Investigating the Convergent, Discriminant, and Predictive Validity of the Mental Toughness Situational Judgment Test

Nicholas Martin Flannery

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Psychology

Neil M. A. Hauenstein, Co-Chair
E. Scott Geller, Co-Chair
Ivan Hernandez, Committee Member
Tina Savla, Committee Member

04/30/2020
Blacksburg, Virginia

Keywords: Mental Toughness, machine learning, job performance

Investigating the Convergent, Discriminant, and Predictive Validity of the Mental Toughness Situational Judgment Test

Nicholas M. Flannery

ACADEMIC ABSTRACT

This study investigated the validity of scores of a workplace-based measure of mental toughness, the Mental Toughness Situational Judgment Test (MTSJT). The goal of the study was to determine if MTSJT scores predicted supervisor ratings 1) differentially compared to other measures of mental toughness, grit, and resilience, and 2) incrementally beyond cognitive ability and conscientiousness. Further, two machine learning algorithms – elastic nets and random forests – were used to model predictions at both the item and scale level. MTJST scores provided the most accurate predictions overall when model at the item level via a random forest approach. The MTSJT was the only measure to consistently provide incremental validity when predicting supervisor ratings. The results further emphasize the growing importance of both mental toughness and machine learning algorithms to industrial/organizational psychologists.

Investigating the Convergent, Discriminant, and Predictive Validity of the Mental Toughness Situational Judgment Test

Nicholas M. Flannery

GENERAL AUDIENCE ABSTRACT

The study investigated whether the Mental Toughness Situational Judgment Test (MTSJT)– a measure of mental toughness directly in the workplace, could predict employees' supervisor ratings. Further, the study aimed to understand if the MTSJT was a better predictor than other measures of mental toughness, grit, resilience, intelligence, and conscientiousness. The study used machine learning algorithms to generate predictive models using both question-level scores and scale-level scores. The results suggested that the MTSJT scores predicted supervisor ratings at both the question and scale level using a machine learning model. Further, the MTJST was a better predictor than most other measures included in the study. The results emphasize the growing importance of both mental toughness and machine learning algorithms to industrial/organizational psychologists.

Acknowledgments

Although I could easily make this the longest section of my dissertation, I would like to express my sincerest appreciation for the effort and support of the following individuals: Virginia Tech faculty members Dr. E. Scott Geller, Dr. Neil Hauenstein, Dr. Roseanne Foti, Dr. Ivan Hernandez, and Dr. Tina Savla; College of Wooster faculty members Dr. Mike Casey, Dr. Bryan Karazsia, and Dr. Gary Gillund; coaches Marcus Colvin, Brad Kassner, and Zach Dennis; State Farm's Organizational Insights Team; Virginia Tech's Office of Institutional Effectiveness, especially Dr. Bethany Bodo; Virginia Tech Psychology's support staff, especially Michelle Wooddell; close friends Daniel Perez, Holli Kossoudji, Brian Beall, Sean Vahldiek, Adam Coppock, Zach Moore, Joe Superick, Paul Kelbly, Mackenzie Kelbly, Dan Misinay, Derek Burns, Zach Mastrich, Jess Gladfelter, Shelby Borowski, Devin Carter, Emma Condy, Kristen Tetreault, Jefferson Salan, Bryan Acton, Brad DeVore, Jaclyn DeVore, Molly Minnen, Alex Peebles, Vivian Zagarese, Jack Wardale, and Emily Rost, among many, many others; and finally (but most importantly) my family – especially Mom, Dad, Cait, and Jake. This accomplishment would not be possible without the help and support of these individuals.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mental toughness (MT) has gained substantial popularity over the last two decades as a predictor of success in many domains, including athletics (Gucciardi & Gordon, 2011), the military (Arthur, Fitzwater, Hardy, Beattie, & Bell, 2015), and the classroom (Lin, Mutz, Clough, & Papageorgiou, 2017). MT is broadly defined as "the personal capacity to produce consistently high levels of subjective (e.g., personal goals or strivings) or objective performance (e.g., sales, race time, GPA) despite everyday challenges and stressors as well as significant adversities" (Gucciardi, Hanton, Gordon, Mallett, & Temby, 2015, p. 28). Researchers generally agree that MT facilitates adaptive mechanisms to cope with stressors and prevent distress (Gucciardi, et al., 2015).

While the foundational research of MT examined the construct as a predictor of success in athletics (e.g., Gucciardi & Gordon, 2011; Loehr, 1986), recent research has applied MT as a predictor of workplace outcomes (Lin et al., 2017). For instance, Gucciardi et al. (2015)

demonstrated that MT scores predicted supervisor ratings of employees. Further, they found that perceived distress and the use of adaptive coping mechanisms mediated this relationship. Subsequent research demonstrated that MT scores correlated negatively with perceived distress among police officers and firefighters (Ward, St. Clair-Thompson, & Postlethwaite, 2018).

Additional researchers found MT to predict income (Lin, Clough, Welch, & Papageorgiou, 2017) and ascension within an organizational hierarchy (Marchant et al., 2009). These results suggest MT is a promising construct for predicting job performance, but beyond the general conceptualization, researchers do not agree on the underlying dimensions of MT, raising challenges to the construct validity of MT measures (Gucciardi & Gordon, 2011; Gucciardi et al., 2015; Lin et al., 2017). As a result, Flannery, Hauenstein, & Geller (2019) introduced the Mental Toughness Situational Judgment Test (MTSJT) to improve the measurement of MT by assessing behavioral expectations in work situations. In their initial validation effort, the authors recovered a three-factor structure consisting of task persistence, emotional control, and utilization of feedback. Further, they observed modest correlations among MTSJT scores and measures of MT, perceived stress, and the Big Five, likely as a result of the different response formats, thereby suggesting the MTSJT captures unique variance in work-related outcomes.

Given these findings, more research is needed to investigate the construct validity of MTSJT scores. To that end, the first aim of the current study was to investigate the degree of convergence among several measures of MT and similar constructs, including grit and resilience. Secondly, the current study estimated the predictive accuracy of MTSJT scores relative to other MT scale scores as well as scores on assessments of related constructs. This research also assessed the incremental validity of MTSJT scores beyond general aptitude and

conscientiousness. Finally, this study applied machine learning techniques to investigate whether predictions based on scores of such measures were maximized at the item or scale level.

MT Issues

Despite the embryonic research suggesting MT scores may predict workplace outcomes, conceptualization and measurement challenges remain.

Conceptual Debate

One major impediment to MT research is the inability of researchers to reach consensus regarding the conceptualization of MT. While the definition from Gucciardi et al. (2015) provides a working synthesis of the field of research, it is distinct from other MT definitions in meaningful ways. For example, Jones, Hanton, and Connaughton (2002) defined MT as "having the natural or developed psychological edge that enables you to: generally, cope better than your opponents with many demands (competition, training, lifestyle) that sport places on a performer and specifically, be more consistent and better than your opponents in remaining determined, focused, confident, and in control under pressure." (p. 209). An inherent flaw in this definition is that MT is suggested to be an other-dependent or comparative construct (e.g., "cope *better than your opponents*…"), which contrasts with Gucciardi et al.'s definition of MT as an intrapersonal resource.

Researchers have also defined MT as inextricably linked to goal achievement (Arthur et al., 2015; Hardy, Bell, & Beattie, 2014). While researchers typically agree that MT relates to behavioral expression (e.g., task persistence), some researchers have defined MT as "the ability to achieve personal goals in the face of pressure from a wide range of different stressors" (Hardy et al., 2014, p. 70). Such a definition creates circularity because the process of MT is dependent

on obtaining desired outcomes. That is, it suggests that if individuals do not achieve their goal, they are not mentally tough. However, given the number of external factors that could affect goal attainment (Proctor & Dutta, 1995), it is untenable to argue that MT does not exist in the absence of goal attainment. Gucciardi et al.'s definition of MT differs from such definitions as it highlights the importance of *striving* to achieve goals, rather than actually achieving the outcomes.

Researchers have also debated the dimensionality of MT (Gucciardi et al., 2015; Lin et al., 2017) and have not reached a consensus regarding: a) whether MT is best conceptualized as unidimensional or multidimensional, and b) if MT is multidimensional, what are the underlying MT subdimensions. Most researchers argue MT is multidimensional (Clough, Earle, & Sewell, 2002; Gucciardi et al., 2015; Jones et al., 2002; Loehr, 1986) and as a result, several multidimensional assessments of MT have been created. For example, the 48-item Mental Toughness Questionnaire (Clough et al., 2002) is intended to measure confidence, emotional control, commitment to goals, and the ability to interpret stressors as challenges rather than threats. Similarly, the 14-item Sports Mental Toughness Questionnaire (SMTQ; Sheard, Golby, & van Wersch, 2009) conceptualizes MT as a combination of confidence, emotional control, and commitment.

Gucciardi et al. (2015) surveyed the major scales to assess MT and identified seven common factors, with no scale capturing all of the relevant factors. The factors were: generalized self-efficacy, buoyancy, success mindset, optimistic style, context knowledge, emotion regulation, and attention regulation. Middleton et al. (2011) suggested the 11 factors of MT were self-concept, potential, self-efficacy, task familiarity, personal bests, value, goal commitment, task focus, perseverance, positive comparison, stress minimization, and positivity. Thus,

although many researchers have treated MT as multidimensional, there is little consensus regarding the underlying measurement structure.

In stark contrast to the hypothesized multi-dimensional measurement models of MT, researchers have recently found support for a unidimensional MT model (Bédard-Thom & Guay, 2018; Gucciardi et al., 2015). Gucciardi et al. created the Mental Toughness Index (MTI) based on the seven factors mentioned above. However, contrary to expectations, a unidimensional solution was superior to the multidimensional solution, and scores from the unidimensional model predicted supervisor ratings, perceived distress, psychological health, and goal progress (Gucciardi et al., 2015).

Psychometric Issues

Given the issues researchers have faced regarding the definition and conceptualization of MT, it is not surprising that issues have arisen regarding the reliability, validity, and psychometric properties of the available measurement tools (Birch et al., 2017; Golby, Sheard, & van Wersch, 2007; Gucciardi, 2012; Gucciardi et al., 2012; Gucciardi & Gordon, 2011; Mack & Ragan, 2008, Middleton et al., 2004, Vaughan, Hanna, & Breslin, 2017). Popular MT scales have been critiqued for specific reasons. For example, researchers have struggled to consistently recover the proposed four-factor solution of the MTQ48 (Birch et al., 2017; Vaughan et al., 2017). Additionally, the SMTQ has been criticized for containing hyperbolic language (e.g., "I have an *unshakeable* confidence in my abilities") and double-barreled items (e.g., "I get angry and frustrated when things do not go my way"). Further, the convergent validity between these two measures has been lower than expected (Crust & Swann, 2011).

More importantly, the general measurement approaches to MT have been plagued by several issues common to nearly all assessments of MT. MT has been measured almost exclusively via self-report-Likert-scale questions whereby the respondent is asked to select the extent to which they believe a domain-general statement is descriptive of them (e.g., "I strive for continued success"). These items lack a frame of reference and are particularly susceptible to social desirability bias, faking, and acquiescence (Donovan, Dwight, & Hurtz, 2003; Soto, Gosling, & Potter, 2008; Van de Mortel, 2008). Additionally, the overreliance on this style of assessing MT makes it difficult for researchers to separate construct from method variance. That is, biases inherent in this style of measurement may be influencing relationships between MT scores and measures of other constructs, yet, because MT has almost exclusively been assessed via the same approach, researchers have been unable to determine the amount of variance attributable to the MT construct versus the method of assessment

MT measures have lacked an emphasis on assessing contextually-bound behavioral expressions, despite research showing MT scores are expected to differ across situations (Gucciardi et al. 2015, Hardy, Imose, & Day, 2014; Horsburgh et al. 2009; Veselka, Schermer, Petrides, & Vernon, 2009). Therefore, MT should be measured within a particular context. While relationships between MT and behavior are evident (e.g., Arthur et al., 2015; Gucciardi, Peeling, Ducker, & Dawson, 2016), current measures of MT focus more on measuring abstract characteristics instead of behavioral expression. Taken together, these critiques suggest that the items used to assess MT are too general, contain hyperbolic language, and should be more connected to patterns of behavior reflective of MT.

