This study demonstrates the need to connect the findings and methodology from the neuroscience literature to the practices and applications of clinical neuropsychological assessment.

**Executive Functioning**

The domain of EF has been increasingly researched in recent years. A relatively nascent and ill-defined concept, the term and theory of EF developed out of the “central executive” described by Baddeley and Hitch (1974). Part of the term’s youth can be attributed to its inherently difficult to understand breadth and scope, and as a result, there is still no agreed upon definition of the term in the literature. Likewise, navigating the construct’s historical trajectory and the identification of its important precursors suffers
from its indefinable nature (Stern, Andersen, & Gavett, 2011). Several have defined the term, and while the definitions reflect similar understanding about the hierarchical nature of this domain in relation to other domains of neuropsychological functioning, its parts and the structure of those parts remain unclear and understudied. Strauss and colleagues describe the collection of functions that fall under the umbrella term of EF as processes within a “system that acts in a supervisory capacity in the overall hierarchy of brain processing” and further define their reach as one that “encompasses skills necessary for purposeful, goal-directed behavior” (Strauss, Sherman, & Spreen, 2006, p. 401). EF has been described as both the pieces of the mind’s structure as well as the overarching structure itself (Stern et al., 2011; Strauss et al., 2006). In the absence of a consensus definition, relevant precursors to the current research include such disciplines as neuroanatomy, psychology, cognitive psychology, and psychometric testing, among others (Ashendorf, Swenson, & Libon, 2013; Barkley, 2012), in which each disparate definition leads also to a slightly or very different understanding of EF’s component parts (Mueller & Dollaghan, 2013).

Many researchers have provided theoretical and research evidence to support the idea that EF abilities play a supervisory role over other domains of functioning and that the EF system itself is made up of multiple parts. Norman and Shallice (1986) described a supervisory attentional system that implied multiple subcomponents. Lezak (1983) delineated possible abilities subsumed by the term EF that included things like volition, planning, purposive action, and effective performance. Delis and colleagues added to those descriptions a longer list of potential abilities: flexibility of thinking, inhibition, problem-solving, planning, impulse control, concept formation, abstract thinking, and
creativity (Delis, Kaplan, & Kramer, 2001). To simplify the evaluation of the abilities associated with EF, Miyake and colleagues (2000) proposed and tested a three component latent variable model that included the often studied and likely related abilities of shifting, updating, and inhibition. Their model, though simplistic in nature, provided evidence for a now foundational aspect of EF research – the subcomponents of EF, the functions themselves, represent independent, but strongly related constructs. That is, each subcomponent contributes a unique portion of the variance to more complex tasks of EF while remaining strongly correlated to the other subcomponents.

The current study mimics the approach and utilizes the framework from Miyake and colleagues’ body of research, in which EF subcomponents represent distinct, but related abilities whose shared and unique variance can be used to predict individual differences in task performance and thus, behavior (Miyake & Friedman, 2012). These closely related constructs pose problems for measurement accuracy, however, as those tasks thought to represent a given subcomponent of EF frequently suffer from task impurity, measuring multiple subcomponents of EF and even other domains of functioning (Miyake et al., 2000). Miyake and Friedman (2012), as well as many others, emphasize the need for additional research on the subcomponents of EF that uses the same latent variable approach to analysis, but that also includes complex tasks and more complex measurement of the purported subcomponents (Mueller & Dollaghan, 2013; Toplak, West, & Stanovich, 2013; Willoughby, Wirth, & Blair, 2012). Such an approach will be necessary for moving beyond their initial conclusions.
Measuring Executive Functioning

Multiple fields of study and areas of research identify the importance of EF and the ability to both understand and accurately measure it. Its relationship to intelligence (Duncan, Emslie, Williams, Johnson, & Freer, 1996) and its role in deciphering the neuroanatomical correlates of behavior within the brain, such as the frontal lobes and associated areas (Barkley, 2012), have made it a critical topic for examination. Clinically, the executive functions may represent one of the first domains of neuropsychological functioning to decline with increasing age (De Luca et al., 2003), and impairment in various subcomponents of EF can be diagnostically indicative of different neurological and degenerative disorders (Lyon & Krasnegor, 1996). More importantly, deficits in the executive functions, perhaps more so than deficits within other areas of functioning, predict challenges in functional tasks and complex behaviors, the hallmarks of daily living (Cahn-Weiner, Malloy, Boyle, Marran, & Salloway, 2000). Traditional tests of EF feature prominently within a list of the most frequently administered assessment instruments of practicing neuropsychologists in the United States and Canada (Rabin, Barr, & Burton, 2005). Tests of EF are interpreted on the basis of one or more summary scores from the test, scaled in reference to a normative sample; this is referred to herein as a performance score. From the clinician’s perspective, assessment of the executive functions is a necessary part of any neuropsychological evaluation and can be used to both conceptualize cases and develop treatment recommendations (Miyake & Friedman, 2012).

Measurement of EF, however, suffers from the ill-defined and difficult to conceptualize nature of the term. The absence of consensus in the term’s definition
extends to its clinical application, in which selection of a given test during administration may be based on factors such as training and familiarity, accessibility, and subjective or anecdotal evidence (Rabin et al., 2005). In the absence of research evidence supporting the reliability and construct and ecological validity of a given measure, clinicians may be making a best or only minimally supported guess at choosing the right instrument or collection of instruments for the assessment of abilities (Rabin et al., 2005). Aside from the conceptual difficulties of the term itself, tests of EF possess the following weaknesses: task impurity, low reliability, reliance on performance scores, and low construct and ecological validity (Jurado & Rosselli, 2007; Miyake, Emerson, & Friedman, 2000).

Task impurity is especially problematic for executive tasks. Some define the executive functions as the abilities that structure the implementation of other cognitive processes (Miyake, Emerson, & Friedman, 2000), making their differentiation from other domains of functioning and their contribution to overall performance scores difficult to discern. Hughes and Graham (2002) describe the task impurity problem from the perspective of the final performance score, where the score itself and the interpretation of the score both contribute to the assumptions made about what abilities and domains are being measured. Because EF is thought to underlie, dictate, and structure most complex tasks, it is believed that complex tasks of behavior must be administered in order to measure its role; however, complex behaviors require the contribution of multiple domains of functioning as well as multiple subcomponents of executive functioning (Burgess, 1997). Trying to assign responsibility for performance to a single domain may be both inaccurate and impossible.
Some theories of neuropsychological assessment look beyond performance scores to help make the distinction among competing contributing domains. The process approach to neuropsychological evaluation relies on observable behaviors and variables that deconstruct the patient’s strategy for a given task. In this approach, an examinee’s errors, approach to the task, and style of responding serve as diagnostic indicators that may evaluate the contribution of EF (Libon, Swenson, Ashendorf, Bauer, & Bowers, 2013). But even that approach is not without challenges. A problem of low process-behavior correspondence may exist. In other words, the same observed performance can be linked to several disparate causes, implicating multiple different domains of functioning. Similarly, a variety of error types are broadly associated with executive dysfunction, but not specifically associated to any one subcomponent of EF (Burgess, 1997). Thus, accurate attribution of errors to domains and/or areas of the brain, especially in the case of impure tasks, requires highly reliable and valid tasks with predictable performance correlates.

According to other cognitive researchers, however, high reliability may never be achieved for tasks of EF. Of core importance when testing the executive functions is the ability to produce a novel problem for solving, a criterion made impossible by the nature of reliability assessment and repeat testing (Burgess, 1997; Denckla, 1996; Salthouse, Atkinson, & Berish, 2003). For complex tasks of behavior, executive control and involvement is considered to be at its strongest during the initial administration (Phillips, 1997), and tasks such as the Wisconsin Card Sorting Test (WCST) exhibit low test-retest reliability estimates with an average of $r = 0.43$ (Bowden et al., 1998). It may also be the case that repeated exposure to complex tasks leads participants to attempt different
strategies, resulting in different overall performances across trials (Phillips, 1997). To make matters worse, the performance scores may limit the information available for understanding the role of EF in a given task, which may have more to do with a patient’s approach to the task and less to do with the overall performance. If it is the executive functions that dictate the structure imposed upon and the strategies utilized for a given task’s completion, then performance scores are lacking in information useful for interpreting EF’s role. There is a need for scores that measure the structure and strategy throughout execution of the task (Anderson, Levin, & Jacobs, 2002). Anderson and colleagues propose the use of scoring methods that highlight process oriented analysis of performance rather than reliance on an overall performance score alone. Though as might be expected, the administration of complex tasks requires a highly structured set of procedures that limits the extent of participant strategy use (Lezak, Howieson, Bigler, & Tranel, 2012).

**Construct validity of executive functioning tests.** The dearth of conceptual clarity plaguing EF also plagues the development of instruments intended to measure the construct. Initial validation of these tests is often dependent upon the criterion that they detect frontal lobe damage; however, the exact subcomponent of executive functioning necessary for successful completion of the task may then remain unspecified (Miyake et al., 2000). Determining the construct validity for a group of tests based on an ill-specified definition results in disagreement about what a given test measures. The WCST, for example, has been described as a measure of mental set shifting, inhibition, cognitive flexibility, problem solving, and organization/categorization (Miyake, Emerson, & Friedman, 2000), though it is likely that all of those subcomponents are
involved to varying degrees, depending on the examinee. Construct validity is particularly important for clinicians who may be selecting instruments for the assessment of a given subcomponent. These choices may end up being the result of assumptions, as there are few independent empirical studies to demonstrate what is being measured by a specific task (Miyake, Emerson, & Friedman, 2000). This is perhaps most problematic when assumptions about task validity and impaired performance are used to inform treatment recommendations, highlighting the need for ecological validity in addition to construct validity.

**Ecological validity of executive functioning tests.** While there is some evidence that performance scores from tasks of EF are predictive of broad categories of behavior (i.e., impaired or not) (Hall, Fong, Epp, & Elias, 2008; Klauer, Schmitz, Teige-Mocigemba, & Voss, 2010), minimal evidence exists to connect specific task performance to specific real life behavioral correlates (i.e., high perseverative errors on WCST indicates dangerous utilization behaviors in the home) (Jurado & Rosselli, 2007). There is only limited evidence that neuropsychological measures of EF inform ecologically valid conclusions. Of the studies demonstrating positive relationships, informant or self-report questionnaires about daily functioning often serve as the proxy for an ecologically valid task. For example, impaired performance on measures of executive functioning (e.g., Trail Making Test (TMT), FAS fluency, and error-related measures) has significantly correlated with caregiver ratings of everyday functioning and served as a stronger predictor of impairment than performance on other neuropsychological measures (i.e., tests of premorbid intellectual function, IQ, memory, or language); however, these EF tests were only moderately correlated to the factor
scores representing the dysexecutive symptoms of inhibition, intentionality, and working memory (Burgess, Alderman, Evans, Emslie, & Wilson, 1998). Within a sample of patients diagnosed with probable vascular dementia, longitudinal declines in caregiver ratings of instrumental (IADL) and basic activities of daily living (BADL) were significantly and solely predicted by declines on the Initiation/Perseveration and Memory subscales of the Dementia Rating Scale (DRS), respectively (Jefferson, et al., 2006).

Examination of executive test batteries such as the Behavioral Assessment of Dysexecutive Syndrome (BADS) (Wood & Liossi, 2006) and the Hayling and Brixton Tests (Odhuba et al., 2005) has yielded only limited and moderate ecological validity, respectively, within brain-injured samples. Other studies, using a variety of patient populations, have shown an absence of correlation between measures of EF and measures of everyday functioning (Amieva, Phillips, & Della Sala, 2003; Bogod, Mateer, & MacDonald, 2003; Chan, 2001; Norris & Tate, 2000; Ready, Stierman, & Paulsen, 2001).

Due to the absence of clear positive relationships between executive tasks and daily functioning, consensus findings about specific tests are necessary goals for further research. Of the Stroop, TMT, The Controlled Oral Word Association Test (COWAT), and the WCST, only Trails B and the Stroop - Color Word interference trial were correlated with informant measures of everyday functioning (The Dysexecutive Questionnaire – DEX and The Brock Adaptive Functioning Questionnaire – BAFQ). These results, however, were attributed to the relationship between speeded processing and everyday functioning, a rather narrow aspect of executive control (Chaytor, Schmitter-Edgecombe, & Burr, 2006). Similar findings occurred within a group of patients with cardiovascular disease, providing further evidence that a measure of
speeded inhibition (the DKEFS Color-Word Interference Tests) could predict informant ratings of IADLs (Jefferson, Paul, Ozonoff, & Cohen, 2006). Given the role of EF within complex behaviors as well as the absence of clear evidence for executive tasks to correlate with real world functions, the current body of executive measures appears inadequate for drawing strong conclusions between performance on tasks and performance in day-to-day life.

**Clinical Judgment and Use of Assessment**

The implications of a clinical assessment extend further than the diagnosis made and the identified errors. A neuropsychological assessment should also produce an evaluation of the patient’s strengths and weaknesses as well as a prediction about everyday functioning (Morgan & Heaton, 2009; Wood, 2007). This evaluation may broadly evaluate all areas of functioning and/or evaluate a narrow scope of functioning such as the fitness to drive, to manage finances, or to make healthcare decisions (Wood, 2007). The need for ecologically valid tests of neuropsychological functioning currently exceeds the research evidence linking task performance to real world behavior. The use of neuropsychological tests to predict behavioral consequences rests “on the assumption that demonstration of an impaired ability in the laboratory indicates difficulty or failure with functions that rely upon that ability in the natural environment,” an argument that relies upon the construct validity of the selected measures (Morgan & Heaton, 2009, p. 638). Even in the presence of research evidence linking test performance to behavior, little work has been done to delineate the cognitive abilities necessary or sufficient for specific real world behaviors (Farias, Harrell, Neumann, & Houtz, 2003). Not only that, but neuropsychological constructs do not map directly onto the legal and medical
constructs they are often asked to determine (Wood, 2007), as is needed for capacity or other forensic evaluations.

A small subset of existing tests attempt to measure performance on functional tasks. These measures estimate ability on the instrumental activities of daily living (IADLs) that are included in the clinical diagnostic criteria for determination of cognitive impairment: driving, financial management, medication management, grocery shopping, cooking, cleaning, etc. (Lezak et al., 2012). Of all the domains of functioning, EF may be the most closely linked to performance of the IADLs (Cahn-Weiner et al., 2007), highlighting the need for additional research linking specific executive functions to performance errors on measures of functional daily behaviors. Among the most complex of the IADLs, driving ability and financial management likely require intact EF across many different EF subcomponents. These two may also be the most amenable to in-vivo research approaches, as the ability to perform both tasks can be directly measured without having to rely upon contrived environmental controls in the laboratory setting that may create artificial advantages/disadvantages. However, just a few research studies have linked on-road driving performance to test performance of that ability (Abularach, Seichepine, Tripodis, Gavett, & Stern (2013); Asimakopulos et al., 2012). Likewise, financial management and bill paying behavior is more frequently researched for the determination of capacity rather than for the purpose of identifying the subcomponents underlying those abilities (Okonkwo, Wadley, Griffith, Ball, & Marson, 2006; Sherod et al., 2012). In response to these research shortcomings, performance-based measures with an emphasis on verisimilitude, in which functional capacity is directly assessed and used as an outcome variable in statistical prediction with other tests, should be embraced
(Morgan & Heaton, 2009). There is a dearth of research currently utilizing this approach, but some growing support for the use of both technology and cognitive neuropsychological approaches to improve the measurement and prediction of ecological behaviors (Parsons, Carlew, Magtoto, & Stonecipher, 2015).

In the absence of improved research approaches, it is clinician training and intuition that will dictate test selection within a flexible battery approach to neuropsychological assessment, utilizing the best available tests for the specific referral question (Rabin et al., 2005). In many cases, these choices are based on the clinician’s desire to confirm family members’ observations and to produce a list of symptomatic behaviors from which an accurate diagnosis can be made (Qualls & Smyer, 2007). Clinician decision-making may then depend on a host of unmeasured or subjective factors. For example, neuropsychological measures account for 10.2% - 78.0% of the variance of the yes/no decision made within a capacity assessment (Gurrera, Moye, Karel, Azar, & Armesto, 2006), and account for approximately 16-30% of the variance in the clinical diagnostic decision between two competing dementia diagnoses (John, Gurnani, Bussell, Saurman, Griffin, & Gavett, under review), highlighting the variability of test efficacy across situations. Research that can identify both the construct and the ecological validity of individual tests within relevant samples may therefore may be able to reduce the situational variability of test efficacy, and should be a goal of the field.

**The Present Study**

**Executive functions in older adults.** Miyake and colleagues attempted to determine the role of several subcomponents of EF within complex tasks of behavior within a laboratory setting (Miyake et al., 2000). Each of their younger adult participants
completed a series of nine computerized versions of neuropsychological tasks of EF as well as two complex tasks of behavior, the Tower of Hanoi (TOH) and the WCST. Each of the nine computerized tasks utilized was administered within a laboratory setting and was selected to represent a clean and highly standardized measure of executive behavior (Hull et al., 2008; Miyake et al., 2000). Utilizing CFA, the researchers determined that the nine simple tasks loaded separately onto one of three EF subcomponents - shifting, updating, or inhibition - and that these three constructs measured distinct but correlated abilities. Utilizing SEM, the researchers then determined that shifting, updating, and inhibition contributed differently to TOH and WCST performance, with a single construct predicting each of the two complex behavioral tasks. Inhibition best predicted TOH performance, whereas shifting best predicted WCST (Miyake et al., 2000). When this procedure was replicated in a cognitively healthy older adult sample, a group known to experience declines in EF ability (Hull et al., 2008; Miyake & Friedman, 2012), a different measurement model of underlying contributions to simple tasks and a different structural model predicting complex task performance emerged (Hull et al., 2008).

Simple tasks of EF loaded onto only two constructs, shifting and updating, while the inhibition construct, previously noted as the main component of “unity” in the younger adult model, failed to fit the sample data. For older adults, the updating subcomponent was the strongest predictor of complex task performance on both TOH and WCST (Hull et al., 2008). Taken together, these two studies reveal the following: 1) the underlying subcomponents of EF contribute to simple task performance differently for different aged cohorts, with younger adults demonstrating stronger evidence for both unity (as measured by the correlation among factors) and diversity (as measured by the presence of multiple
supported underlying factors); 2) the identified subcomponents of EF differentially contribute to and predict complex task performance; 3) these contributions vary by both task demand and sample population; and 4) complex behavioral tasks may therefore be approached differently by different aged samples, explaining the observed differences in model fit for younger and older adults and the need for process-oriented examination of task performance.

The research of Miyake and colleagues (2000, 2012) and Hull and colleagues (2008) provides the basis for improved understanding of the sub-components of EF and their relationship to age-related cognition, but they unfortunately provide only minimal insight about the clinical neuropsychological assessment of these abilities and their subsequent role in ecological behaviors. Computerized measures of EF are rarely used within clinical assessments (McGuire, 2014; Tanguay, Davidson, Guerrero Nuñez, & Ferland, 2014) and they may be measuring different constructs in different aged samples, or different constructs than their paper and pencil equivalents (Weigold, Weigold, Drakeford, Dykema, & Smith, 2016). There is a need for the cognitive neuropsychological approach to embrace the use of existing, commonly used clinical measures in order to translate the findings of Miyake and Hull into real-world application. This is particularly true for an older adult sample for whom the prediction and measurement of ecological behaviors is common in clinical practice (Carr & Ott, 2010).

Clinical neuropsychology has made attempts to design tasks that measure and represent ecologically valid behaviors, such as financial management, driving ability, and capacity to consent (Besnard et al., 2015), but the relationship between these tasks and
other measures of EF implies that a measurable level of impairment already exists and that patient errors in performance are necessary for nuanced measurement. Ceiling effects on these tasks make them less sensitive to subtle changes in cognition, and measurable errors in performance do a better job of classifying patient diagnostic status than they do at predicting performance on ecological tasks. A verisimilitude approach linking EF tasks to an ecologically valid measurement of real world behavior would be the ideal process for determining task success for the fitness to drive or management of finances, but many clinicians are asked to make these determinations using only neuropsychological tasks (Besnard et al., 2015; Carr & Ott, 2010).

**Combining approaches.** In the clinical neuropsychology research literature, little of the necessary statistical evidence exists for guiding task selection and making patient recommendations during and after an evaluation (Sweet, Nelson, & Moberg, 2006). There is an absence of literature confirming construct validity of tasks and a dearth of ecologically valid measures (Miyake et al., 2000; Miyake & Friedman, 2012). Though instruments may demonstrate sensitivity and specificity for differential diagnosis, less is known about the relationship between patient performance on those instruments and daily functioning. This gap in the literature is most pervasive for the domain of executive functioning. Those suffering from a neurodegenerative disorder are often in great need of treatment recommendations at the time they present for outpatient evaluation, demonstrating a present and ongoing demand for recommendations that coincide with referral questions pertaining to functional performance (Qualls & Smyer, 2007). Evidence linking tasks of EF to ecologically valid conclusions about one’s ability to perform complex behaviors is needed.
Cognitive neuroscience has tended to study the simplest of tasks, the closest things to “pure” measures of a given ability, to make sense of the building blocks involved in EF (Miyake et al., 2000). Clinical neuropsychology has tended to focus on the ecological relevance and diagnostic implications of impaired functioning, where patterns of errors on tests are extrapolated to inform predictions about a patient’s diagnosis, and his or her ability to successfully function in daily life (Delis, Jacobson, Bondi, Hamilton, & Salmon, 2003; Qualls & Smyer, 2007). Few researchers have attempted to meld the two approaches. As an important concept in both research and clinical practice, EF research may benefit from a combination of the two approaches, utilizing quantifiable aspects of functioning through testing, but with an emphasis on the types of tests that capture qualitative behaviors and inform clinical conclusions (Ashendorf et al., 2013). Meta-analyses examining the correlations between executive task performances and functional abilities may provide an introductory step in melding the two approaches. Asimakopulos and colleagues conducted a meta-analysis comparing EF task performance to on-road driving behavior and identified the following constructs as necessary contributors: planning, working memory, and cognitive flexibility (Asimakopulos et al., 2012). While not a direct extension of Miyake’s model, the meta-analysis captures and defines three subcomponents of clinically measured EF that relate to Miyake’s cleaner subcomponents of shifting, updating, and inhibition. Further research is needed to evaluate the predictive power of these three clinically measured constructs and their relationship to the cognitive neuroscience models of EF.