The Social Cognitive MT Model and MT Situational Judgment Test

In response to the conceptual and measurement-related issues above, Flannery et al. (2019) introduced a new measurement model of MT - the Social Cognitive MT model (SCMT). The SCMT is based on previous MT research that suggests there is as much within-person variability in MT scores as there is between-person variability (Gucciardi et al., 2015). Moreover, research of twins suggests that around half of the variance in MT scores is attributable to genetics, while the other half is attributable to non-shared environmental factors (Horsburgh et al., 2009; Veselka et al., 2009). Thus, the SCMT takes an interactionist approach by conceptualizing mentally-tough behavior and motivation as the product of both individual-difference characteristics and situational factors. Therefore, MT is conceptualized as a situation-bound chronic preference (rather than a trait), whereby individuals have a dispositional preference to display MT (or lack thereof), but situational factors influence the behavioral expression of the construct.

The foundation of the SCMT was taken from Mischel and Shoda's (1995) Cognitive-Affective Processing Model (CAPS), which suggests that personal factors (cognitive-affective processing units) interact with each other and one's environment to produce behavior. Individual differences exist among people in terms of the ease with which certain units are activated, as well as the stability of the MT dimensions within an individual. In this way, the SCMT accounts for both dispositional and situational determinants of MT, supporting the conceptualization of the construct as a chronic preference.

In addition to the research showing that MT scores are heavily influenced by the environment (Gucciardi et al., 2015; Horsburgh et al., 2009; Veselka et al., 2009), the social-cognitive approach to assessment is warranted because MT assessments that have contextualized

the items (e.g., "I have confidence in my *cricket* abilities"; Gucciardi & Gordon, 2009) have shown promise for predicting outcomes such as scores on measures of flow and hardiness. For example, Hardy et al. (2014) demonstrated that a domain-specific measure of MT predicted learning outcomes better than a domain-general measure. Beyond MT, research has generally supported the use of frame-of-reference wording on personality assessments, demonstrating higher predictive validity for contextualized assessments than general assessments (Lievens, De Corte, & Schollaert, 2008). However, such sport-specific assessments of MT are too narrow to be used beyond the specific sport for which they were developed.

The SCMT defines five cognitive-affective processing units presumed to underlie MT - task persistence, attention regulation, utilization of feedback, emotional control, and approach motivational orientation (Table 1 describes each processing unit). These units were identified and defined based on combining core factors of MT (Gucciardi et al., 2015) with tenets of social-cognitive (Mischel & Shoda, 1995) and self-regulatory (Bandura, 1991) theory.

Table 1

The Five Subcomponents of the Social Cognitive Mental Toughness Model

Dimension	Definition
Attention Control	The ability to focus on relevant stimuli while minimizing the intrusion of irrelevant stimuli
Emotional Control	The awareness and ability to use emotionally relevant processes to facilitate optimal performance and goal attainment.
Utilization of Feedback	The ability to utilize both positive and negative feedback from a variety of sources, including both objective standards and social responses, to direct behavior towards a goal.
Task Persistence	The ability to remain committed to a specific task despite challenges. This involves adjusting one's goals if they are met too easily, and exerting more effort to achieve challenging goals.

Approach Motivational Orientation	One's likelihood to set goals to achieve a favorable outcome rather than avoid an unfavorable outcome.

Based on the SCMT, Flannery et al. (2019) constructed the MTSJT to assess MT in the context of a workplace. Situational judgment tests provide respondents with scenarios and ask them to select response options that reflect how they would behave or think, typically measuring either behavioral-intentions or context knowledge. In the initial validation effort involving exploratory and confirmatory factor analyses, Flannery et al. obtained a three-factor solution that included task persistence, emotional control, and utilization of feedback (shown in Figure 1).

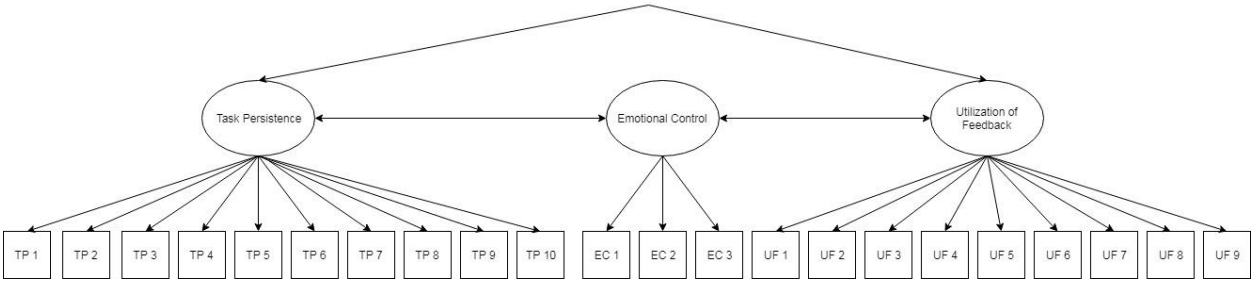

Figure 1. Mental Toughness Situational Judgment Test measurement model.

Further, the MTSJT demonstrated modest correlations (r = -.04 to .35) with the MTI, measures of the Big Five (Goldberg, 1999), the Regulatory Focus Questionnaire (Lockwood, Jordan, & Kunda, 2002), and the Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983). Two potential explanations for the magnitude of these correlations were noted. First, it's possible the MTSJT demonstrated high degrees of discriminant validity (evident via small correlations with Big Five measures) and modest convergent validity (evident via small-to-moderate correlations with the MTI) while associating modestly with approach motivational orientation and perceived distress, two common correlates of MT. Alternatively, the modest

correlations may have resulted because the MTSJT avoids common-method bias that typically inflates correlations among traditional Likert-style self-report measures (especially those lacking frame-of-reference wording), given that the MTSJT was the only measure included in the study that was not a traditional Likert measure. Therefore, it seems further investigation is needed regarding the nomological net surrounding the MTSJT, as well as its predictive validity.

Convergent and Discriminant Validity of MT Scores

As the nomological net surrounding the MTSJT is investigated, it is important to discuss issues about the convergent and discriminant validity of MT scores. Particularly, the jangle and jingle fallacies affect this network of measures (Credé et al., 2017). The jangle fallacy occurs when constructs are deemed to be distinct simply because they have different names (i.e., lack discriminant validity; Kelley, 1927). In contrast, the jingle fallacy occurs when two measures are assumed to assess the same construct simply because they have the same name (i.e., lack convergent validity; Thorndike, 1904). Both are particularly relevant to the assessment and application of MT.

Constructs that have been identified as contributing to the jangle fallacy in relation to MT include (but are not limited to) grit and resilience (Credé, Tynan, & Harms, 2017; Joseph, 2009). Grit is defined as a personality trait reflecting passion and perseverance to achieve long-term goals (Duckworth, Peterson, Matthews, & Kelley, 2007). Resilience is the capacity of a dynamic system to adapt successfully to disturbances that threaten its function, viability, or development (Masten, 2014). Given the conceptual similarities among these constructs, an investigation into the degree of convergence as well as the predictive accuracy of the various scale scores is warranted to help determine whether these constructs are distinct.

In addition to the modest correlations with MTI and MTSJT scores, empirical research has suggested the SMTQ and MTQ48 do not correlate as strongly as hypothesized (Crust & Swann, 2011). The researchers ultimately concluded that these scales could be capturing distinct yet related constructs, showing evidence of the jingle fallacy. As indicated by Johnson, Rosen, and Levy (2008), one manner of determining whether the same underlying construct is being assessed by multiple scales is to determine if the scales differentially predict outcomes. Thus, in conjunction with an assessment of the convergence among MT scores, evidence investigating this issue may be accrued by determining whether various MT assessments equally predict outcomes in an organizational context. By doing so, a better understanding of the nomological net surrounding the MTSJT would be advanced.

Item v. Scale Predictions

Developments in organizational sciences and predictive analytics have provided an innovative tool kit for examining the predictive validity of such constructs. Traditionally speaking, assessments of constructs such as MT have been created in accordance with classical test theory (CTT). CTT assumes that an observed test score decomposes into a true-score variance and measurement-error variance. Random measurement error "cancels out" when averaging across items to create a scale composite score. Given that each scale item likely contains unique variance, unique variance is treated as random-error variance in CTT. However, Putka, Beatty, and Reeder (2018) suggest unique variance likely accounts for additional variance in outcomes. Specifically, scores on individual items decompose into common variance (shared with other items in the scale), unique variance, and measurement error. An item-level approach may be particularly useful when assessing the predictive utility of the MTSJT because each

scenario contains contextual information; therefore, each item may carry a greater degree of unique variance than traditional MT items.

Recent statistical advances (i.e., machine learning models) have allowed for organizational researchers to investigate item-level predictions. Because item-level predictions require a potentially larger number of predictor variables to be entered into a model, larger sample sizes are typically required for traditional estimation methods, such as ordinary least squares regression. However, the advent of new methods such as the elastic net regression (Zou & Hastie, 2005) and random forests (Breiman, 2001a) provide a means of predicting outcomes even when the sample-size-to-predictor ratio is quite small. Putka et al. (2018) demonstrated that item-level analyses using modern predictive analytic methods account for more variance in outcomes than a) scale-level analyses and b) ordinary least squares regression analyses.

Incremental Validity over Common Predictors

Although MT has been established as a predictor of supervisor ratings and other organizational criteria (Lin et al., 2017), one question that has not been examined is whether MT accounts for incremental variance above the generally accepted predictors of job performance - cognitive ability and conscientiousness. Meta-analyses have estimated the corrected relationship between job performance and cognitive ability to range from .48 to .75 (Hunter, 1986; Hunter & Hunter, 1984), while the corrected relationship between conscientiousness and job performance is estimated to range from .24 to .38 (Barrick, Mount, & Judge, 2001; Salgado et al., 2015).

If MT is to be used as a predictor of job performance, it is important to determine if MT provides incremental validity in predicting outcomes beyond cognitive ability and conscientiousness. Research has shown MT to be generally unrelated to cognitive ability

(Flannery, Glasgow, Torian, DeVore, & Acton, 2017), but relationships between MT scores and conscientiousness scores have ranged from .16 to .43 (Flannery et al., 2019; Horsburgh et al., 2009).

The Current Study

In summary, the aims of the current study were to: 1) investigate the construct validity of MTSJT scores in relation to additional measures of MT and measures of related constructs, 2) determine if MT provides incremental prediction in comparison to well-established predictors of job performance, and 3) apply machine learning techniques to assess whether scale- or item-level models of MT scores are better predictors of job performance.

Chapter 2

Literature Review

Industrial/Organizational (I/O) psychologists have recently examined constructs that facilitate coping mechanisms in response adversity as predictors of workplace outcomes (Gucciardi et al., 2015; Lin et al., 2017). Such constructs have traditionally included grit, resilience, and conscientiousness. However, researchers have recently examined the use of MT as a predictor as well, finding that MT scores predict supervisor evaluations (Gucciardi et al., 2015), income (Lin et al., 2017), and ascension within the organizational hierarchy (Marchant et al., 2009). While MT has traditionally been established as a predictor of success in athletic contexts (Gucciardi & Gordon, 2011) and the military (Arthur et al., 2015; Godlewski & Kline, 2012; Gucciardi et al., 2015), the introduction of MT to the workplace holds promise for I/O psychologists but poses a set of new challenges.