This project is a starting point for a larger body of research necessary for determining the construct and ecological validity of clinical EF tasks and their use in
developing patient recommendations across different levels of functioning. In navigating the gaps between the approaches of cognitive neuroscience and clinical neuropsychology, there is a myriad of questions that could be asked; this project highlights a single example of how methods from cognitive neuroscience can be applied to a clinical question. As such, this study followed the suggestions laid out by Miyake and Friedman (2012) and utilized their approach of CFA and SEM to identify the subcomponents of EF and to predict complex task performance using clinical neuropsychological tasks of EF in an older adult sample. Given the dearth of research linking clinical tasks to ecologically valid behaviors, the results of Asimakopulos and colleagues’ meta-analysis were utilized to hypothesize the best fitting models of clinical neuropsychological tasks in the prediction of complex, ecological behaviors, in this case, driving fitness and financial management. More specifically, the models evaluated the relationship among three hypothesized subcomponents of EF thought to underlie driving fitness - planning, working memory, and cognitive flexibility - and a group of nine EF tests often administered clinically to inform fitness to drive decisions (Asimakopulos et al., 2012). Though somewhat messier than the constructs utilized in previous research (Hull et al., 2008; Miyake et al. 2000), the three identified constructs of planning, working memory, and cognitive flexibility are frequently cited subcomponents of EF that may more closely map onto the abilities measured by several clinical neuropsychological measures of EF (Asimakopulos et al., 2012; Miyake & Friedman, 2012; Strauss et al., 2006). They have also been linked to performance on tasks of financial management (Martyr & Clare, 2012).
**Hypothesized models.** CFA and SEM models were hypothesized for the older adult sample according to the previously supported models utilized by Miyake et al. (2000) and Hull et al. (2008); however, the use of impure, clinical neuropsychological tasks was expected to produce weaker model fit as well as cross-loadings among the measured variables onto the latent factors. These considerations led to the hypothesis of additional models and approaches which might better account for the messiness of the data. The present study hypothesized the following set of potential findings and associated models, which were tested in this order: 1) CFA and SEM on an older adult sample will demonstrate the unity and diversity principle of the executive functions within neuropsychological tasks used clinically, and this may take the form of 1a) a three-factor model that follows directly from Miyake et al. (2000) with relationships existing among all three latent constructs and with measured variables loading more strongly, though perhaps not independently, onto one of those three distinct constructs – planning, working memory, and cognitive flexibility (see Figure 1); or 1b) a three-factor model that represents the underlying contribution of a single *unifying* factor to all tasks of EF, with separate factors contributing additional variance in performance to subsets of EF tasks (see Figure 2); 2) CFA and SEM on a cognitively heterogeneous older adult sample will demonstrate the pattern of model fit found by Hull and colleagues (2008), in which only two subcomponents of EF, planning and updating, will contribute to simple EF task performance, as the component of *inhibition* (or cognitive flexibility here) was not previously supported in older adults (see Figure 3); 3) Clinical measures of EF are too impure to be used for the isolation of EF subcomponents, implicating a single-factor model of EF which underlies the performance on all tasks (see Figure 4); and 4) The
subcomponent of working memory, the clinical correlate to Miyake’s and Hull’s
updating will be most predictive of complex task performance, in this case, older adult
driving ability and financial management as measured through proxy assessment
measures of those two abilities (see Figure 1) (Asimakopulos et al., 2012; Hull et al.,
2008; Miyake et al., 2000).
Figure 1. Hypothesized structural model with three correlated latent factors based on the original Miyake model.
CHAPTER 2

METHOD

Participants

Participants were 192 community-dwelling older adults between 54 and 92 years of age. All participants included in the study met the following eligibility criteria: 1) no more than moderate memory impairment as measured by the Memory Impairment Screen-Telephone version (MIS-T score of 3 or higher); 2) intact visual acuity measured through the Snellen eye chart; 3) intact auditory processing measured through ability to respond accurately to questions over the phone; 4) able to speak and read in English; 5) still driving and managing own finances; 6) free of diagnosis of a neurological or neurodegenerative disorder (e.g., multiple sclerosis, Parkinson’s disease, Alzheimer’s disease, and other dementias); and 7) living independently in the community.

Participants were recruited from a variety of sources, including local community advertisement, older adult community housing, the UCCS Gerontology Center registry, and through local organizations, including the Colorado Springs Lions Club, Argonauts, and Senior Center visitors. Participants received financial compensation for their time and participation at the rate of $10 per study session, which was roughly one hour in length.

To ensure a cognitively heterogeneous sample of older adults, participants were evaluated with a brief screening tool of cognitive function via telephone (described
below) when they were contacted to participate. Cognitively healthy participants as well as those exhibiting mild to moderate levels of impairment, as determined by cut-off scores on the screening instrument, were scheduled to participate. Those exhibiting severe impairment based on a total score cut-off were excluded from participation given the burden of extensive testing and the absence of driving fitness applicability. These individuals were informed of other study options at UCCS for which they qualified.

**Sample size.** Target recruitment was based on the proposed CFA and SEM models and the number of parameters to be estimated. Simulated data and model testing using R and M-Plus software was performed prior to study start to identify a target sample size sufficient for model fit. The full hypothesized SEM model included 30 free parameters for estimation. Using parameter estimates from the existing relevant literature to create a hypothetical model (Miyake et al., 2000; Hull et al., 2008), a Monte Carlo Simulation of 10,000 studies suggested that similar or slightly weaker relationships among variables could be estimated at a power of .8 or higher with a sample size of 170. This number was the recruitment goal. When additional resources became available to assist with participant payment, recruitment continued until funds were depleted, giving an overall sample size of 192.

**Measures**

The following listed measures made up the battery of neuropsychological tasks and questionnaires administered to all participants. Neuropsychological measures were selected on the basis of frequent clinician use (Rabin et al., 2005), research evidence of ecological validity in fitness to drive evaluations (Asimakopulos et al., 2012), accessibility for the project, and amenability to future process analysis and scoring.
(Ashendorf et al., 2013). All selected tests have been standardized and normed for administration on adult and older adult samples; however, normative corrections were not applied to participant scores.

**Demographic questionnaire.** Participant demographics were collected as recommended by the American Psychological Association (APA; 2009) and included 1) date of birth and age (reported by the participant and measured in chronological years), 2) sex (self-reported as male or female), 3) educational attainment (measured in years of formal schooling), 4) ethnicity (self-reported as Hispanic/Latino/Spanish origin, Not Hispanic/Latino/Spanish Origin, and Unknown), and 5) race (self-reported as American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White or Caucasian, and Other), and handedness (self-reported as Right, Left, or Ambidextrous). Race and ethnicity data were categorized according to the recommendations of the Office of Management and Budget used by the most recent U.S. Census guidelines (2010 Census guidelines, as set forth by Sink, 1997). Participants also self-reported their health status according to a five-point Likert scale. Self-rated health was assessed using the item, “In general, would you say your health is…” with a scale from 1 = excellent to 5 = poor. The specific wording of this item was drawn from the National Health Interview Survey (Idler & Benyamini, 1997). This item has demonstrated good test-retest reliability ($r = .92$; Lorig et al., 1996).

**Memory Impairment Screen – Telephone version (MIS –T).** This measure was administered over the phone during the recruitment and scheduling phase for older adult participants as a brief measure of memory and screen for possible impairment. The Memory Impairment Screen – Telephone version is a short, five-minute memory
screening tool involving categorization, as well as free recall and cued recall (if necessary). The researcher placing the call presented two words and then asked the participant to repeat them aloud. The participant was then asked to choose the word, of the two presented, that fit a category cue provided. The researcher then stated the category cue for the other target word. This procedure was then repeated with two new words. Following a two to three minute interference period, the participant was then asked to recall the target words. The total MIS-T score was calculated by doubling the number of items retrieved during free recall and then adding the number of items retrieved during cued recall to that total. Scores range from 0-8. Scores of 0-5 are considered impaired. Participants scoring greater than or equal to 3 on the screen were invited to participate. At a cut-off score of 3, sensitivity is .67 and specificity is .98 for discriminating between healthy controls and individuals with dementia (Lipton et al., 2003).

**Montreal Cognitive Assessment (MoCA).** This measure was administered to older adult participants as a brief measure of global cognitive functioning. It was included in the battery to help characterize the sample of participants and ensure cognitive heterogeneity within the sample. There is evidence that the MoCA can detect subtle changes in cognition and may accurately identify those with mild cognitive impairment (MCI) who may score 25 or higher on the MMSE (Smith, Gildeh, & Holmes, 2007). With a cut-off score of 26, the MoCA had a sensitivity for detecting MCI of 83% and a specificity of 50% within a community mental health memory clinic sample (Smith, Gildeh, & Holmes, 2007).
Neuropsychological Assessment Battery (NAB) – Driving Scenes. The Driving Scenes subtest served as the outcome variable in the SEM. It served as the clinical assessment proxy for driving fitness. Driving Scenes subtest is one of 33 tasks within the Neuropsychological Assessment Battery (NAB; Stern & White, 2003). The Driving Scenes subtest is the Daily Living task taken from the Attention module. In the task, the participant was presented with a series of visual images depicted as though from behind the steering wheel of a vehicle. The participant was asked to view each subsequent image and compare it to the image previously viewed to identify content that is new or modified. The task was timed with each new scene allotted a minimum of 60 seconds and a maximum of 120 seconds of viewing time. The number of correctly identified modifications in each scene was summed to produce a total performance score. Scores range between 0 and 69. White and Stern note that performance on the full NAB battery is affected by age (White & Stern, 2003b) and to a lesser extent both education and gender (White & Stern, 2003a). The battery was standardized on a census-matched sample from the United States in 2001-2002.

The Driving Scenes subtest, as measured within the standardization sample, exhibits marginal ($r = .60 -.69$) test-retest reliability (White & Stern, 2003b). Content validity of all NAB subtests was guided by a review of research literature in neuropsychology as well as the results of a survey of needs and practices of current clinicians (Stern & White, 2000). Independent and subsequent research evaluating the Driving Scenes subtest has provided initial evidence of ecological validity. Brown and colleagues correlated test performance to on-road performance ($r = .55$) and demonstrated a difference in performance between healthy participants and those
suffering from a mild dementia (Brown et al., 2005). Abularach and colleagues evaluated a mixed clinical sample of adults, aged 55-90, on both an on-road driving evaluation as well as neuropsychological measures of functioning. Of the tests administered, both NAB Driving Scenes and NAB Mazes, among others, independently predicted on-road driving performance; Driving Scenes along with another test of visual perception predicted on-road driving status with 90% accuracy (Abularach et al., 2013).

**NAB Bill Payment.** The Bill Payment subtest served as an outcome variable in a secondary SEM. It served as the clinical assessment proxy for financial management. The Bill Payment subtest is one of 33 tasks within the Neuropsychological Assessment Battery (NAB; Stern & White, 2003). The Bill Payment subtest is the Daily Living task taken from the Language module. The task required participants to examine a utility bill statement and check ledger and to fill out a check and envelope in order to accurately pay a bill. To successfully complete the task, participants must follow a series of eight commands that require oral and written responses. The functional task is a measure of auditory language comprehension, reading comprehension, writing, simple calculations, and speech output (NAB; Stern & White, 2003). Scores on the test range from 0-19, with higher scores indicating better performance. The test has adequate internal consistency within a 70-97 year old sample ($\alpha = .72$-$79$) and low ($\leq .41$) test-retest reliability across age groups (White & Stern, 2003b). Information regarding the standardization and content validity of the NAB is included in the section above.

**NAB Mazes.** The Mazes subtest was one of three observed measures in the estimation of the latent variable of *Planning* within the proposed CFA models. The Mazes subtest is one of four tasks within the Executive module of the NAB (Stern &
White, 2003). In the task, the participant was presented with a series of seven increasingly more difficult timed visual mazes and asked to trace a path from start to finish. Scores for each maze were allotted on the basis of time to completion, and scores may range from 0-2 for easier mazes and 0-5 for more difficult mazes. The performance score for the task is the summed total of points across all mazes and will range between 0 – 26. Information regarding the standardization and content validity of the NAB is included in the section above. The Mazes subtest, as measured within the standardization sample, exhibits adequate ($\alpha = .70-79$) internal consistency and low ($\leq .59$) test-retest reliability (White & Stern, 2003b).

**Rey-Osterrieth Complex Figure Test – copy trial (ROCF).** The ROCF copy trial was one of three observed measures in the estimation of the latent variable of Planning within the proposed CFA models. The ROCF is a measure of visuospatial processing, planning, and visual memory. The full test consists of four separate tasks, a copy, immediate, delay, and recognition trial. There are a number of versions as well as scoring systems developed for the ROCF (Strauss et al., 2006). The version utilized for the present study required examinees to copy a complex figure. The ROCF has been normed for ages 18-89 years. The scoring system suggested within the manual assigns a value of 0, 0.5, 1, or 2 to each of 18 component parts of the overall figure; the scores allotted to each part are then summed to obtain the overall performance score for each trial. As such, raw scores range from 0.0 to 36.0 for the copy trial. A score of 2 for a given component of the figure is only achieved when a participant both accurately draws and accurately places the component of the figure relative to the overall drawing (Meyers & Meyers, 1996). Mitrushina and colleagues suggest that an age-related decline in
performance on the copy trial as well as increased variability in scores is demonstrated after age 60 (Mitrushina, Boone, Razani, & D’Elia, 2005). Others have suggested that the primary error associated with advanced age is omission of details from the drawing (Hartman & Potter, 1998; Mitrushina, Satz, & Chervinsky, 1990). The effects of gender and education are less well established (Strauss et al., 2006).

Both split-half as well as alpha coefficients have been used to evaluate internal reliability with \( r > .60 \) for the copy condition (Berry, Allen, Schmitt, 1991; Fasteneau, Bennett, Denburg, 1996). Mitrushina and Satz report much lower values for test-retest reliability among older adults, with values ranging from .56 to .68 (1991). Factor analytic and correlational studies of construct validity show evidence for the underlying contributions of visual-constructional ability as well as visual memory; however, the exact contribution of EF ability and its subcomponents are not well understood (Strauss, Sherman, & Spreen, 2006).

**Clock Drawing Test – to command.** The Clock Drawing Test was one of three observed measures in the estimation of the latent variable of *Planning* within the proposed CFA models. The Clock Drawing Test exists in more than a dozen versions and is intended to screen for dementia as well as measure visual-spatial, constructional, and executive abilities (Strauss et al., 2006). The free-drawing version of the task asks examinees to draw a clock, including all of the numbers, and then to set the time to “10 after 11” (Strauss et al., 2006). Given the somewhat subjective scoring of reconstructed stimuli, several quantitative scoring systems exist for the task, including a simple 3-point system (Goodglass, Kaplan, & Barresi, 2001), a 10-point system (Libon, Swenson, Baroski, & Sands, 1993; Sunderland et al., 1989), and a 20-point system (Mendez, Ala, &
Two different scoring methods were utilized for the Clock Drawing Test in the present study. A simple 3-point system was utilized as a result of the clock’s inclusion within the global cognitive screening measure used in the study, the Montreal Cognitive Assessment (MoCA; see above), which relies upon 3-point scoring. A more complex 10-point system was utilized within all of the tested models because of the evidence that more complex scoring related directly to planning and driving ability (Rouleau, Salmon, Butters, Kennedy, & McGuire, 1992). Previous research has noted that the different scoring methods of the Clock Drawing Test tend to correlate highly with one another. Royall and colleagues found evidence of construct validity among six different scoring systems, with values ranging from -.73 – .95 (Royall, Mulroy, Chiodo, & Polk, 1999). There is significant evidence that the task is effective at differentiating healthy older adults from those suffering from a neurodegenerative disorder (Aprahamian, Martinelli, Neri, & Yassuda, 2009; Manos, 1999; Mittal, Gorthi, Rohatgi, 2010; Royall, Cordes, & Polk, 1998; Tuokko, Hadjistavropoulos, Miller, & Beattie, 1992), some evidence that the score on the task can inform severity of impairment (Nair et al., 2010; Royall et al., 1998), but less evidence that the task can be used effectively for differential diagnosis (Heinik, Solomesh, Raikher, & Lin, 2002).

**Wechsler Adult Intelligence Scale –III (WAIS-III) - Digit Span Backwards.**

The Backwards Digit Span subtest was one of three observed measures in the estimation of the latent variable of *Working Memory* within the proposed CFA models. The Digit Span task is taken from the WAIS-III and represents one of two tests included within the Working Memory subscale on the Verbal IQ index. The full WAIS-III battery is normed for ages 16-89 and is intended to provide a measure of general intellectual functioning; it
is a revision of the earlier published WAIS-R (Wechsler, 1981). The standardization sample is based upon the 1995 U.S. Census (The Psychological Corporation, 1997). The Digit Span test is made up of two tasks, Digits Forward and Digits Backward. The Backward task requires participants to repeat the number sequences in the reverse order from the oral presentation. There are two trials for each span length, from a series of 3 digits to 9 digits, producing a total of 14 trials and total score from 0 to 14. In some cases, the scores from both Digits Forward and Digits Backward are combined to create a single total score, but other researchers advocate the separation of the two tasks, as they may be measures of different constructs, with Digits Forward representing more strongly the skill of auditory attention and Digits Backward representing more strongly the skill of working memory (Lezak et al., 2012; Strauss et al., 2006). As a result, only the Digits Backward portion of the subtest was utilized in this study. The Digit Span task possesses very high (α = .90+) internal consistency and high (r = .80 - .89) test-retest reliability. Factor analytic studies consistently demonstrate the Digit Span’s relation to working memory and other tasks intended to measure that construct (Strauss et al., 2006).

**NAB Dots.** The NAB Dots subtest was one of three observed measures in the estimation of the latent variable of *Working Memory* within the proposed CFA models. The test is taken from the Attention module of the NAB (Stern & White, 2003). In the task, the participant was presented with a visual image of colored dots for ten seconds, followed by a distractor page of the same color for five seconds, followed by a second image of the same-colored dots in the same arrangement as the first page. The third page contains one additional colored dot and participants were asked to point out the new dot. The number of correctly identified single dots from across all three-page sequences was
summed to produce a total performance score. Scores range between 0 and 12.

Information regarding the standardization and content validity of the NAB is included in the section above. The Dots subtest, as measured within the standardization sample, exhibits adequate ($\alpha = .79$) internal consistency and low ($\leq .44$) test-retest reliability (White & Stern, 2003b).

**Wechsler Memory Scale –III (WMS-III) – Mental Control.** The Mental Control subtest was one of three observed measures in the estimation of the latent variable of *Working Memory* within the proposed CFA models. The test is an optional measure from the Wechsler Memory Scale – version III and younger. Participants were asked to recite a series of familiar sequences (e.g., days of the week, numbers 1-20, months of the year) in forward and reverse orders as well as recite and manipulate novel sequences. Participant responses are timed and scored on the basis of both time in seconds as well as the occurrence of errors. Scores range from 0-40, with higher scores indicating better performance. Mental Control has a high generalizability coefficient (.80-.89) (Strauss, et al., 2006).

**Delis-Kaplan Executive Function System - Design Fluency.** The D-KEFS Design Fluency subtest was one of three observed measures in the estimation of the latent variable of *Cognitive Flexibility* within the proposed CFA models. The Design Fluency subtest is one of nine tasks within the Delis –Kaplan Executive Function System (D-KEFS; Delis, Kaplan, & Kramer, 2001). The Design Fluency test was created as a non-verbal analogue to verbal fluency procedures (Jones-Gotman & Milner, 1977). In the original design fluency task, examinees draw as many different designs within a five-minute time limit as possible that are neither namable forms nor scribbles (Strauss et al.,
The version within the D-KEFS pulls methodology from a revision created by Regard, Strauss, and Knapp (1982). The D-KEFS Design Fluency is composed of three conditions in which rows of boxes containing an array of dots are presented to the examinee. The examinee must connect the dots, with four lines only, to make different designs. For the study, participants were administered all three conditions. In the first condition, Filled Dots, each response box contains five filled dots and the examinee is given 60 seconds to draw as many designs as possible. For condition 2, Empty Dots Only, each response box contains five filled dots and five unfilled dots. The examinee must only connect the unfilled dots, also within a 60 second time frame. Condition 3, Switching, represents a new procedure in which the examinee is presented with response boxes that contain both filled and unfilled dots and must draw designs by alternating between filled and unfilled dots for each adjacent connection. Condition 3 is intended to test both fluency and cognitive flexibility. The total number of correctly drawn designs in each trial is tallied, and these values can be summed across trials to create total scores (Delis, Kaplan, & Kramer, 2001). Item interdependence precluded the use of internal consistency procedures; as such, reliability was investigated through test-retest procedures only. Test-retest values vary by age group, but fell within the low range ($r \leq .59$) (Strauss et al., 2006).

**FAS – phonemic fluency.** Phonemic fluency was one of three observed measures in the estimation of the latent variable of *Cognitive Flexibility* within the proposed CFA models. There are several different versions of verbal phonemic fluency. Of those, F-A-S is the most commonly used and has been normed on ages 7-95 years (Strauss et al., 2006). Examinees are typically given 60 seconds to produce as many words as possible
that begin with a given letter. The rules of the task also specify that examinees should not provide words that are proper names nor those words that possess the same root but with different endings. Scores are determined by the number of admissible words produced for each letter with the overall total number of words produced across all three trials summed to produce a total score (Strauss et al., 2006). Performance on phonemic fluency may decline with increasing age, after reaching peak performance at age 30 to 39 (Backman et al., 2004; Delis et al., 2001, Mitrushina et al., 2005). Higher levels of education are typically associated with better performance (Backman et al., 2004; Steinberg, Bieliauskas, Smith, & Ivnik, 2005; Kave, 2005). Internal consistency reliability for FAS was evaluated by Tombaugh and colleagues and found to be high ($\alpha = .83$) (Tombaugh, Kozak, Rees, 1999). Test-retest reliability is also high and typically above $r = .70$ (Strauss et al., 2006; Tombaugh et al., 1999).