Namely, researchers have struggled to 1) reach consensus regarding a conceptualization and operational definition of MT, 2) measure the MT construct validly and reliably, 3)

distinguish MT from similar constructs such as grit and resilience, and 4) determine if MT is a useful predictor of workplace outcomes in relation to other common predictors, such as cognitive ability and conscientiousness. Thus, these challenges suggest that a critical evaluation of the construct validity of MT scores is warranted before I/O psychologists can fully leverage MT scores as predictors of workplace outcomes.

In response to these issues, Flannery et al. (2019) introduced a novel assessment of MT – the Mental Toughness Situational Judgment Test (MTSJT). The MTSJT is grounded in socio-cognitive and self-regulatory theory, positing that the expression of mentally tough behavior is contingent upon both dispositional characteristics and contextual information. Thus, the MTSJT can be viewed as an improvement upon previous MT assessments for use in the workplace because it is domain-specific and assesses behavioral expectations rather than abstract traits. However, further investigation of the construct validity of MTSJT scores is needed for the assessment to be properly applied by I/O psychologists. This study was designed to accomplish this by investigating the convergent, discriminant, and criterion-related validity of MTSJT scores in the workplace via the use of machine learning algorithms.

Theoretical Development of the MTSJT

Issues regarding MT Definitions

Table 2 lists prominent definitions of MT. Each of these definitions has been critiqued as a deficient encapsulation of the MT construct. Specifically, Clough et al.'s (2002) definition has been criticized as a simple rebranding of the hardiness construct, considering that three of the four components of MT (control, challenge, commitment) are identical to the factors of hardiness (Gucciardi et al., 2011). Clough et al. propose MT to be distinct from hardiness as a result of the

inclusion of confidence as an MT dimension; however, research has shown confidence to correlate highly with hardiness (Gucciardi et al., 2011).

Table 2

Popular Definitions of MT in Academic Literature

Author(s)	Definition	Factors
Clough et al., 2002	"Mentally tough individuals tend to be sociable and outgoing; as they are able to remain calm and relaxed, they are competitive in many situations and have lower anxiety levels than others. With a high sense of self-belief and an unshakeable faith that they control their own destiny, these individuals can remain relatively unaffected by competition and adversity." (p.38)	Challenge, commitment, control, confidence
Jones et al., 2002	"Mental toughness is having the natural or developed edge that enables you to: i) always, cope better than your opponents with many demands (competition, training, lifestyle) that sport places on a performer; ii) specifically, be more consistent and better than your opponents in remaining determined, focused, confident, and in control under pressure. (p. 209)"	Self-belief, desire/motivation, dealing with pressure and anxiety, performance-related focus, lifestyle-related focus, dealing with pain/hardship
Sheard et al., 2009	"the ability to bounce back from stressful experiences, such as competitive sport, quickly and effectively" (p. 188)	Constancy, emotional control, commitment
Gucciardi et al., 2015	"personal capacity to produce consistently high levels of subjective (e.g., personal goals or striving) or objective performance (e.g., sales, race time, GPA) despite everyday challenges and stressors as well as significant adversities" (p. 28)	Self-belief, attention regulation, emotion regulation, success mindset, situational knowledge, buoyancy, and optimism.

Jones et al.'s (2002) definition of MT included 12 key attributes of a mentally tough performer, including: a) "having an unshakeable belief in your ability to achieve your competition goals," b) "having an insatiable desire and internalized motives to succeed," and c) "remaining fully focused on the task at hand in the face of competition-specific distractions." Jones et al.'s definition has been critiqued as other-dependent (e.g., "cope *better than your opponents*"), which implies that one's level of MT is contingent upon the performance of others (Gucciardi et al., 2011). From the perspective of a researcher seeking to use MT scores as a predictive tool, this conceptualization creates difficulties in comparing across individuals.

Additionally, Sheard et al.'s (2009) definition of MT was criticized for being too parsimonious (Gucciardi et al., 2011). That is, Sheard et al.'s conceptualization suggests MT is a combination of constancy, confidence, and emotional control. However, this conceptualization lacks other commonly identified MT dimensions such as attention control, success mindset, and contextual intelligence, among others.

In an attempt to synthesize the myriad of MT definitions in psychology, Gucciardi et al. (2015) reviewed the literature on MT and defined it as "the personal capacity to produce consistently high levels of subjective (e.g., personal goals or strivings) or objective (e.g., sales, race time, GPA) performance despite everyday challenges and stressors as well as significant adversities" (p. 28). Further, they identified seven core factors of MT presumed to be integral to the construct, including generalized self-efficacy, buoyancy, success mindset, optimistic style, context knowledge, emotion regulation, and attention regulation. Considering the extensive literature review conducted, Gucciardi et al.'s definition appears to be the most comprehensive and accepted definition of MT to date.

Unidimensional vs. Multidimensional Conceptualization

Although Gucciardi and colleagues (2015) identified seven core factors of MT, they were unable to empirically recover these factors when assessing MT via their Mental Toughness Index (MTI). Rather, the authors found a unidimensional model to fit the data best. As a result, Gucciardi et al. questioned decades of research suggesting MT is a multidimensional construct, stating "Interpretations regarding the dimensionality of MT that scholars have made of performers' perceptions of this construct from qualitative research may not be entirely accurate, including our own early work, and therefore require reconsideration and examination in future research" (p. 40). Thus, research investigating further the dimensionality of MT is necessary to fully understand the construct validity of MT scores.

Trait vs. Chronic Preference

Similarly, researchers have debated whether MT is best conceptualized as a trait or a chronic preference influenced by context. If MT is a trait, then the scores will be stable across time and context, suggesting that global, infrequent assessments of MT will predict future outcomes. If MT is conceptualized as a chronic preference, measuring MT within a context will be necessary to predict outcomes within that context.

Empirical research to-date has suggested that MT is a chronic preference (Gucciardi et al., 2015; Lin et al., 2017). Specifically, Gucciardi et al. (2015) assessed individuals daily for two weeks and found MT scores to have as much within-person variance (56%) as between-person variance (44%). Relatedly, research of identical twins suggests that environmental influences have roughly as much influence on MT scores as genetic factors (Lin et al., 2017). Specifically, Horsburgh et al. (2009) demonstrated that genetics accounted for 52% of the variance in MT

scores among twins, while nonshared environmental factors accounted for 48%. Interestingly, these authors found that the subscales of emotional control and commitment to goals were particularly influenced by the environment. Similarly, Veselka et al. (2009) found that nonshared environmental factors explained 47% of the variance in MT scores among twins.

Social Cognitive Mental Toughness Model

In response to the major issues surrounding the conceptualization and measurement of MT, Flannery et al. (2019) introduced the Social Cognitive Mental Toughness Model (SCMT) that guided the development of the MTSJT. This model integrates the core factors of MT within the broader literature established by social-cognitive and self-regulatory theories. In conjunction with empirical research, MT is conceptualized as a chronic preference such that individuals tend to express mentally tough behavior (or lack thereof), yet situational factors influence their likelihood of demonstrating MT. In this manner, stable between-individual differences, as well as within-person variability in MT scores, are captured.

Mischel and Shoda's (1995) CAPS model provides a strong theoretical basis in which to ground the SCMT and explain the situational sensitivity of MT. In the CAPS model, behavioral expressions are the product of interactions between an individual's cognitive and affective processes and environmental factors. Specifically, Mischel and Shoda suggest individuals possess cognitive-affective units (CAUs) that are personal variables that interact with each other and with the environment to produce behavior. Individual differences exist regarding the ease of access, usability, patterns of connections, and stability of CAUs, which explains why the same situational factors do not elicit the same behavior from each individual.

According to the SCMT, MT consists of five CAUs: task persistence, emotional control, attention control, utilization of feedback, and approach motivational orientation. Table 1 summarizes these CAUs. These five components were selected based on an extensive review of the qualitative and quantitative literature pertaining to MT (e.g., Gucciardi & Gordon, 2011; Gucciardi et al., 2015, Jones et al., 2002, Lin et al., 2017). Each factor has been consistently identified throughout the MT literature and aligns with the social-cognitive and self-regulatory theories upon which the SCMT was developed (Bandura, 1991; Mischel & Shoda, 1995). However, not all factors have been directly assessed by previous MT measurement models (e.g., task persistence and utilization of feedback).

Broadly speaking, each CAU influences goal-directed behavior that enables successful performance and increases the likelihood of goal attainment. Attention control facilitates goal-attainment by channeling attention despite distractions from serious obstacles (Bandura, 1991). Regarding emotional control, self-regulatory theory suggests individuals who can manage emotions are likely to have higher self-perceptions and more confidence, which positively influences motivation and behavior in pursuit of a goal (Bandura, 1991). Further, the utilization of feedback is important in goal attainment because to enhance their performance individuals must self-observe their motivation and behavior in conjunction with information from the environment.

Task persistence focuses on the effort component of self-regulation theory, such that when individuals identify a discrepancy between their behavior and their personal standard, they must exert some degree of effort to improve their behavior or increase their goal standards. Finally, approach-avoidance theorists argue that the overarching motivation affecting all behavior is the desire to either acquire positive consequences or to avoid negative consequences,

and such goals are an essential part of the self-regulation process (Elliot & Thrash, 2002). Research suggests that individuals in a state of high MT tend to use approach coping styles to address issues (Kaiseler et al., 2008; Nicholls et al., 2008). In sum, from a social-cognitive perspective, these five factors reflect crucial aspects of MT that direct goal-relevant behavior. However, as discussed in detail below, the MTSJT was only able to capture three of these dimensions – task persistence, emotional control, and utilization of feedback.

Improving MT Measurement

The lack of agreement regarding how best to define and conceptualize MT has led to a multitude of self-report assessments of MT. Several major critiques exist regarding specific scales, such as the 48-item Mental Toughness Questionnaire (MTQ48, Clough et al., 2002), the MTI (Gucciardi et al., 2015), and the Sports Mental Toughness Questionnaire (SMTQ; Sheard et al., 2009). The critiques of these scales include poor factor structures, poorly worded items, a lack of construct validity, and lack of a frame of reference, among others (Anderson, 2011, Birch et al., 2017, Crust & Swann, 2011; Gucciardi & Gordon, 2011, Gucciardi et al., 2012, Perry et al., 2013; Vaughan et al., 2017). Thus, researchers have called into question the validity of the findings using such scales, and suggested that alternative approaches to measuring MT should be explored (Birch et al., 2017, Gucciardi & Gordon, 2011, Gucciardi et al., 2012).

MT has been nearly exclusively assessed via self-report scales in which respondents assess the extent to which domain-general statements (e.g., "I have an unshakeable confidence in my ability") are descriptive of them. Such scales are particularly susceptible to various self-report biases, including socially-desirable responding, acquiescence, and faking (Donovan et al., 2003; Soto et al., 2008; van de Mortel, 2008). The fact that most studies have assessed MT from

this perspective (Lin et al., 2017) makes it difficult to disentangle construct-relevant variance from common-method variance.