**Trail Making Test A & B (TMT).** The TMT Part B was one of three observed measures in the estimation of the latent variable of *Cognitive Flexibility* within the proposed CFA models. The Trail Making Test, now within the public domain (Lezak, Howieson, Bigler & Tranel, 2012), is intended to be a measure of attention, speed, and mental flexibility. It was adapted from the “Partington’s Pathways” or “Divided Attention Test” (Partington & Leiter, 1949) within the Army Individual Test Battery (1944) by Reitan and included in the Halstead Battery (1955). The test is normed for adults 15 to 89 years of age. The task requires examinees to connect visual numeric (Trails A) and alphanumeric (Trails B) stimuli by drawing a line from target to target in ascending order. Part A requires the examinee to simply connect numbers, but Part B requires the examinee to alternate between letters and numbers (Strauss et al., 2006).
Administration is timed and scores are measured as the time in seconds to complete the trial. Discontinuation occurs at 180 seconds for Part A and 300 seconds for Part B. Participants in the study were administered both Parts; however, TMT Part B was considered a better measure of cognitive flexibility given the requirement of alternating between stimuli. In some cases, a derived variable taken from the subtraction of Part A from Part B is utilized to describe shifting, inhibition, or flexibility of thinking ((TMT B – TMT A); Christidi, Kararizou, Triantafyllou, Anagnostouli, & Zalonis, 2015)). Therefore, both Parts were administered to allow for a derived variable approach to measurement. Performance on the task declines with age (Backman et al., 2004; Drane, Yuspeh, Huthwaite, & Klinger, 2002; Hester, Kinsella, Ong, & McGregor, 2005; Mitrushina, Boone, Razani, & D’Elia, 2005). Test-retest reliability varies by age group and population, but is mostly adequate across studies for Part B (Strauss et al., 2006). With an 11 month interval in a healthy adult sample, Dikman and colleagues found adequate reliability for Part A ($r = .79$) and high reliability for Part B ($r = .89$) (Dikmen, Heaton, Grant, & Temkin, 1999). These results were similar to those found by Levine and colleagues (Levine, Miller, Becker, Selnes, & Cohen, 2004). Older adult, one year reliabilities were low for Part A ($r = .53 – .64$) and slightly better for Part B ($r = .67 – .72$) (Mitrushina & Satz, 1991; Snow, Tierney, Zorzitto, Fisher, & Reid, 1988). Part A and Part B exhibit a moderate correlation to one another across studies ($r = .31 - .60$), indicating measurement of similar, though distinct abilities (Heilbronner, Henry, Buck, Adams, & Fogle, 1991; Pineda & Merchant, 2003; Royan, Tombaugh, Rees, & Francis, 2004).
Procedure

Older adults expressing interest in study participation were contacted via phone and scheduled for study slots after partaking in the initial phone screen. A brief introduction to the study was provided and questions specific to the exclusion criteria were asked over the phone. Older adults were administered the MIS-T to help ensure that participants fell within the boundaries set for the identification of a cognitively heterogeneous sample (i.e, MIS-T score of 3-8). Those older adults scoring 3 or higher on the measure were scheduled to participate.

All participants were administered consent prior to the start of the study, which was given human subjects approval by the UCCS IRB. The consent process included a brief description of the study with an orientation to the basic procedure of one on one testing, including an explanation of the standardized administration process and absence of feedback, and the risks and benefits of participating.

Study procedures loosely followed those set out by Miyake and colleagues (2000). Following the consent process, participants were administered an hour-long battery of nine neuropsychological tests thought to measure different subcomponents of EF as well as the MoCA, NAB Driving Scenes, and Bill Payment subtests. The tests of the battery were administered to each participant in the same order. Following the administration of all of the tasks, basic demographic data were collected through the demographic questionnaire. Participants were then debriefed about the study’s purpose and compensated for their participation. Each neuropsychological task was scored following the testing administration and raw scores were entered into a participant database without any identifying information. Signed consent forms were kept separate
from raw data. Performance scores on the tests served as the observed variables within the CFA and SEM models. Those individuals who expressed concern about their cognitive functioning were referred to the clinical services offered by the UCCS Aging Center.

**Task Scoring and Data Analysis**

Each of the twelve neuropsychological measures can be scored to produce a performance or total score reflective of overall success on the task. Given the diversity of tasks administered as well as the relative impurity of each task, performance scores utilized unique and individual scoring systems and differing scales of measurement. This resulted in possible score ranges that differed greatly from task to task and which caused unequal variances across tasks. As a result, the hypothesized structural and measurement models utilizing performance scores were standardized. This produced standardized parameter estimates, which facilitated direct comparisons between tests in terms of each test’s contribution to the model. Use of standardized scores was also utilized in the previous research (Hull et al., 2008; Miyake et al., 2000).

Analyses followed the procedure utilized by Miyake et al. (2000) in which the hypothesized measurement models with the EF subcomponents were tested using CFA. Subsequent SEM model testing with the added NAB Driving Scenes and Bill Payment scores as outcome variables was dependent upon CFA results. CFA determined the relationships among nine tasks of executive functioning and tested the hypothesized underlying subcomponents for the tasks within the older adult sample. SEM was intended to determine if and how the subcomponents of executive functioning related to complex neuropsychological task performance, represented through the addition of the
manifest variables on the right side of the model. SEM would evaluate the extent to which each executive functioning subcomponent contributed to the measured behavior of NAB Driving Scenes and Bill Payment performance.

CFA and SEM models were hypothesized a priori according to both previous research (Hull et al., 2008; Miyake et al., 2000) and the anticipated model fit difficulties implied by the use of impure neuropsychological tasks. Of the 192-participant sample, 190 participants had complete data across measures. The incomplete participant cases were due to participant refusal to complete specific measures within the battery. Missing values were not associated with any one particular test. The full sample was retained for the CFA analyses and full information maximum likelihood (fiml) was used to handle the missing values. When appropriate model fit statistics were not attained through any of the hypothesized models, additional exploratory analyses were performed, including exploratory factor analysis and multiple regression. These methods were utilized in an attempt to better identify relationships among variables within this older adult sample and to guide future research.

Statistical analyses were performed in R version 3.0.2 (R Core Team, 2013). Test scores were standardized by the software and used as indicators. The variances of the latent variables were set to 1.0 so that factor loadings could be estimated. A robust maximum likelihood estimator accounted for potential non-normality in the indicators, as a subset of the tests have a restricted range with possible floor or ceiling effects and others are allowed to vary quite significantly. Latent variables were correlated in the initial hypothesized models, but this varied in subsequent models (as described above in Present Study).
Model fit was evaluated using the comparative fit index (CFI), Tucker-Lewis Index (TLI), and the root mean square error of approximation (RMSEA) and its 90% confidence intervals. CFI and TLI values of 0.95 and above are suggestive of good fit; RMSEA values of 0.06 and lower are suggestive of good fit (Hu & Bentler, 1999; Weston & Gore, 2006). There is evidence and argument in the literature that because smaller sample size may influence a model’s ability to fit the data that a less stringent set of fit statistic criteria may be used for samples with less than 500 cases (Marsh, Hau, & Wen, 2004). Therefore, a more lenient set of fit statistics was considered, where necessary, including CFI > .90 and RMSEA < .10 (Weston & Gore, 2006).
CHAPTER 3

RESULTS

Participant Characteristics

The full sample consisted of 192 community-dwelling older adults between the ages of 54-92 ($M = 69.80$, $SD = 8.06$). The sample was primarily Caucasian (92.2%), non-Hispanic (88.5%), female (68.6%), right-handed (85.9%), college educated ($M = 15.84$ years, $SD = 2.25$), and in good health ($M = 2.21$ rating, $SD = 0.88$). The sample was cognitively heterogeneous, as measured through the MoCA, with scores ranging from 12-30 ($M = 26.19$, $SD = 3.08$). Descriptive statistics for the full set of demographic questions are shown in Table 1.

Latent Variable Relationships

Suitability of the data for CFA and exploratory factor analysis (EFA) was assessed through inspection of the correlation matrix, as well as through examination of the Kaiser-Meyer-Oklin value and Bartlett’s Test of Sphericity. Each of the three latent variables within the hypothesized CFA models was estimated through three observed test scores. Three cognitive tests were used to measure the EF subcomponent of planning: ROCF, Clock Drawing Task, and NAB Mazes. Total performance scores from the three measures were not strongly correlated to one another (ROCF to Clock $r = .27$; ROCF to Mazes $r = .10$; Clock to Mazes $r = .17$), suggesting that the variables estimating this factor may not be sufficient for good model estimation (per suggested cut-offs of $r > .30$;
Table 1

*Means, Standard Deviations, and Percentages for Demographic Variables*

<table>
<thead>
<tr>
<th>Demographic factors (n = 192)</th>
<th>M</th>
<th>SD</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>69.80</td>
<td>8.06</td>
<td></td>
</tr>
<tr>
<td>Gender (% female)</td>
<td></td>
<td></td>
<td>68.6</td>
</tr>
<tr>
<td>Race (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>1.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Native Hawaiian or Pacific Islander</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White or Caucasian</td>
<td></td>
<td></td>
<td>92.2</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Ethnicity (% Hispanic)</td>
<td></td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>Education (years)</td>
<td>15.84</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>Handedness (% Right)</td>
<td></td>
<td></td>
<td>85.9</td>
</tr>
<tr>
<td>Self-Rated Health</td>
<td>2.21</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>MoCA</td>
<td>26.19</td>
<td>3.08</td>
<td></td>
</tr>
</tbody>
</table>

Pallant, 2013). The correlation between ROCF and Clock was statistically significant at $p < .01$. The correlation between Clock and Mazes was statistically significant at $p < .05$. ROCF was not significantly correlated to Mazes. Three cognitive tests were used to measure the EF subcomponent of working memory: Digit Span Backward, Mental Control, and NAB Dots. Total performance scores from the three measures were moderately and significantly correlated to one another (Digit Span to Mental Control $r = .33$; Digit Span to Dots $r = .31$; Mental Control to Dots $r = .32$). All three correlations were statistically significant at $p < .01$. Three cognitive tests were used to measure the EF subcomponent of cognitive flexibility: TMT B, D-KEFS Design Fluency, and FAS. Total performance scores from the three measures were moderately and significantly correlated to one another (TMT B to Design Fluency $r = -.49$; TMT B to FAS $r = -.38$; Design Fluency to FAS $r = .35$). All three correlations were statistically significant at $p < .01$; relationships with TMT B were negative to represent the inverse direction of
performance scores, in which higher scores on TMT B indicate worse performance. For all models tested, the Kaiser-Meyer-Oklin value was greater than the recommended value of .6 at .81 (Kaiser 1970, 1974) and Bartlett’s test of Sphericity was statistically significant ($p < .001$), indicating sufficient correlations among the variables to pursue factor extraction (Bartlett, 1954). Summary statistics for each test within the CFA and SEM models are presented in Table 2 and relationships among tests are presented through a correlation matrix in Table 3.

**Measurement Models**

The four hypothesized CFA models were tested and compared to find the best fitting model for describing the underlying latent structure of the subcomponents of EF. The first tested model was a replica of the original Miyake model (2000) with three correlated latent factors with measured variables loading more strongly, though perhaps not independently, onto one of three constructs - planning, working memory, and cognitive flexibility. The second tested model was a bi-factor modification of the original Miyake model in which a single unifying factor was an underlying contributor to all task of EF, while the remaining two factors contributed additional variance in performance score to a subset of the EF tasks. The third tested model was a replica of the older adult model supported by Hull and colleagues (2008). This model, a two-factor CFA, evaluated the presence of planning and updating only, as previous research has indicated that the component of inhibition (or cognitive flexibility here) does not underlie older adult performance. The fourth model tested was hypothesized on the basis of task impurity, in which no clinical measure of EF was expected to accurately represent a single domain of functioning or single subcomponent of EF. This single-factor model of
Table 2

Descriptive Statistics of the Measured Variables Used in the Hypothesized Measurement Models

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>Range</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Latent Factors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Planning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clock Drawing</td>
<td>8.31</td>
<td>2.23</td>
<td>2-10</td>
<td>192</td>
</tr>
<tr>
<td>NAB Mazes</td>
<td>12.27</td>
<td>5.77</td>
<td>1-25</td>
<td>192</td>
</tr>
<tr>
<td>Rey Copy</td>
<td>28.56</td>
<td>5.25</td>
<td>10.5-36</td>
<td>191</td>
</tr>
<tr>
<td><strong>Working Memory</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digits Backward</td>
<td>7.03</td>
<td>2.40</td>
<td>2-13</td>
<td>191</td>
</tr>
<tr>
<td>NAB Dots</td>
<td>5.93</td>
<td>2.50</td>
<td>0-12</td>
<td>191</td>
</tr>
<tr>
<td>WMS Mental Control</td>
<td>25.19</td>
<td>5.28</td>
<td>10-37</td>
<td>192</td>
</tr>
<tr>
<td><strong>Cognitive Flexibility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAS</td>
<td>41.21</td>
<td>13.30</td>
<td>15-92</td>
<td>192</td>
</tr>
<tr>
<td>DKEFS Design Fluency</td>
<td>6.88</td>
<td>2.78</td>
<td>10-46</td>
<td>192</td>
</tr>
<tr>
<td>TMT B</td>
<td>76.03</td>
<td>37.84</td>
<td>24-254</td>
<td>191</td>
</tr>
<tr>
<td><strong>Complex Functional Tasks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAB Driving Scenes</td>
<td>47.98</td>
<td>7.41</td>
<td>27-67</td>
<td>191</td>
</tr>
<tr>
<td>NAB Bill Payment</td>
<td>17.88</td>
<td>1.81</td>
<td>11-19</td>
<td>190</td>
</tr>
</tbody>
</table>

EF suggested that one unifying factor underlies the performance on all clinical tasks of EF. Hypothesized models are depicted in Figures 2-5.

The three-factor Miyake model produced a covariance matrix that was not positive definite, making the results un-interpretable. Examination of the covariance matrix and the correlations among the factors showed that the three latent factors were too highly correlated to one another. Estimated correlations between each pair of factors was greater than 1.00 (Planning to Working Memory $r = 1.02$; Planning to Cognitive Flexibility $r = 1.01$; Working Memory to Cognitive Flexibility $r = 1.10$). These results provided support for the continued testing of the other hypothesized models, as well as the need for fewer latent factors.
### Table 3

**Correlation Matrix of Neuropsychological Tests' Total Performance Scores**

<table>
<thead>
<tr>
<th>Test</th>
<th>Clock</th>
<th>Mazes</th>
<th>Rey</th>
<th>Digits</th>
<th>Dots</th>
<th>Mental Control</th>
<th>FAS</th>
<th>Design Fluency</th>
<th>TMT B</th>
<th>Driving Scenes</th>
<th>Bill Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock</td>
<td>1.00</td>
<td>.17*</td>
<td>.27**</td>
<td>.18*</td>
<td>.15*</td>
<td>.10</td>
<td>.14</td>
<td>.14</td>
<td>-.22**</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>Mazes</td>
<td>--</td>
<td>1.00</td>
<td>.10</td>
<td>.22**</td>
<td>.46**</td>
<td>.32**</td>
<td>.25**</td>
<td>.46**</td>
<td>-.42**</td>
<td>.33**</td>
<td>.18*</td>
</tr>
<tr>
<td>Rey</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>.20**</td>
<td>.27**</td>
<td>.10</td>
<td>.12</td>
<td>.02</td>
<td>-.31**</td>
<td>.07</td>
<td>.14</td>
</tr>
<tr>
<td>Digits</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>.31**</td>
<td>.33**</td>
<td>.42**</td>
<td>.13</td>
<td>-.25**</td>
<td>.15*</td>
<td>.11</td>
</tr>
<tr>
<td>Dots</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>.32**</td>
<td>.27**</td>
<td>.40**</td>
<td>-.51**</td>
<td>.40**</td>
<td>.28**</td>
</tr>
<tr>
<td>Mental Control</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>.49**</td>
<td>.51**</td>
<td>-.52**</td>
<td>.30**</td>
<td>.21**</td>
</tr>
<tr>
<td>FAS</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>.35**</td>
<td>-.38**</td>
<td>.21**</td>
<td>.10</td>
</tr>
<tr>
<td>Design Fluency</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>-.49**</td>
<td>.40**</td>
<td>.15*</td>
</tr>
<tr>
<td>TMT B</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>--</td>
<td>.46**</td>
<td>.31**</td>
</tr>
<tr>
<td>Driving Scenes</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>.60**</td>
<td>--</td>
</tr>
<tr>
<td>Bill Payment</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>--</td>
</tr>
</tbody>
</table>

*Note: * < .05; ** < .01*
Figure 2: Hypothesized three-factor measurement model with three correlated latent factors based on the original Miyake model.
Figure 3. Hypothesized bi-factor measurement model with three uncorrelated latent factors based on the Miyake modification model.
Figure 4: Hypothesized two-factor measurement model with correlated latent factors based on the Hull model.
Figure 5. Hypothesized one-factor measurement model representing a unifying global factor for executive functioning.
The modified three-factor model, a bi-factor model with three uncorrelated factors, consisted of a single unifying factor underlying all EF tasks and two other independent factors that contributed to a distinct subset of the EF tasks. This bi-factor model also produced a not positive definite variance matrix, making the results uninterpretable. These results also provided support for the continued testing of the other hypothesized models; however, further exploration of the data was performed to evaluate the underlying reasons for the not positive definite errors and to ensure data fidelity. These methods included double-scoring of the test variables and double-checking of data entry, visual inspection of the individual variables and all descriptive statistics, and running the model on a random half-split of the overall sample. As these exploratory methods did not change the results nor reveal significant issues in the data, the subsequent hypothesized models were tested.

The Hull replica model, a two-factor CFA with two correlated factors, planning and updating, resulted in an interpretable model. The two factors were strongly related to one another, \( r = .996 \), suggesting the possible presence of a superior one-factor model. The Hull replica model provided poor model fit \( (\chi^2 = 96.03, \ df = 26, \ p < .001; \ CFI = .84; \ TLI = .78; \ RMSEA = .12, \ 90\% \ CI [0.09, \ 0.14], \ p < .001) \). Examination of the latent factors revealed that all of the factor loadings were statistically significant at or below \( p = .01 \) (see Figure 6). For the updating factor, all standardized factor loadings were .44 or higher, providing further evidence of a well-estimated latent factor. In contrast, the planning factor demonstrated weaker standardized factor loadings, with both Clock and Rey falling below .350.
Figure 6. Two-factor CFA solution based on the Hull model, with standardized estimates and latent factor variances set to 1.00.
The one-factor model produced nearly identical model fit statistics to the two-factor model ($\chi^2 = 96.04, df = 27, p < .001; \text{CFI} = .84; \text{TLI} = .79; \text{RMSEA} = .12, 90\% \text{CI [}.09, .14], p < .001$). Examination of the relationships between neuropsychological test variables and the single latent factor revealed that the neuropsychological test variables were statistically significant estimators of the common EF factor (see Figure 7). Pre-determined model fit criteria were not met by any of the hypothesized models, as a result, SEM predicting performance on the two proxy measures of ecologically valid abilities was not pursued with any of the hypothesized versions of the measurement models. Instead, an exploratory factor analysis approach to the measurement model was taken to examine other possible structural models that were not hypothesized in advance.

**Exploratory Factor Analysis.** None of the hypothesized models produced model fit statistics that met the pre-determined guidelines for identifying a sufficient model of the underlying subcomponents of EF (Hu & Bentler, 1999; Weston & Gore, 2006). Given the insufficient model fit of each CFA, it appears that the expectations of the fourth hypothesis are supported. In other words, total scores may be poor representations of any one particular aspect of functioning and the proposed models may not capture the true underlying structure of the tests. Therefore, exploratory factor analysis was pursued to identify the best fitting structural model for the nine tests of EF.

While there are many different possible techniques used for factor extraction, an exhaustive list of the results of each technique will not be provided here. Each method of factor extraction utilized produced similar results and provided some evidence for the unity and diversity of the performance scores, with most evidence supporting the presence of a one-factor model and some evidence supporting the presence of a two- or
Figure 7. One-factor CFA solution with a global executive functioning factor depicted with standardized estimates.
three-factor model. Principal axis factoring with Oblimin rotation produced evidence for a one-factor model that explained 38.03% of the variance. Inspection of the screeplot (Figure 8) revealed a clear break in the data after the first factor, which was further supported by the magnitude of the corresponding eigen values. However, two less prominent breaks occurred after the second and third factors, respectively. Examination of the factor loadings revealed that a three-factor model, for which there was minimal support (eigen value greater than 1), resulted in one test which strongly cross-loaded onto two-factors (standardized factor loadings greater than .3 on two factors) as well as several negative factor loadings, evidence of a not positive definite matrix. Forcing a one-factor model produced sufficient factor loadings for all but two of the tests, Rey Figure and Clock (both lower than at < .3) and removed the negative factor loadings for measures other than the TMT B (which is reversed scored and therefore predicted to produce a negative relationship to the factor). Factor loadings for both the three and one factor EFA models are depicted in Table 4.

Table 4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKEFS Design Fluency</td>
<td>.757</td>
<td>-.216</td>
<td>.063</td>
<td>.615</td>
</tr>
<tr>
<td>TMT B</td>
<td>-.633</td>
<td>-.263</td>
<td>-.082</td>
<td>-.773</td>
</tr>
<tr>
<td>NAB Mazes</td>
<td>.619</td>
<td>.027</td>
<td>-.036</td>
<td>.553</td>
</tr>
<tr>
<td>NAB Dots</td>
<td>.589</td>
<td>.236</td>
<td>-.008</td>
<td>.645</td>
</tr>
<tr>
<td>WMS Mental Control</td>
<td>.413</td>
<td>-.112</td>
<td>.404</td>
<td>.642</td>
</tr>
<tr>
<td>Rey Copy</td>
<td>.061</td>
<td>.607</td>
<td>-.032</td>
<td>.281</td>
</tr>
<tr>
<td>Clock Drawing</td>
<td>.004</td>
<td>.400</td>
<td>.150</td>
<td>.292</td>
</tr>
<tr>
<td>FAS</td>
<td>.031</td>
<td>-.052</td>
<td>.763</td>
<td>.552</td>
</tr>
<tr>
<td>Digits Backward</td>
<td>-.021</td>
<td>.180</td>
<td>.523</td>
<td>.452</td>
</tr>
</tbody>
</table>
Figure 8: Scree plot from the exploratory factor analysis using Principal Axis Factoring and Oblimin rotation.
Models of Prediction

In the absence of a good fitting measurement model, there was no evidence that one of the hypothesized measurement models would serve as an effective foundation for the structural model. Given the insufficient fit of both the hypothesized and exploratory measurement models, an extension of those into the hypothesized structural model did not seem warranted. Instead, multiple regression was utilized to evaluate separate predictive models for each of the two complex tasks serving as proxy measures for ecological abilities, Driving Scenes and Bill Payment. Prediction models were supported by a combination of previous research findings and examination of the correlation matrix of all the cognitive tasks. Preliminary analyses ensured that there were no violations of the assumptions of normality, linearity, multicollinearity, and homoscedasticity. Evaluation of Mahalanobis distance revealed that two cases, participants 906 and 1080, were possible outliers in the data; however, their values for Cook’s Distance suggested that neither was exerting an undue influence on the statistic. In addition, neither case was noted to have any unusual circumstances or behaviors on the day of testing; neither case had evidence of significant departure from the required inclusion criteria. As a result, neither case was removed from the sample for analysis.