Although most MT assessments are devoid of frame-of-reference wording, empirical research in the field of MT has supported the importance of assessing MT in context. For example, Hardy et al. (2014) compared the predictive validity of a domain-general MT measure and a domain-specific MT measure on a complex video game learning task. Specifically, the authors found that domain-specific MT predicted task-related self-efficacy, enjoyment, knowledge, and performance whereas domain-general MT scores did not. Further, the domain-specific assessment provided incremental validity in predicting outcomes above control variables, including cognitive ability, core self-evaluations, and goal orientation.

Sport-specific MT assessments have also been developed, with empirical research indicating these assessments are effective in differentiating players of different abilities and predicting athlete burnout while demonstrating moderate correlations with measures of flow, hardiness, and resilience (Gucciardi & Gordon, 2011; Gucciardi et al., 2009a). Despite these encouraging findings, these scales are not applicable outside of the narrow context in which they were developed.

Finally, MT assessments have lacked an emphasis on behavioral expression. As Arthur et al. (2015) suggested, MT is inextricably linked to behavior and should be assessed at the behavioral level. Likewise, Hardy et al. (2014) indicated that MT must be conceptualized and measured only as an outward behavior, defining it as "the ability to achieve personal goals in the face of pressure from a wide range of stressors" (p. 70). Further, both Arthur et al. and Hardy et al. introduced behavior-based assessments of MT (for the military and sports, respectively) which predicted outcomes incrementally over existing MT measures. However, critiques of these

scales include doubt regarding the extent to which the behaviors listed epitomize MT, the fact that MT reflects achieving (rather than striving to achieve) goals, and a lack of generalizability across contexts.

MTSJT

To address these shortcomings yet maintain ease of administration, Flannery et al. (2019) introduced the MTSJT. This assessment captures MT in the workplace within a situational judgment test framework in which participants are provided with a scenario and asked to assess the likelihood they would engage in each of the associated response options. This assessment addresses previous limitations regarding MT assessments by placing respondents directly in the workplace and asking them to evaluate behavioral expectations in that context.

In the creation of the original item bank of the MTSJT (consisting of 40 scenarios and 120 total response options), subject-matter experts relied on knowledge of the construct, personal experience, as well as consultations with colleagues to determine critical incidents that individuals might experience in the workplace. A guiding framework of contexts was created to generate situations (See Table 3 for a description of the framework). All situations were classified as either acute or long-term situations. Acute situations require immediate action to resolve (e.g., a mishap during a presentation), whereas long-term situations require a persistent plan of action (e.g., working to achieve a positive performance appraisal).

Table 3

Taxonomy for Scenario Development of the Mental Toughness Situational Judgment Test

Classification	Description	Example
Acute – High-stakes	A scenario requiring performance pertaining to a situation in which the consequences are high-stakes	Being asked to give an important presentation
Acute – Threat to self-image	A scenario involving a reaction to an immediate potential threat to self-image	Dealing with disrespectful behavior from your coworker
Acute – Self-discipline	A scenario involving dedication to a mundane task to succeed	Working on a necessary project instead of attending a social event
Long term – High-stakes	A scenario involving performance despite chronic pressure and important outcomes	Adjusting to a new promotion
Long term – Threat to self-image	A scenario involving a threat to self-image that persists in one's mind over time.	Working to improve behavior after a negative performance appraisal
Long term – Self-discipline	A scenario involving dedication to achieve an overarching goal	Exerting effort to achieve a new promotion

Within these two broad domains, each situation was designed to represent the following three types of issues: situations with high-stakes consequences, threats to self-image, and situations requiring self-discipline. Scenarios could be characterized by several issues – for example, a scenario might involve both a high-stakes consequence and a threat to self-image (e.g., receiving corrective feedback on a make-or-break project). High-stakes situations involve consequences contingent upon one's performance. Threats to self-image involve a degree of personal adversity relating to one's perception of him/herself. Finally, self-discipline situations require dedication and adherence to repetitive or mundane tasks to be successful.

Situations were designed to cover a wide variety of issues, including both intrapersonal issues (e.g., being personally challenged by a difficult task) and interpersonal issues (e.g.,

dealing with inappropriate coworker behavior). Eighty-four of the 120 total response options were designed to represent adaptive responses that would help to remedy the situation. In contrast, 36 response options were reverse coded to represent maladaptive behaviors in response to the situation. A behavioral-tendency (i.e., "would" instead of "should") response format was chosen to measure the construct with regard to intent (i.e., real vs. ideal). Studies have found that behavioral-tendency response options can have better criterion-related validity than knowledge-based response options (Ployhart & Ehrhart, 2003)

As indicated by the conceptualization of Gucciardi et al. (2015), MT includes not only relevant behaviors but also attitudes and beliefs tied to behavioral expression. Thus, while 23 scenarios asked respondents to indicate the likelihood they would engage in each of the behaviors indicated, 17 scenarios included a behavior in the item stem and asked how the respondent would think and feel in the particular situations. In this way, the attitudes and beliefs reflected in MT would be presumably captured, providing a more comprehensive assessment of the construct space. However, contrary to existing self-report measures of MT (i.e., the SMTQ and MTI), the MTSJT measures attitudinal responses tied directly to specific elements of the context. Thus, rather than measuring domain-general attitudes of MT, the MTSJT assesses context-bound attitudes of MT that underlie specific behaviors.

Originally, the MTSJT was designed to assess a hierarchical MT factor in addition to the five subcomponents identified in the SCMT. However, an exploratory factor analysis on a sample of 466 undergraduates suggested (after item reduction) that a three-factor solution (consisting of task persistence, utilization of feedback, and emotional control) fit the data best ($\chi 2$ (150) = 241.98, p <.01, CFI = .93, RMSEA = .04, SRMR = .04). A subsequent analysis on a holdout sample (N = 251) attempted to confirm the factor structure. However, a hierarchical

structure could not fit the data because a hierarchical model with three factors is just-identified at the second-order level, disallowing for the existence of a hierarchical factor (Credé & Harms, 2015).

Instead, a correlated factors model was fit to the data (as shown in Figure 1) and demonstrated adequate fit (x^2 (186) = 314.74, $p < .01$, CFI = .90, RMSEA = .05, SRMR = .06; Hu & Bentler, 1999). This model fit better than a unidimensional model and a scenario-based model wherein all response options within a given scenario loaded onto the same factor. Ultimately, these results suggested that the MTSJT was capturing three salient factors. The final form of the MTSJT consisted of 11 scenarios and 21 response options - 3 representing emotional control, 10 related to task persistence, and 8 reflecting utilization of feedback.

Additionally, Flannery et al. (2019) observed correlations among MTSJT scores and other measures, including the MTI (r = .29), the Perceived Stress Scale (PSS; r = -.09), the Regulatory Focus Questionnaire-Approach Motivation (RFQ; r = .35), the Regulatory Focus Questionnaire-Avoidance Motivation (r = .13) and the Big Five (r = -.04 to .24). Ultimately, correlations among MTSJT scores and these additional measures were small to moderate in magnitude. Notably, the correlations among MTSJT scores and MTI scores represented only modest convergence

However, the MTI scores demonstrated stronger correlations with all other measures in the battery (.16 to .49). It is hypothesized that the correlations among MTI scores and PSS, RFQ, and Big Five scores were inflated due to common method bias. Further, the correlations among MTSJT scores and PSS, RFQ, and Big Five scores may be attenuated because the MTSJT was the only assessment that used a situational judgment test framework while focusing on behavioral expectations.

Although approach-motivation orientation and attention control are factors of MT proposed by the SCMT, the empirical validation effort was unable to recover these factors. Two main explanations may account for this. First, the MTSJT is distinct from previous MT measures because it directly assesses task persistence and utilization of feedback. Given the prominence of task persistence and utilization of feedback in this model, other traditional components (e.g., attention control) may be less salient.

Second, it may be that attentional control is indistinguishable from the other factors as a result of the SJT method used to assess MT. That is, the use of traditional self-report methods may enable participants to distinguish the attention control dimension, but this dimension becomes obfuscated when using the SJT approach, perhaps due to the behavioral and contextual emphasis of that measure.

A final consideration is that MT may be best modeled as a reflective, causal indicator model, which combines aspects of both reflective and formative models (Bollen & Bauldry, 2011). With such an approach, task persistence, utilization of feedback, and emotional control items would be considered reflective, while assessments of approach-motivational orientation and attention control would be considered formative. Considering that more foundational research is needed to investigate such an approach, the current study evaluated the validity of the three-factor, reflective model found by Flannery et al. (2019); therefore, attention control and approach-motivational orientation were not measured. However, these dimensions may still be modeled and measured as part of the SCMT in additional research.

Convergent and Discriminant Validity

The modest correlations among MTSJT scores and all other self-report measures observed by Flannery et al. (2019) call for further investigation into the convergent and

discriminant validity of MT assessment tools. Research has generally suggested that the convergent and divergent validity of MT scores depend largely on the measurement scale used (Arthur et al., 2015, Crust & Swann, 2011, Flannery et al., 2019). As a result, two major measurement fallacies are relevant for the measurement of MT - the jangle and jingle fallacies. Investigating the extent to which such fallacies apply to MTSJT scores (and existing measures of MT) is critical to understanding the construct validity of scores on MT measurements.

The Jangle Fallacy

The jangle fallacy occurs when two measures are assumed to measure different constructs simply because they have different names (i.e., lack discriminant validity; Kelly, 1927). For the current study, two prominent constructs presumed to have overlap with MT were selected to explore further - grit and resilience (Joseph, 2009). Grit is defined as a personality trait capturing passion and perseverance for long term goals (Duckworth et al., 2007). Resilience is defined as the capacity of a dynamic system to adapt successfully to disturbances that threaten its function, viability, or development (Masten, 2014).

Scholars have suggested theoretical distinctions among these constructs. Specifically, Gucciardi (2017) suggested that grit and MT are distinct because a) grit is defined explicitly as a personality trait, while MT has both trait and state qualities, and b) grit enables an individual to remain persistent in pursuit of a certain goal, whereas MT reflects the ability to prioritize, juxtapose, and integrate multiple goals simultaneously. The main distinguishing factor between MT and resilience is that resilience is defined as a reactive construct, but MT can be both reactive and proactive (Gucciardi, 2017).

A recent meta-analysis of the research on grit (k = 6) suggested the population correlation between MT and grit scores to be .49, with the 90% confidence interval ranging from .35 to .56, indicating significant overlap (Credé et al., 2017). While meta-analytic estimates of the association among MT scores and resilience do not exist, empirical research has indicated that the correlation between these constructs ranges from .10 to .55 depending on the scale and sample used (Arthur et al., 2015, Flannery et al., 2017; Gucciardi, Gordon, & Dimmock, 2009a).

There is preliminary evidence suggesting that both grit and resilience are predictors of workplace performance (Dugan, Hochstein, Rouziou, & Britton, 2018; Suzuki, Tamesue, Asahi, & Ishikawa, 2015; Youssef & Luthans, 2007). According to Johnson et al. (2008), one approach to determining whether a set of measures reflect the same construct rather than distinct yet related constructs is to determine whether the scales show unique effects. Theoretically, measures of the same construct should show similar levels of predictive accuracy for a given outcome. If the constructs are distinct, assessments of such constructs might not predict equally.

The Jingle Fallacy

In addition to the unintended empirical overlap among MT measures and measures of grit and resilience, many measures of MT have not overlapped to the extent hypothesized (Arthur et al., 2015; Crust & Swann, 2011, Flannery et al., 2019). This is known as the jingle fallacy, wherein assessments are assumed to measure the same construct simply because they have the same name (i.e., lack convergent validity; Thorndike, 1904), explaining in part why correlations among MT scores and grit/resilience scores vary as a function of the measures used.