Driving Scenes. Previous research has suggested that both the Clock and TMT B are among the strongest neuropsychological predictors of on-road driving behavior (Asimakopulos et al., 2012). Of these two tests, however, only TMT B was correlated with Driving Scenes. As a result, TMT B was hypothesized to be the strongest predictor of Driving Scenes performance. In addition, the format and administration of NAB Dots is similar in format and administration to NAB Driving Scenes and both tasks are taken
from the Attention module of the NAB. As such, Dots and Driving Scenes were
designed together and with construct validity in mind (Stern & White, 2003, 2003b).
Standard multiple regression was used to assess the ability of the nine tests of EF to
predict performance on a proxy measure of on-road driving behavior, Driving Scenes.
The total variance explained by the model was 29.9%, $F(9, 176) = 8.34, p < .001$.
Several of the EF test scores were statistically significant predictors of Driving Scenes
performance, including TMT B (beta = -.30, $p = .001$), DKEFS Design Fluency (beta = 
.17, $p = .04$), and NAB Dots (beta = .18, $p = .02$), representing the hypothesized factors of
working memory and cognitive flexibility. The most predictive task of Driving Scenes
performance, TMT B, only accounted for approximately 4.5% of the variance, providing
evidence that much of the variance in Driving Scenes performance cannot be explained
through EF tests alone.

**Bill Payment.** While the executive functions, broadly, are the most cited
predictors of financial literacy and management abilities within an older adult cohort,
efforts to isolate more specific aspects of EF have identified the abilities of selective
attention, self-monitoring, and integration of temporally sequenced information
(Okonkwo et al., 2006) as necessary for prediction. Previous research has suggested that
visuomotor sequencing and set-alternation were predictive of financial management
abilities within an older adult cohort, regardless of disease status: normal, MCI, or
dementia (Sherod et al., 2009). As a result of these findings, TMT B, Digit Span, and
Dots were hypothesized to be the best predictors of Bill Payment performance. Standard
multiple regression was used to assess the ability of the nine tests of EF to predict
performance on a proxy measure of financial management and capacity, Bill Payment.
The total variance explained by the model was 12.6%, $F(9, 180) = 2.89, p = .003$. Only one of the EF test scores was a statistically significant predictor of Bill Payment performance, TMT B (beta = -.23, $p = .03$). As the most predictive task of Bill Payment performance, TMT B only accounted for approximately 2.4% of the variance, providing evidence that much of the variance cannot be explained through EF tests alone.
CHAPTER 4
DISCUSSION

The present study evaluated the underlying factor structure of nine commonly
used neuropsychological tests of EF in a sample of 192 cognitively heterogeneous older
adults from the Colorado Springs community. The study sought to improve the
understanding of the ill-defined construct of EF (Stern et al., 2011) by validating existing
models of construct validity (Hull et al., 2008; Miyake et al., 2000, Miyake & Friedman,
2012) with clinical neuropsychological tasks of EF and then deconstructing those existing
models to propose an improved approach to the measurement of EF. Previous research
from cognitive neuroscience (Friedman et al., 2011; Mischel et al., 2011; Moffit et al.,
2011) has provided evidence that the subcomponents of EF are strongly related constructs
that contribute differentially to complex tasks and behaviors, depending upon the
demands of the task as well as the sample tested. While previous research provided
evidence for a model of functioning best described by unity and diversity (Miyake et al.,
2000, Miyake & Friedman, 2012), extension of this model to older adult populations
revealed that diversity, more than unity, prevailed (Hull et al., 2008). The distillation of
EF subcomponents was made possible through the use of computerized measures of EF
that could capture, sometimes with millisecond accuracy, a clean measurement of
performance. While this body of research serves as the foundation for studies of
construct validity in the executive functions, it only contributes minimally to the clinical
understanding of these constructs. These computerized tests, the limited number of subcomponents evaluated, and the laboratory settings in which tasks were administered to participants bear only minimal resemblance to the wide array of neuropsychological measures most commonly used to evaluate EF within a clinical evaluation. This limits the generalizability of these results within clinical settings. The present study tested the previously established models of construct validity by utilizing confirmatory factor analysis on a set of commonly used clinical measures of EF that represented the constructs of planning, working memory, and cognitive flexibility.

Three previously supported models of the underlying structure of EF were evaluated, two of which were supported in a younger adult only sample (Miyake et al., 2000, Miyake & Friedman, 2012), and one of which was supported in an older adult only sample (Hull et al., 2008), along with a fourth hypothesized model. All four of these models were tested with the current sample of cognitively heterogeneous older adults and none of the models provided adequate fit. As hypothesized, the models derived from computerized measures of EF did not hold within a sample of performance scores taken from clinical neuropsychological measures of EF. Instead, values from the tested models appeared to suggest that total performance scores taken from clinical measures, which reflect the issues of both poor EF construct validity and poor measurement, could not be used to differentiate subcomponents from one another. The tasks themselves, and their total score values of performance, likely represent multiple shared underlying subcomponents (and other domains of functioning) that clouded the distinction among factors. In fact, the two models attempting to replicate the findings of Miyake and colleagues contained factors that were too highly correlated with one another to be fully
interpreted, suggesting unity more than diversity. This was a somewhat expected finding, given the nature of impure clinical tasks; however, it stands in contrast to the established finding that older adults demonstrate diversity of subcomponents more than unity (Hull et al., 2008). Thus, the two remaining hypothesized models anticipated these two possibilities, a two-factor model of diversity, replicating the findings of Hull and colleagues (2008), and a one-factor model of unity, representing the impurity of neuropsychological tasks and the all-encompassing nature of EF. The two- and one-factor models produced nearly identical model fit statistics that were better than the fit statistics for the first two tested models, but still not adequate according to the predetermined cut-offs. Given that none of the hypothesized models provided adequate fit, exploratory factor analysis (EFA) was utilized to find a superior-fitting model. Most of the evidence from the EFA suggested a one-factor model, though there was some evidence for both a two- and three-factor model of EF from evaluation of the eigen values, scree plot, and arrangement of test scores onto factors. Three key findings emerge from this series of analyses and each is supportive of the study’s original hypotheses: 1) Existing models of EF subcomponents do not hold for the clinical neuropsychological measures currently used to estimate these abilities; 2) With regard to clinical measures of EF, CFA and EFA provide evidence for unity and diversity among the executive functions and within a single sample of participants; and 3) The present models are incomplete and cannot account for a large proportion of the variance in performance scores, given the task impurity of these measures.

Total performance scores taken from each of the nine measures of EF do not map onto distinct subcomponents of EF. In fact, even their ability to map onto a global EF
factor is somewhat limited, as one-factor models did not significantly improve the model fit. Neuropsychological measures utilized in this study were selected, in part, on the basis of frequent clinician use (Rabin et al., 2005) and research evidence of ecological validity. It is therefore unlikely that the present results are an artifact of the specific tests selected. Rather, the measures selected more likely represent an issue inherent to most, if not all, clinical measures of EF. These results call into question the current practice of test-selection for the explicit purpose of evaluating specific aspects of EF. The results appear to indicate that the best estimate of functioning that can be gleaned from overall performance scores is an estimate of global EF ability and that even this estimate is somewhat mired by the contribution of other domains of functioning (e.g., visuospatial processing, language, memory). The clinical practice of evaluating specific components of EF through test performance on one or even up to three measures drastically misrepresents the actual findings from testing and likely overestimates the possible ecological implications of said performances.

From the present study, there is evidence that measures of EF provide an interpretation of global EF ability and that they may therefore indicate the presence or absence of decline or impairment when interpreted together. The present study also strongly suggests that interpretation of a single, low, test score cannot provide accurate information about the specific nature of that given low score. Instead, total performance scores are providing information about both local and global patterns of functioning and how the two are working together. In other words, a total score value of EF has multiple underlying contributors, including multiple subcomponents of EF and one to several other domains of functioning, as well as a more unifying or global measure of how these
contributors work together. This helps to illustrate why EF as a domain has been described through various levels of integration, as a set of higher-order functions, as a supervisory system to other domains of functioning in the brain, and as a hierarchical system of abilities which build upon each other or fit together to ensure task completion and goal-directed behavior (Barkley, 2012; Stern et al., 2011; Strauss et al., 2006). In the absence of better, more pure measures of these subcomponent abilities, a different approach to the evaluation of test performance and utilization of scores is necessary.

The present study also sought to find an accurate predictive model of ecological task performance by utilizing two additional neuropsychological tests of more complex abilities within structural equation models that could represent an extension of the best-fitting model of EF construct validity. Unfortunately, a model of EF subcomponents with sufficient model fit could not be identified and SEM was therefore not pursued. Instead, multiple regression was used to find the measures of EF that best predicted performance on these proxy measures for real world behavior, Driving Scenes and Bill Payment. Results of the multiple regression demonstrated that TMT B, DKEFS Design Fluency, and NAB Dots all predicted performance on the Driving Scenes subtest, with TMT B accounting for the largest proportion of the variance, 4.5%. Based on initially hypothesized placement of test scores onto latent factors, this would suggest that a combination of cognitive flexibility and working memory are required for successful performance on the Driving Scenes test, which is likely true; however, most of the variance in performance was unaccounted for and results of the CFA models suggested that conclusions about subcomponents should not be drawn from test performance on these measures. Instead, it is likely that TMT B represents the best predictor of local and
global contributors necessary for successful Driving Scenes performance. Evaluation of task descriptions supports this idea. Both the TMT B and Driving Scenes tasks require visual attention and scanning, the ability to hold onto task instructions and maintain set, intact psychomotor function, and efficient processing speed (Lezak et al., 2012; Strauss et al., 2006; White & Stern, 2003). The integration of those abilities is also required for both.

For the Bill Payment test, results of the multiple regression demonstrated that TMT B was the only predictor of performance and that it accounted for 2.4% of the variance of the Bill Payment score. This result speaks strongly to the inability of current EF tests to predict ecological behavior. Bill Payment is a highly structured proxy for the real-world behaviors of writing a check, balancing a checkbook, and correctly addressing an envelope for the sending of payment to the appropriate location, all based on the accurate reading of a bill (Stern & White, 2003). For those without impairment, the task is simple to complete. Visual inspection of the data revealed that performance scores on the task were negatively skewed with a ceiling effect occurring at the maximum end of the test’s restricted range (0-19). However, more than half of the participants received a score of 18 or lower, with most participants falling between the range of 15-18 on the task with a full range of scores from 11-19 for the overall sample, suggesting adequate variability of performance for prediction. The fact that only one test was able to predict this performance is both disappointing and illustrative of the problems plaguing EF’s and neuropsychology’s struggle in predicting ecological behavior. Evidence from both predictive regression models suggests that TMT B is a sensitive measure of complex task performance. Unfortunately, more specific conclusions about the underlying
subcomponents of this task and its relationship to performance on other measures cannot be determined. Its common use in clinical practice is likely the result of its wide reaching sensitivity (Rabin et al., 2005). The evidence from all analyses in the present study points to the same conclusion: the current methods for measuring EF are extremely limited in their ability to draw valid conclusions about subcomponents of functioning or to make accurate predictions about behavior.