Examining the item content across several MT scales highlights potential differences (See Table 4 for examples). As evident in Table 4, these MT items capture very different

dimensions of MT, including outperforming other individuals, surviving pain, and attaining relevant skills. The wide range of content captured by these scales contributes to limited convergent validity among various MT scales. More specifically, Crust and Swann (2011) investigated the extent to which two multidimensional measures of MT, the SMTQ and the MTQ48, assessed the same underlying construct. While they found strong correlations among the hierarchical factors (r = .75), the correlations among lower-order factors were smaller in magnitude than expected, despite the theoretical overlap among the factors. For example, the emotional control subscales across both assessments correlated at .49. Further, each measure contained a factor designed to assess striving toward goals, yet these scores correlated at .61. The authors concluded that these scales captured similar yet distinct constructs.

Table 4

Construct Interference Evident in Items of MT Scales

Item	Scale	Construct Interference
He has been reprimanded/punished	Military Training Mental Toughness Inventory	Heavy influence from context and others' behavior
He is in pain	Military Training Mental Toughness Inventory	Physical influence
I can find a positive in most situations	Mental Toughness Index	Strongly associated with optimism
I am able to execute appropriate skills or knowledge when challenged	Mental Toughness Index	Context knowledge and skills are required
Player X is able to maintain a high level of performance in competitive matches when people are relying on him to perform well	Mental Toughness Inventory	Other-dependent
Player X is able to maintain a	Mental Toughness Inventory	Physical influence

high level of performance in competitive matches when he is struggling with an injury		
Uncontrollable events like the wind, cheating opponents, and bad officials/judges get me very upset	Psychological Performance Inventory	External contingencies
I use images during play that help me perform better	Psychological Performance Inventory	Use of mental imagery

MT as a Predictor of Success

In addition to investigations of the convergent and discriminant validity of MT scores, an additional approach to establish the construct validity of MT scores involves an investigation of the predictive validity of such scores. Traditional research has focused on examining MT as a predictor of success among athletes, with substantial research supporting MT as a predictor of competitive success (Gucciardi & Gordon, 2011; Gucciardi et al., 2015, Jones et al., 2002), level of performance (e.g., amateur vs. professional; Chen & Cheesman, 2013; Gucciardi et al., 2009), performance under pressure (Bell, Hardy, & Beattie, 2013; Goldberg, 1998), and recovery from injury (Petrie, Deter, & Harmison, 2014). These findings resulted in the development of sport-specific MT training programs that have been shown to increase both MT scores and athletic performance (Gucciardi, Gordon, & Dimmock, 2009b; Gucciardi, Gordon, & Dimmock, 2009c; Lin et al., 2017; Sheard & Golby, 2006). Overall, empirical studies of MT among athletes have concluded that MT facilitates performance and goal attainment by lowering perceived distress and increasing an athlete's use of adaptive coping strategies (Gucciardi et al., 2015).

Researchers have begun investigating the potential merits of MT in other contexts, including the military, education, and the workforce. Regarding military training, Gucciardi et al.

(2015) demonstrated that MT scores predicted training outcomes in the Australian Defence Force, whereas hardiness and self-efficacy did not. Among the Canadian Armed Forces, MT scores predicted voluntary turnover as mediated by normative and affective organizational commitment (Godlewski & Kline, 2012). Finally, Arthur et al. (2015) demonstrated that the superordinate's behavioral ratings of subordinates' MT predicted the performance ratings of infantry personnel.

MT has also been shown to predict performance on cognitive tasks (Dewhurst, Anderson, Cotter, Crust, & Clough, 2012; Hardy et al., 2014). More specifically, Dewhurst et al. found that individuals scoring high in MT performed better on a directed-forgetting paradigm, suggesting such individuals are better at preventing the intrusion of irrelevant stimuli from distracting them from the pursuit of their goals. Further, Hardy et al. (2014) found that those scoring high in MT performed better at a video game learning task, particularly when MT was assessed via a domain-specific measure.

Within the realm of academic achievement, MT scores have been shown to predict progress over time (Gucciardi et al., 2015), as well as attendance, teacher evaluations, frequency of counterproductive classroom behavior, and quality of peer relationships among students (St. Clair-Thompson, et al., 2015), self-esteem and confidence among students at a new school (St. Clair-Thompson et al., 2017), and overall grades among both student-athletes and non-athlete students (Crust et al., 2014; Lin et al., 2017).

Regarding MT workplace research, Gucciardi et al. (2015) found MT scores to predict supervisor ratings of employees' job performance, mediated by lower levels of perceived distress and higher levels of adaptive coping. Among junior, mid, and senior-level managers, Marchant et al. (2009) found that MT scores correlated positively with one's position in an organizational

hierarchy. Moreover, in a sample of 100 full-time employees from across the world, Lin et al. (2017) found MT scores to correlate positively with financial income, even after controlling for age and gender.

Incremental Validity

As research regarding the predictive validity of MT scores has grown, practitioners have become more interested in using assessments of MT for various purposes (Gucciardi et al., 2015; Lin et al., 2017). Shoenfelt (2016) advocated that resilience and other oft-studied constructs in sports psychology be used by I/O psychologists. Also, the conclusion of a review of the literature on MT advocated that researchers and practitioners further apply MT to the workplace (Gucciardi & Gordon, 2011).

Despite such sentiments, researchers and practitioners need to be careful when applying such constructs to the workplace. Moreover, I/O psychologists should resist using such constructs to predict workplace outcomes before critically examining their validity and utility. Such a critical evaluation should involve comparisons among assessments of established predictors to determine if new predictors account for incremental variance.

Consider the research on grit as an example. Researchers quickly assembled a substantial body of research to show that grit scores predict a variety of outcomes in sports, work, military training, and education (e.g., Credé et al., 2017; Duckworth et al., 2007; Duckworth, Kirby, Tsukayama, Berstein, & Ericcson, 2011, Eskreis-Winkler, Duckworth, Shulman, & Beal, 2014; Kelly, Matthews, & Bartone, 2014). Further, grit obtained a large degree of media coverage (e.g., Ted Talks, interviews, magazine articles, podcasts) touting this construct as an extremely important predictor of meaningful outcomes (Duckworth, 2013; Sclefo, 2016). Such research and

media coverage influenced the U.S. Department of Education to recommend that grit be taught in schools - a policy that has been adopted by some schools already (Credé, 2018).

Despite this, there is little evidence to suggest that grit training-programs are effective (Credé, 2018). More importantly, a recent-meta analysis on grit found that grit scores offer little value in prediction once cognitive ability and conscientiousness have been controlled (Credé et al., 2017). Across over 70 studies and 60,000 individuals, the authors found that grit scores provided no incremental prediction in overall academic performance, college GPA, or high-school GPA after controlling for cognitive ability and conscientiousness. However, when grit was entered first into the model, conscientiousness scores did capture unique variance beyond that accounted for by grit. Ultimately, these authors concluded, "Grit as a predictor of performance and success and as a focus of intervention holds much intuitive appeal, but grit, as it is currently measured does not appear to be particularly predictive of success and performance…" (p. 504).

While the meta-analysis of Credé et al. (2017) focused on predicting educational outcomes, the implications for I/O psychologists interested in predicting workplace performance are clear: the predictive accuracy of MT scores (and scores of similar constructs) must be compared with established predictors to determine if they capture incremental variance before advocating their use. The current study assessed cognitive ability and conscientiousness, which are generally considered to be among the strongest predictors of job performance (Barrick et al., 2001; Barrick & Mount, 1991; Hunter & Hunter, 1984).

Cognitive Ability

Cognitive ability has been identified as the strongest predictor of job performance, with decades of research supporting this conclusion (Hunter, 1983; Hunter & Hunter, 1984; Schmidt, 2002; Van Iddekinge, Aguinis, Mackey, & DeOrtentiis, 2017). Theoretically, scholars have suggested that cognitive ability is a strong predictor of performance because individuals scoring high in cognitive ability can quickly and effectively develop greater degrees of job knowledge, which ultimately allows them to perform well on the job and earn higher supervisor ratings (Hunter, 1983). However, cognitive ability also has a direct effect on job performance (even when controlling for job knowledge) suggesting that it plays a critical role in immediate performance. Empirical research has found that cognitive ability serves as a consistently strong predictor of job performance across outcome measures (e.g., supervisor ratings, objective performance criteria, training performance) and occupations (e.g., civilian jobs, military jobs, entry-level jobs, managerial positions; Hunter, 1986; Hunter & Hunter, 1984). Such estimates were moderate to strong in magnitude, with corrected relationships estimated to be between .48 and .75 (Hunter, 1986; Hunter & Hunter, 1984).

The relationship between MT scores and cognitive ability scores has not been fully investigated. Dewhurst et al. (2012) found that those scoring high on measures of MT performed better on a directed-forgetting paradigm and concluded they had identified a cognitive basis of MT, such that those scoring high in MT can prevent irrelevant stimuli from impacting their cognitive functioning. Despite this potential cognitive basis, research has suggested that MT scores are orthogonal to cognitive ability scores (Flannery et al., 2017). In a sample of 242 undergraduate students, Flannery and colleagues found that scores on the Wonderlic intelligence test were unrelated to scores on the SMTQ (r = .09) and the MTI (r = .03). To date, no other

empirical investigations comparing cognitive ability and MT scores have been published. Furthermore, the literature on grit indicates this construct is unrelated to cognitive ability (Credé et al., 2017), implying it is likely that MT scores are also unrelated. These results suggest that MT scores are capable of capturing unique variance in outcomes beyond the variance accounted for by cognitive ability scores.

Conscientiousness

While MT and cognitive ability appear to be distinct, conceptual and empirical overlap exists among measures of MT and conscientiousness. Several researchers have argued that conscientiousness is a construct contributing to the jangle fallacy surrounding MT, grit, and related constructs (Credé et al., 2017). Observed correlations among MT scores and conscientiousness scores have ranged from small to moderate (.16 to .43; Flannery et al., 2019; Horsburgh et al, 2009). Horsburgh et al. suggest MT and conscientiousness likely share a similar construct space, because those who remain strongly committed to setting and achieving goals are likely to be more conscientious individuals. However, MT captures components not traditionally associated with conscientiousness (e.g., confidence, emotional control); MT is also presumed to be unrelated to certain aspects of conscientiousness, such as orderliness. Like grit, conscientiousness is proposed to be a stable trait, while MT is presumed to vary as a function of context. Despite such distinctions, an empirical investigation comparing the predictive accuracy of assessments of each construct is certainly needed.

It is well-established that conscientiousness scores are a generalizable predictor of job performance (Barrick & Mount, 1991; Barrick et al., 2001; Hurtz & Donovan, 2000; Salgado et al., 2015; Rojon, McDowell, & Sanders, 2015; Shaffer & Postlethwaite, 2012). Given the

potential overlap between MT and conscientiousness, comparisons of predictive accuracy for these constructs are warranted.

Modern Predictive Analytics - Item v. Scale Predictions

The literature reviewed here suggests the need for research to compare the predictive accuracy of measures of MT, grit, resilience, cognitive ability, and conscientiousness for workplace outcomes. A primary aim of the current study was to compare the validity of these measures to predict supervisor ratings of employees. This methodology requires the recruitment of dyads (supervisors and employees), which typically results in smaller sample sizes. Furthermore, examinations of the content of the aforementioned scales (MTI, SMTQ, MTSJT) suggest that the items themselves may carry unique variance that ultimately becomes washed out upon aggregating to scale-level scores. For example, MTSJT items contain a large degree of contextual information which may increase the predictive validity of each item. Theoretically, this suggests that creating item-level models may increase the predictive accuracy of such assessments.

However, traditional modeling methods such as ordinary least squares regression (OLS) are ill-suited for item-level predictions because they require many predictor variables to be entered into the model, creating issues regarding multicollinearity, model parsimony, and overfitting. Further, OLS necessitates a greater sample size when including many predictor variables in the model, which is not always feasible for researchers. Recently, I/O psychologists have leveraged machine learning algorithms capable of addressing such issues and create accurate predictions that are stable across data sets. Two algorithmic approaches used in this study were elastic net regression and random forests.

The Bias-Variance Tradeoff

One major advantage of modern predictive analytics is a remedy to the bias-variance tradeoff (Putka et al., 2018). Model bias refers to error introduced by approximating a real-life problem via a much simpler model (James, Witten, Hastie, & Tibshirani, 2013). In contrast, model variance refers to the extent to which the model parameters change when they are fit on a different data set (James et al., 2013). Essentially, this tradeoff is based on the concept of model flexibility. Flexibility reflects the extent to which the model closely fits each point in the data. For example, it is possible to draw a line of best fit through a data set such that each point fits perfectly on the line, thereby overfitting the data. Such a model would be characterized as highly flexible and would make no errors in predicting the values of the data set upon which it was fit. However, this model would suffer from high variance — the model would perform poorly if used to predict the same outcomes in an independent data set because it overfit the training data.

In contrast, a high-bias model is too simple and underfits the data. Consider a model that is drawn as a straight horizontal line through a fixed point on the y-axis, despite a clear positive relationship between the predictor and the criterion. This model is not flexible and has no variance — it will produce the same model parameters in a new data set, yet the prediction will also suffer because it underfit the data and did not make use of the systematic variability in the data. Thus, the goal of many machine learning algorithms is to balance this tradeoff. Simply put, there is a "sweet spot" researchers should strive to hit to maximize prediction.

Sample Size/Predictor Ratio

Another contribution of modern predictive methods is that they outperform traditional methods when the sample size is small and the number of predictors is large (N/P ratio; Putka et

al., 2018). With traditional methods, low N/P ratios are problematic because such models tend to overfit the data and create issues of multicollinearity, ultimately resulting in highly variable model parameters. This advantage offered by modern predictive analytics is important given the difficulty of obtaining a large number of supervisor evaluations in research.

Tuning Parameters

Modern predictive analytics address the bias-variance tradeoff by using tuning parameters, which are model features that help researchers scale the complexity of the model and constrain estimates of the model parameters to reduce the degree to which the parameters are sensitive to noise in the data. Tuning parameters allow such methods to make accurate predictions when the N/P ratio is small. Within the training data, the optimal values for tuning parameters are selected via a method known as *k*-fold cross-validation, which is described in detail in the Analysis section.

Elastic Net

The elastic net is a regression-based analytic method that addresses the bias-variance tradeoff by using two tuning parameters: α and λ. The elastic net is an extension of ridge and least absolute shrinkage and selection operator (LASSO) regressions as it combines aspects of both models. By introducing some bias into the model, the elastic net reduces variability and provides stable predictions on holdout samples (Putka et al., 2018).

The purpose of a basic OLS model is to minimize the sum of squared errors (SSE) from a prediction line. Ridge regression is an extension of OLS that deals with issues of multicollinearity by incorporating the tuning parameter λ. The λ parameter is multiplied by the sum of the squared βs, thereby regularizing (i.e., penalizing) them. By adding this tuning

parameter, the βs need to make a large contribution to a more accurate prediction to ultimately reduce the SSE.

LASSO regression is similar to ridge regression, but rather than multiplying the tuning parameter by the sum of the squared β weights, the λ parameter is multiplied by the sum of the absolute values of the β weights. This change allows βs to assume a value of zero, thereby dropping them from the model. In this manner, the LASSO regression incorporates a variable selection feature such that only predictor variables that are significantly related to the criterion variable are retained, providing a parsimonious model that does not overfit the data.

The elastic net combines the tuning parameters of both ridge and LASSO models and is touted as the optimal modeling solution for linear relationships (Putka et al., 2018). The α parameter controls the extent to which the elastic net performs like a LASSO or a ridge regression; $\alpha = 0$ sets the model to a ridge regression, while $\alpha = 1$ sets the model to a LASSO regression. An α value in between allows the model to be considered "elastic" and mixes the two approaches.

Elastic net models include pockets of highly collinear predictor variables but regularize the parameters and distribute the β values among the predictors. This is a particularly important feature when examining item-level predictions in organizational sciences, as such items tend to be highly correlated (potentially causing multicollinearity issues) yet may still capture unique variance in outcomes. By handling multicollinearity in this manner, the elastic net offers a clear improvement over the LASSO model. By incorporating a variable selection feature, the elastic net offers a clear improvement over the ridge regression model. Therefore, it is a very useful analysis for examining scale- and item-level predictions in the organizational sciences.

Random Forests

Random forests are ensemble classification algorithms based on tree-structured classifiers, wherein a random subsample of the predictors is selected as candidate variables to split the data at each node (Breiman, 2001a). For example, a researcher building a model with 35 predictors may specify the random forest model to randomly select five predictor variables to serve as candidate variables to split at each node in the tree. Among those five variables, the variable that produces the split which best reduces the sum of squared errors is selected. In this fashion, fully-grown trees are developed. Typically, 500-1000 trees are fit using this procedure, and the predictions across these trees are aggregated in the final model. The number of variables to randomly subsample is the primary tuning parameter that must be optimized for this approach. Random forests address the bias-variance tradeoff by randomly subsampling the available predictors, thereby creating a wider range of diversity in the predictions generated by each tree in the forest. This diversity cancels out some noise in the data and prevents the model from overfitting.

Random forests have numerous advantages relative to linear models, such as OLS and the elastic net. Notably, random forests can account for nonlinear trends in the data and are also designed to automatically detect interactions among variables. In contrast, linear models are not suited to detect nonlinear trends and the researcher must specify interaction terms in the model (which may go otherwise unnoticed). Given these advantages, among others, research has empirically demonstrated that random forests are consistently among the most predictive machine learning algorithms available to researchers, often achieving a relatively high degree of predictive accuracy — outperforming more traditional approaches as well as other machine

learning algorithms (e.g., Caruana & Niculescu-Mizil, 2006; Couronné, Probst, & Boulesteix, 2018; Fernandez-Delgado, Cernadas, Barro, & Amorim, 2014).

However, one potential drawback of using the random forest approach is that the model is less interpretable and explainable, in comparison to linear models. This can create complications when deducing the practical and theoretical implications of a predictive model. Therefore, the current study included both algorithmic approaches (i.e., the elastic net and random forest) to maximize predictive accuracy and explore the interpretability of each model.

The Current Study

The current study aimed to further investigate the convergent, discriminant, and predictive validity of MTSJT scores. Specifically, it was proposed that MTSJT scores would account for more variance in supervisor ratings than scores on the SMTQ and MTI. Further, the predictive accuracy of MT scores was compared with that of grit, resilience, cognitive ability, and conscientiousness to help researchers determine if MT scores were distinct from scores of those assessments and if MT scores had incremental value when predicting workplace outcomes. These predictions were modeled at both the scale and item level to determine if unique variance was captured by item-level predictions.

As a result, the following four hypotheses were proposed. Hypothesis 1: All measures of MT, grit, resilience, and conscientiousness will be moderately correlated. Hypothesis 2: At both the scale and item level, scores on MT measures (and related constructs) will account for additional variance in supervisor ratings above that accounted for by cognitive ability and conscientiousness; Hypothesis 2a: MTSJT scores will be the strongest of these predictors. Hypothesis 3: More variance will be accounted for in supervisor ratings when predictor variables

are modeled at the item rather than scale level. Additionally, the study explored the relative predictive accuracy of the elastic net and random forest approaches.

Chapter 3

Methods

Participants

Five hundred and sixty-six employees served as participants (after cleaning the data, described in the Demographics section). Employee participants consisted of two samples: 1) students at a large Southeastern university who were employed at least part-time, and 2) full-time employees recruited online. Supervisor participants were recruited directly from the contact information provided by the employee upon completion of the survey battery. Usable evaluations were obtained from 122 supervisors. All student participants were compensated with course credit for their time; full-time employees (and all supervisors) were entered into a drawing for an online gift card.

Procedure

All employees provided informed consent, demographic information, and responses to the measures described below via Qualtrics Online Survey Software. Employees were informed that they were participating in a study aimed at further understanding the nature of workplace behavior. All employees were free to complete the survey at a time and location of their choosing and most completed it in 30-45 minutes. Upon completion, employees were partially debriefed and compensated for their time. Each employee provided the name and email address of their supervisor, who was contacted to gather ratings of job performance. Participants were assured that their responses would remain confidential and would not be released to their employer.

All supervisors provided informed consent and ratings of in-role task performance for their respective employees. Supervisors were assured that their evaluations of their employees will not be shared with anyone outside of the research team. Completion of the brief performance evaluation was designed to take less than five minutes.

Measures

Predictors

MTSJT. This scenario-based assessment of MT consists of 11 scenarios and a total of 21 response options (given in Appendix A). Participants were instructed to read each scenario carefully and then rate each of the following three response options on a scale of 1 (extremely unlikely) to 7 (extremely likely), indicating the likelihood they would engage in each response option. Each response option was designed to represent one of the three dimensions of the SCMT — task persistence, utilization of feedback, and emotional control.

Although each scenario contained three response options, twelve response options were not scored because these responses did not fit the measurement model in previous research (Flannery et al., 2019). As a result, task persistence was assessed via ten response options; utilization of feedback was measured with eight response options; emotional control was assessed with three response options. Further, an overall scale score was constructed by averaging across responses.

The MTSJT assesses both the behavioral expectations and motivations underpinning MT. Eight scenarios provide an example of a workplace issue, and ask the respondent, "What is the likelihood you would engage in each of the following behaviors?" Further, three scenarios provided an example of a workplace issue and included a behavior in the item stem, and ask the respondent, "What is the likelihood you would think and feel in each of the following manners?" thereby assessing motivational aspects of MT. It is important to note that a behavioral intention response format was used (i.e., "would") in favor of a context knowledge response format (i.e., "should"). In this manner, the tool assesses behavioral intentions rather than knowledge.

MTI. Gucciardi et al.'s (2015) eight-item MTI was also included (as provided in Appendix B). This measure asks participants to indicate the extent to which they believe each statement is true of themselves, from 1 = false, 100% of the time, to 7 = true, 100% of the time. This assessment yields a unidimensional MT score by averaging all eight items.

SMTQ. The 14-item SMTQ is provided in Appendix C (Sheard et al., 2009) and was included as another measure of MT. This scale measures three components of MT — confidence, emotional, control, and constancy in achieving goals — on a Likert scale ranging from one (not at all true) to four (very true). By averaging items, three subscale scores were generated in addition to an overall MT score.

Grit Scale. Duckworth et al.'s (2007) twelve-item Grit Scale (provided in Appendix D) was included. This measure assesses grit as a combination of two factors — passion and perseverance for long-term goals. This tool asks respondents to indicate (on a scale from one to five) the extent to which the statements provided are descriptive of them. Total grit scores were created by averaging across all twelve items, but subscale scores of passion and perseverance are also derived.

Brief Resilience Scale. The six-item Brief Resilience Scale (BRS) provided in Appendix E (Smith et al., 2008) assesses the ability to bounce back from distress. This assessment tool asks participants to indicate the extent to which (one - strongly disagree, to five - strongly agree) resilience-related statements are descriptive of them. A total resilience score was created by averaging all six items.

Conscientiousness. Conscientiousness was assessed using the ten-item International Personality Item Pool measure (IPIP), provided in Appendix F (Goldberg, 1999). This measure asks respondents to indicate the extent to which (one - very inaccurate, to five - very accurate) conscientiousness-related statements are true of them. An overall conscientiousness score was created by averaging all ten items.