**Implications**

As a result of the present findings, this study strongly supports further research into the clinical measurement of EF. The study findings are consistent with the overwhelming amount of dispute within the existing research literature about both the measurement and definition of the domain of EF (Mueller & Dollaghan, 2013; Stern et al., 2011; Strauss et al. 2006). However, this study provides evidence for the conceptual and measurement problems inherent to EF that are frequently alluded to, though not addressed, in clinical practice. Through this evidence, an alternative approach is suggested. Clinical neuropsychology’s assumptions about and practices around EF tests may be inaccurate and should be more stringently addressed in clinical practice and research.

**Implications for clinical practice.** Given the evidence for task impurity and the associated problems with making attributions to specific aspects of functioning, it is important to include more precise measures of ability and to very cautiously make interpretations about functioning. In the absence of new, better tasks, more precise measurement must be attained through new and different variables that can extract specific aspects of functioning. Error analysis and process approach scoring techniques
may allow for a more fine-tuned interpretation of behavior and performance (Ashendorf et al., 2013); however, not all tests are designed with these scores in mind and there is little evidence linking specific variables to specific subcomponents, particularly in the early phases of disease development or across the developmental lifespan. Many skilled clinicians will draw conclusions about patients by interpreting patterns of error across a battery of tests, but these interpretations tend to be done informally and on the basis of clinical experience (i.e., having seen the same patterns of error consistently within a given patient population). This practice may be useful for confirming diagnostic hypotheses that have well-established patterns of error, just as, for example, the presence of false positive errors, intrusions, and repetitions provide support for the diagnosis of Alzheimer’s disease (Karantzoulis & Galvin, 2011); however, clinical judgments are routinely outperformed by actuarial predictions and suffer from a greater degree of error (Dawes, Faust, & Meehl, 1989). As a result, the field of clinical neuropsychology is presently focused on the pre-clinical, early identification of disease, an area currently dominated by biomarker and other medical approaches (McKhann et al., 2011). However, neuropsychology has an opportunity to contribute more to this goal if it embraces a different approach and focuses on the development of evidence-based testing variables that represent derived scores and more specific aspects of functioning.

**Implications for research.** From the nine-test EF battery in the present study, over 134 different variables were scored and extracted and these likely represent a cleaner picture of construct validity in the executive functions. The next step in this line of research is to investigate these more detailed variables as a means of discovering construct validity, which, “must be investigated whenever no criterion or universe of
content is accepted as entirely adequate to define the quality to be measured.” (Cronbach & Meehl, 1955, pp. 282). Measures of reaction time or processing speed are the most likely measures to achieve reliability and many computerized tasks utilize this approach to measurement (Miyake & Friedman, 2012); however, this is a limited view of a given ability and seems to unfairly assess the full picture of older adult functioning, as some older adults may possess the necessary subcomponent abilities for completing a given task, but may not be able to complete such a measure within a restricted time frame. As the field moves toward computerized and even remote testing of cognition, the clinician is losing a host of information that may be necessary for truly understanding patient functioning. Computerized tasks must therefore be developed to account for more precise aspects of functioning, and unfortunately, there is a dearth of evidence guiding the selection of variables in computerized task development (Jansari, Devlin, Agnew Akesson, Murphy, & Leadbetter, 2014; Parsons, Carlew, Magtoto, & Stonecipher, 2015).

Limitations

The present study’s most important finding is the presence of multiple limitations in the current approach to EF measurement. The design of the study captured these limitations and provided concrete evidence for their existence by utilizing a model-testing approach that dissected the current prevailing models in the literature. As evidence from all of the analyses suggests, the present study did not find an adequate model of construct validity for EF. Total scores of performance are impure measures of subcomponents and likely include contributions from both hierarchical aspects of functioning, as well as additional domains of functioning. Therefore, too many of the possible contributors to each performance score were not directly accounted for by the models tested. The
present models evaluated were not powered enough to consider all of the possible underlying contributors to performance. Adding additional latent factors to the models, such as visuospatial processing to account for the visual aspects of Rey, Clock, TMT B, Dots, etc. and processing speed to account for the timed aspects of TMT B, Mazes, FAS, Design Fluency, etc. would have captured additional explanatory variance in performance scores, but would have decreased model fit with the current sample of 192 participants. The target sample of 170 participants was based on values from the published literature, the best available information for study planning. This value was met and exceeded, but unfortunately, could not identify a sufficient model. As a result, a truly comprehensive measure of all underlying contributors to performance is beyond the scope of this project and institution. Large-scale publicly available databases might provide the sample size necessary for a comprehensive evaluation of functioning; however, most of these do not include comprehensive batteries of EF tests or the level of detail in variable extraction that this project provided. In addition, it is possible that the true model of construct validity in EF is best represented through a hierarchical model and this would be particularly true for modeling the unity aspect of the domain. However, the EFA techniques utilized here do not identify hierarchical models. It is therefore possible that exploratory methods were limited in their ability to find an adequate model of construct validity.

Other limitations may exist as a result of variable characteristics and influences. Additional variables may have been unaccounted for in the model, including the effects of continuous age and/or other demographic characteristics. However, these are not expected to greatly influence model fit. Demographically corrected test scores were not
used, as the model was intended to represent a heterogeneous sample of older adults with a wide-range of characteristics and levels of functioning. In addition, performance scores were not intended for the categorization of participants into diagnostic or qualitative groups for comparison. Different normative criteria exist for each of the measures utilized and a single or uniform approach would not have been possible across tests. Instead, given that scores possessed widely varying possible ranges and as a result, discrepant variances in performance score, raw scores were utilized in the models and standardized for clarity in the depicted figures.

Finally, scoring on the included measures, as with all clinical measures of EF, required some subjective judgments and may therefore have differed by examiner, creating inconsistencies in the measurement of certain abilities. To prevent this from occurring, each examiner received training in administration and scoring from the same, primary researcher. Before interacting with participants, each examiner administered the entire battery to the primary researcher and had scoring procedures double-checked. A randomly selected subset of participant data was then subsequently double-scored by the primary researcher to maintain procedure fidelity and identify any discrepancies. Despite these policies, the average scores on several measures differed by examiner and that may be a result of differences in testing and scoring procedures that went undetected. It should be noted, however, that examiners tested somewhat different subsets of the sample (i.e., some tested only those recruited from the Gero Registry, while others tested off-site at locations and organizations around the city that may have represented differences in levels of functioning, impairment, and other unmeasured participant characteristics), and that this too likely influenced participant score averages by examiner. In addition,
despite the highly standardized and structured nature of test administration, there are likely to be distinct differences in administration and scoring procedures used clinically across the country, and in that respect, the present study is a representation of yet another aspect inherent to the larger field of neuropsychology (Gwet, 2014).

**Future Directions**

Additional variables beyond the total performance scores were measured and derived according to a few factors, such as research literature supporting their efficacy (Ashendorf et al., 2013), research literature supporting their weaknesses or limited use in the field, clinician intuition, and the need for uniformity across measures. Consideration of these factors was taken at the outset of the present study. As a result, 134 different variables measuring performance can be scored and extracted from the present set of tests and these will be used in future research addressed at identifying better models for subcomponents of functioning.

This approach is somewhat driven by theory, but is far more exploratory than the present study, which had the benefit of existing published models to guide the research methodology. The goal of this larger body of research is to identify and provide evidence for a set of variables that can be readily calculated and used within clinical practice to improve the interpretation of test performance and inform treatment recommendations. There is evidence that subtle, early changes to cognition may occur in middle and late adulthood, but that these changes can go unnoticed during the very early phases of disease development (Brandt et al., 2009). Executive functioning should be the domain that would best capture those changes, given its supervisory and hierarchical reach across other domains of functioning. In addition, evaluation of error patterns at early phases of
disease may help to predict later manifestations of disease (Brandt et al.), which could then be used for diagnosis and treatment planning. If a model of EF subcomponents can be identified through process-oriented variables, then the longitudinal stability or instability of this model might also provide critical information about disease trajectories. And while there is benefit to the development of more sensitive measures of cognition, the identification and validation of new variables that can be derived from existing tests would allow for greater stability of measurement across time, strengthening current longitudinal studies and providing some level of methodological fidelity across different longitudinal projects (Gross et al., 2015).

There is a move toward the use of longitudinal latent variable modeling approaches in the measurement of cognition and for these methods to incorporate multidisciplinary factors represented by a variety of biomarker and neuropsychological evidence (Gross et al., 2015). Discrete measures of specific abilities in neuropsychology would help to clarify discrete measurements of change in volume and metabolism within specific brain areas, for example. This combined neuroscience and neuropsychology approach would ensure the continued relevance of the field and would better identify patients in an at-risk state, far earlier than the current methods. The present study serves as the foundation for such future projects and provides evidence for continued research in the construct and ecological validity of EF that combines approaches from neuroscience and neuropsychology, challenges current clinical practices, and identifies techniques for improved measurement.


APPENDIX

University of Colorado
Colorado Springs

Institutional Review Board (IRB) for the Protection of Human Subjects

Date: 12/19/2015

IRB Review

IRB PROTOCOL NO.: 15-098
Protocol Title: Measuring Executive Functioning in Older Adults
Principal Investigator: Samantha John
Faculty Advisor if Applicable: Brandon Gavett
Application: Renewal
Type of Review: Expedited 7
Risk Level: No more than Minimal Risk
Renewal Review Level (If changed from original approval) if Applicable: N/A No Change
This Protocol involves a Vulnerable Population: N/A (No Vulnerable Population)
Expires: 5 January 2017

*Note: if exempt: If there are no major changes in the research, protocol does not require review on a continuing basis by the IRB. In addition, the protocol may match more than one review category not listed.

Externally funded: ☒ No ☐ Yes
OSP #: Sponsor:

Thank you for submitting your Request for IRB Review for renewal of an approved protocol. The protocol identified above has been reviewed according to the policies of this institution and the provisions of applicable federal regulations. The review category is noted above, along with the expiration date, if applicable.

Once human participant research has been approved, it is the Principal Investigator’s (PI) responsibility to report any changes in research activity related to the project:

- The PI must provide the IRB with all protocol and consent form amendments and revisions.
- The PI must approve the IRB changes prior to implementation.
- All advertisements recruiting study subjects must also receive prior approval by the IRB.
- The PI must promptly inform the IRB of any anticipated serious adverse (within 24 hours). All anticipated adverse events must be reported to the IRB within 1 week (see 45CFR46.103b). Failure to comply with these federally mandated responsibilities may result in suspension or termination of the project.
- Renew study with the IRB prior to expiration.
- Notify the IRB when the study is complete

If you have any questions, please contact Research Compliance Specialist in the Office of Sponsored Programs at 719-255-3903 or irb@uccs.edu

Thank you for your concern about human subject protection issues, and good luck with your research.

Sincerely yours,

Melissa J. Benton
Melissa Benton, PhD
IRB Committee Member