Cognitive ability. To assess cognitive ability, an eleven-item matrix-reasoning task (ICAR; The International Cognitive Ability Resource Team, 2014) was included but not appended due to copyright privacy. This task provides participants with a series of 3x3 matrices containing 8 geographic shapes and one missing cell. Participants are then provided with six geometric shapes as response options and asked to select which of the six shapes fits the pattern in the matrix. The items progress in difficulty to identify individuals with different underlying levels of cognitive ability.

In a series of studies, Condon and Revelle (2014) demonstrated the internal reliability and the convergent and discriminant validity of ICAR assessments in unproctored settings, suggesting they are valid measures of cognitive ability. Specifically, the researchers found that ICAR items were saturated by a general factor, demonstrated consistent factor loadings, correlated strongly with existing commercial measures of cognitive ability and standardized student achievement scores, among other measures. Overall scores are generated by summing the number of correct responses.

Supervisor Ratings

Supervisors were asked to rate employees using Williams and Anderson's (1991) seven-item measure of in-role performance, given in Appendix G. Employees were rated from one (strongly disagree) to seven (strongly agree) on a series of descriptive statements. A total score of job performance resulted from averaging across all items.

Analyses

The analyses for this study were conducted in two stages. First, a confirmatory factor analysis (CFA) was conducted on the MTSJT. Next, both elastic net and random forest models were used to predict job performance, and two rounds of analyses were conducted — one at the item level and one at the scale level. The procedure outlined here was adapted from Putka et al.'s (2018) approach. As Putka et al. noted, machine learning algorithms such as elastic nets and random forests are still considered novel to many organizational researchers, and therefore an in-depth review is provided here.

In Step 1, the data were split into a training set used to fit the model parameters (consisting of 80% of the overall *N*) and a test set, used for validation (consisting of the

remaining 20%). Step 2 optimized model tuning parameters (for both the elastic net and random forest) via ten-fold cross-validation, using only the training data. In the process of ten-fold cross-validation, the training data were split into ten random partitions. Using data from partitions one through nine, a series of elastic net and random forest models were fit using all possible combinations of the tuning parameter values[1]. Then, these models were used to predict values in the tenth partition, and *RMSE* values were recorded. Subsequently, a series of models were fit again using partitions one through eight and partition ten, then used to predict values in the ninth partition. This was repeated until all partitions served as the holdout sample.

The cross-validation process was repeated three times since the splitting of data into partitions was arbitrary. Thus, this process yielded thirty *RMSE* values for each possible combination of tuning parameters. Due to sampling error and the need to avoid overfitting, researchers suggest selecting the most parsimonious model that produced an average RMSE within one standard error of the lowest average *RMSE* value yielded by the ten-fold cross-validation (Breiman, Friedman, Stone, & Olshen, 1984). In this manner, Step 2 yielded optimal values for the tuning parameters for the elastic net (α and λ) and random forest (m – the number of variables to subsample at each node).

In Step 3, the entire training data set was used to estimate the final model parameters (before validating the model on the test data). That is, using the optimal tuning parameters determined from Step 2, a full model was fit to the training data to generate βs for each predictor

[1] In the current study, α ranged from .01 to .99 (testing 50 values), while λ ranged from .1 to 10 (testing 100 values). The rule of thumb for m in a random forest model is to subsample a third of the available predictors ($P/3$). However, for each model the range for m began with two and ended with the maximum number of variables in the model, testing every other value of m. The exact values tested is contingent upon the number of predictors, which varies based on each model in the current study.

variable (for the elastic net) and decision criteria at each node for the random forests. In Step 4, the full model from Step 3 was fit to the test data and the *RMSE*, R^2, and variable importance metrics[2] were recorded. Finally, for Step 5 the entire process was repeated until all Steps 1-4 had been followed 100 times. This was done because of the arbitrary nature of splitting the full *N* into training and test data sets and to provide more stable estimates of the model parameters and performance.

To determine the incremental validity of MT, grit, and resilience scores in predicting supervisor ratings, this process was followed first using only cognitive ability and conscientiousness scores. Then, using the same split to the data (i.e., the same training and test data), the model was trained and tested multiple times by adding the respective MT, grit, or resilience scores to the model. Scores from the various MT, grit, and resilience scales were not entered into the models together — they were each tested independent of each other to determine each's unique contribution beyond cognitive ability and conscientiousness.

Incremental validity of each scale was tested by determining if the average R^2 value of the models containing MT, grit, and resilience scores was significantly higher than the average R^2 value of the models containing only the cognitive ability, and conscientiousness scores. Similarly, models were also compared to determine if the average *RMSE* values differed significantly. Finally, this entire process was repeated at both the scale and item level. All

[2] For the elastic net model, the variable importance is the absolute value of the *t*-statistic associated with the regression coefficient. Variable importance for the random forest model is calculated by computing the difference between the final model's sum of squared errors and the sum of squared errors calculated after permuting (i.e., randomizing observed values in the data) the variable of interest. For both the elastic net and random forest approach, the variable importance metrics was scaled to range from 0-100.

analyses were conducted in R (R Core Team, 2020). Specifically, predictive analyses were conducted using the caret package (Kuhn, 2008).

Chapter 4

Results

Demographics and Data Cleaning

Six hundred and seventy-six employed individuals completed the initial assessment battery. Of these individuals, 110 (16.27%) were removed for failing to pass both attention checks. Of the remaining individuals, 534 (94.34%) were students at a large Southeastern University, and 26 respondents (4.59%) were full-time employees recruited via a snowballing procedure. Six individuals (1.06%) did not indicate their student status. As shown in Table 5, this sample served as the CFA sample to examine the psychometric properties of the MTSJT.

Table 5

Age, Gender, Student Status, Hours Worked, and Ethnicity of Each Sample

	CFA Sample	Predictive Sample
Age in Years *M*(*SD*)	20.58 (4.43)	21.78 (6.33)
Gender (%)		
Female	457 (80.74)	98 (80.32)
Male	104 (18.37)	24 (19.67)
Other	2 (0.00)	0 (0.00)
Prefer not to respond	1 (0.00)	0 (0.00)
Student Status		
Yes	534 (94.34)	110 (90.16%)
No	26 (4.59)	12 (9.84%)
Hours worked per week	18.27 (11.73)	18.7 (13.91)
Race (%)		
White	425 (75.09)	97 (79.51)
Hispanic or Latino	30 (5.30)	8 (6.56)
Black or African American	25 (4.42)	7 (5.73)
Native American	4 (0.01)	1 (0.82)
Asian/Pacific Islander	59 (10.42)	5 (4.10)
Other	23 (4.06)	4 (3.28)
N	566	122

Two hundred and thirty-nine (42.22%) employees provided usable emails to contact their immediate supervisors. Of these, 123 (51.46%) supervisors completed the employee evaluation. One evaluation was completed for an employee who failed the attention checks, and therefore was removed. This left 122 employee-supervisor dyads to be analyzed in the predictive analyses. The employee demographics of the prediction sample are also presented in Table 5. As can be seen by comparing the demographics of the samples, the predictive sample was generally representative of the full sample, with no large differences in age, gender, hours worked, student status, or ethnicities.

The occupations of the 122 employees in the predictive sample varied widely. Broadly speaking, employees represented the service industry (e.g., baristas, waitresses, cashiers), business professionals (e.g., salesmen, financial analysts, accountants), interns, administrative assistants, and student-workers (e.g., equipment managers, tutors), among many other areas. Given that the majority of respondents were students, there were more entry-level positions, internships, and student-workers than professionals. However, given the wide range of occupations represented within the relatively small sample size, there was no systematic manner to account for occupational type.

Confirmatory Factor Analysis of MTSJT

A CFA was conducted to confirm the factor structure of the MTSJT reported by Flannery et al. (2019). Missing data (0.3% of full data) were imputed using maximum likelihood imputation. Mahalanobis distance values indicated that 39 observations (7%) exceeded the critical value of 46.79 for 21 degrees of freedom and were considered multivariate outliers. However, further inspection of these observations indicated these respondents were genuine responses because they had passed the attention checks and exhibited normal response times (i.e., were not careless responders). All MTSJT items had univariate skew values between negative two and two, and univariate kurtosis values between negative seven and seven. However, Mardia's test of multivariate normality revealed that items were not multivariate normal. Therefore, maximum likelihood estimation with robust standard errors was used.

The hypothesized three-factor structure fit the data well, $\chi 2$ (186) = 603.31, $p < .01$, *CFI* = .84, *RMSEA* = .06, *SRMR* = .06 (Hu & Bentler, 1999). Strong factor correlations were evident, with correlations of .77 between task persistence and emotional control, .71 between task persistence and utilization of feedback, and .74 between emotional control and utilization of

feedback. While these strong correlations could suggest a unidimensional model, the three-factor solution provided a better fit to the data than the unidimensional solution, $\chi 2$ (189) = 778.57, $p <$ .01; *CFI* = .77, *RMSEA* = .07, *SRMR* = .06. Given this empirical support in conjunction with the theoretical rationale supporting the three-factor solution (Flannery et al., 2019), the three-factor model was retained over the unidimensional model.

Factor loadings from the three-factor solution are presented in Table 6. The overall mean factor loading was .49, and the means for each of the subfacets were: .52 for task persistence; .32 for emotional control; and .53 for utilization of feedback. It is noteworthy that the mean factor loading for the emotional control dimension was influenced by one item (EC 3) with a factor loading of .67 and the factor loadings for the other two items were at or below .15. The modification indices for these items were small (i.e., <= 9.52), suggesting these items were not indicators of either task persistence or utilization of feedback.

Table 6

Standardized Factor Loadings for the MTSJT

	Factor		
Item	TP	UF	EC
Scenario 1			
TP 1	.56	-	-
EC 1	-	-	.14
Scenario 2	-		
TP 2	.57	-	-
EC 2	-		.15
Scenario 3		-	
EC 3	-	-	.67
TP 3	.52	-	-
UF 2	-	.60	-
Scenario 4			
UF 3	-	.06	-
Scenario 5			
TP 4	.48	-	-
UF 4	-	.61	-

Scenario 6			
UF 5	-	.37	-
TP 5	.48	-	-
Scenario 7			
TP 6	.45	-	-
UF 6	-	.69	-
Scenario 8			
TP 7	.51	-	-
UF 7	-	.68	-
Scenario 9			
UF 8	-	.72	-
TP 8	.46	-	-
Scenario 10			
TP 9	.55	-	-
Scenario 11		-	
UF 9	-	.47	-
TP 10	.59	-	-

Note: * significant at .01 level; TP = Task Persistence, EC = Emotional Control, UF = Utilization of Feedback.

Although the factor loadings for some of these items were low, all items were retained when calculating scale scores and conducting the predictive analyses for several reasons: 1) the global fit for the model was good; 2) SJT items are not typically known to have high factor loadings (Whetzel & McDaniel, 2009); and 3) prior research has provided theoretical and empirical evidence supporting the three-factor MTSJT model. Further, as an exploratory analysis, scale scores without the four lowest-loading items were calculated. The reduced MTSJT scores and the UF subscale scores correlated with the respective original scores (using all items) at or above .97, while the EC 3 item correlated with the three-item EC subscale score at .67. Thus, it appeared the lowest-loading items were not drastically influencing the scores, especially for the overall MTSJT score and the UF subscale. Given this, in conjunction with the theoretical rationale, all items were retained so the scale-level and item-level predictive analyses could be compared directly.

Scale-level Descriptive Statistics

Data Cleaning

Scale scores were calculated using responses from the 122 employees whose supervisors provided usable evaluations. Among the predictor variables, Mardia's test for multivariate normality found the scores to be multivariate normal in terms of skew, but not kurtosis at both the scale and item-level. While multivariate normality is an assumption of the elastic net, it is not required for a random forest. Given this, in conjunction with the fact that the data were multivariate in terms of skew values, no transformations were conducted.

Mahalanobis values calculated using all predictor variables suggested that nine observations (7%) exceeded the critical value of 166.41 for 120 degrees of freedom and were considered multivariate outliers. Upon further inspection, these observations were deemed genuine responses because they had passed the attention checks and exhibited normal response times (suggesting they were not careless responders). For the predictive analyses, missing data were imputed using a bagged tree algorithm. Bagged tree algorithms generate predictions by creating bootstrapped samples of data, fitting decision trees to each bootstrapped sample, then aggregating the predictions of each tree to create a final prediction for each missing value. This imputation method is known for its predictive accuracy (Kuhn, 2008). At the item level, only 0.7% of the data were missing. Finally, before running the predictive models, scores were also mean-centered and scaled such that each predictor value was divided by its standard deviation.

In general, the performance appraisals were high with a mean of 6.33 and a standard deviation of .71. Observed scale scores ranged from 3.14 to 7, despite the scale anchors ranging from 1 to 7. Figure 2 depicts the distribution of this variable. Examination of this figure shows a degree of negative skew and leptokurtosis. However, univariate skew and kurtosis values

indicated that the skew and kurtosis values fell within an acceptable range, with a skew value of -1.88 and a kurtosis value of 5.20 (Griffin & Steinbrecher, 2013). As an exploratory analysis, several transformations were conducted to reduce skew and/or kurtosis, including square root, cube root, and logarithmic transformations. For each of these transformations, the skewness and kurtosis values increased. This was likely because the mode of the distribution (seven) was also the maximum of the scale. In sum, although there was not a large degree of variance in job performance, the skewness and kurtosis values could not be improved via any common transformations. Therefore, the original job performance scores were used for all analyses.

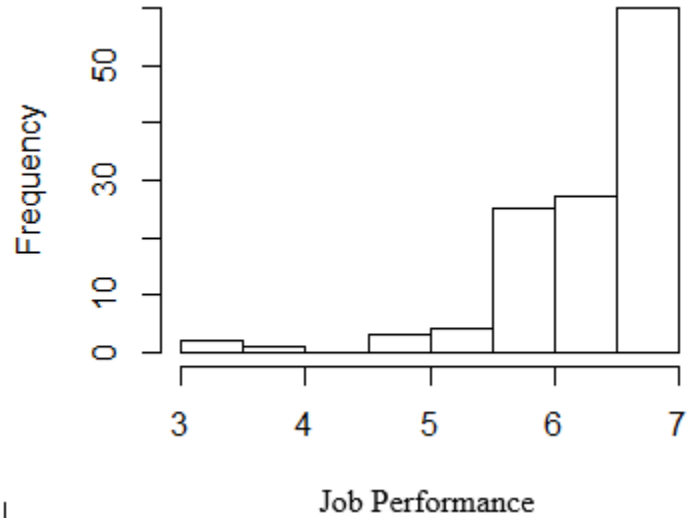

Figure 2. Histogram of job performance scores. Visual inspection of this figure suggests the scores are negatively skewed and leptokurtotic. However, the skewness and kurtosis values fell within a normal range for linear modeling. Further, square root, cubic, and logarithmic transformations could not decrease the skew or kurtosis in the distribution.

Correlations between Predictors and Job Performance/Hours Worked.

Table 7 shows the descriptive statistics, correlations, and coefficient alpha values for the scale-level scores. No scales correlated with job performance at the .05 level. Interestingly, job performance correlated negatively with cognitive ability (r = -.17, p = .052), conscientiousness

(r = -.14, p = .11), and resilience (r = - .13, p = .15). Although not significant, the correlations between job performance and cognitive ability/conscientiousness were particularly surprising, given that the positive validities of aptitude and conscientiousness scores are the only two constructs shown to generalize to all jobs (e.g., Barrick et al., 2001, Hunter & Hunter, 1984). Figures 3 and 4 provide the scatterplots of both cognitive ability and conscientiousness as they relate to job performance.

Table 7

Means, Standard Deviations, Internal Consistency, and Correlations among Scale Scores

		Mean	SD	1	2	3	4	5	6	7	8
1	Job Performance	6.33	0.71	.73							
2	Cognitive Ability	6.08	2.41	-.17	-						
3	Con.	3.82	0.55	-.14	.11	.82					
4	SMTQ Total	2.83	0.40	-.02	.20*	.55*	.80				
5	SMTQ Constancy	2.91	0.48	.00	.17	.37*	.82*	.68			
6	SMTQ Control	3.21	0.43	-.05	.15	.68*	.79*	.47*	.62		
7	SMTQ Confidence	2.26	0.65	-.02	.15	.24*	.75*	.43*	.40*	.69	
8	MTI	5.60	0.72	-.04	.18*	.58*	.66*	.56*	.64*	.34*	.81
9	Grit Total	3.80	0.56	-.03	.14	.57*	.46*	.33*	.60*	.14	.51*
10	Grit Passion	3.39	0.69	.05	.13	.25*	.12	.07	.22*	-.01	.13
11	Grit Perseverance	4.21	0.73	-.05	.10	.65*	.59*	.44*	.72*	.22*	.67*
12	BRS	3.40	0.78	-.13	.13	.37*	.57*	.44*	.40*	.51*	.38*
13	MTSJT Total	5.83	0.52	.07	.04	.38*	.19*	.17*	.37*	-.11	.42*
14	MTSJT TP	5.71	0.72	.08	-.04	.32*	.14	.09	.32*	-.09	.26*
15	MTSJT EC	5.42	0.96	.03	.01	.14	-.05	.03	.03	-.19*	.17
16	MTSJT UF	6.13	0.53	.03	.13	.35*	.25*	.23*	.37*	-.04	.49*

Table 7 (continued)

		9	10	11	12	13	14	15	16
9	Grit Total	.71							
10	Grit Passion	.79*	.65						
11	Grit Perseverance	.80*	.26*	.69					
12	BRS	.24*	.08	.29*	.87				
13	MTSJT Total	.29*	.09	.36*	-.03	.84			
14	MTSJT TP	.21*	.03	.29*	.00	.87*	.77		
15	MTSJT EC	-.04	-.13	.05	-.12	.56*	.36*	.22	
16	MTSJT UF	.36*	.20*	.37*	-.02	.77*	.40*	.30*	.68

Note: *N* = 122. Coefficient alpha values are along the diagonal; SMTQ = Sports Mental Toughness Questionnaire; MTI = Mental Toughness Index; Con = Conscientiousness; BRS = Brief Resilience Scale; MTSJT = Mental Toughness Situational Judgment Test; TP = Task Persistence; EC = Emotional Control, UF = Utilization of Feedback

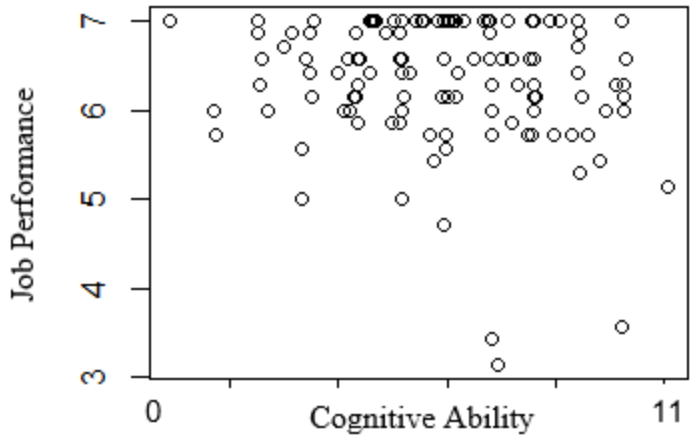

Figure 3. Scatterplot between cognitive ability scores and job performance scores. Three individuals were identified as outliers as they scored high on cognitive ability yet were rated relatively low on job performance, and one individual was identified as an outlier because s/he scored low on cognitive ability and received the maximum performance rating possible.

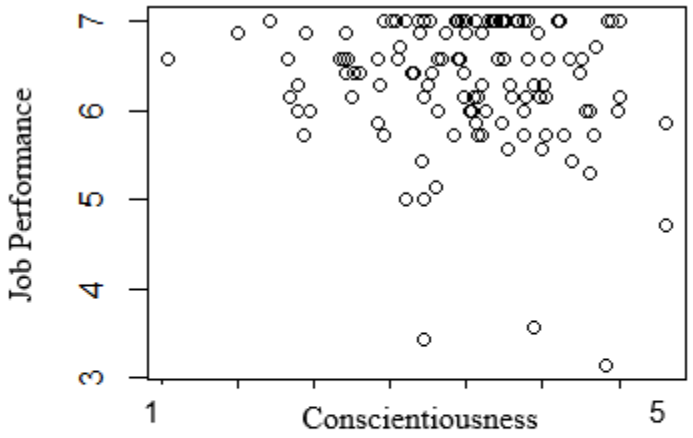

Figure 4. Scatterplot between conscientiousness and job performance scores. Three individuals were identified as outliers as they scored high on conscientiousness yet were rated relatively low on job performance, and one individual was identified as an outlier because s/he scored low on conscientiousness and received the maximum performance rating possible.

As can be seen in these plots, three individuals scored high on both cognitive ability/conscientiousness yet received relatively low job performance ratings. One individual received the maximum possible job performance rating despite scoring very low on cognitive

ability/conscientiousness. When these individuals were removed from the sample, the correlation between job performance and cognitive ability was reduced to r = -.10, p = .27; the correlation between job performance and conscientiousness reduced to r = -.13, p = .11. However, given the fact that the original correlations did not reach significance and the fact that these individuals were deemed to be genuine responses (as described above), they were retained in the predictive sample. Potential explanations regarding the relationship between cognitive ability/conscientiousness and job performance are explored in the Discussion section.

The number of hours an employee worked each week was unrelated to every scale score except for the SMTQ Confidence subscale score (r = .20, p = .03). Thus, because hours worked were unrelated to job performance ratings and nearly every other predictor and there was no strong theoretical rationale for retaining the variable, the number of weekly hours worked was omitted as a control variable.

Hypothesis 1

The first hypothesis was that measures of conscientiousness, MT, grit, and resilience will correlate moderately. Table 7 provides the correlations among the scale scores. With one exception, all correlations among overall scale scores for these measures were moderate to high, ranging from .19 to .66 and significant at the .05 level. The exception was the correlation between the overall MTSJT score and the BRS score (r = -.03, p = .72). Overall, the average correlation among the scale-level MT, grit, and BRS scores was .39. The average correlation among the overall MTSJT, MTI, and SMTQ scores was .42. MTSJT scores correlated strongest with MTI scores (r = .42, p < .001) but also correlated similarly with conscientiousness scores (r = .38, p < .001).

Further, the MTSJT scores were weakly correlated with SMTQ (r = .19, p = .03), moderately with grit (r = .29, p < .001) scores, but were unrelated to BRS scores. Moreover, the average among all total and subscale scores of MT measures was .27. Further examination of Table 7 shows that the correlations between MTSJT scores and all other self-report measures were weaker than the correlations between other MT scales and all other self-report measures, supporting Flannery et al.'s (2019) conclusion that the MTSJT avoided common-method bias in responding due in part to the difference in response format.

With few exceptions, these results generally provided support for Hypothesis 1, which stated that measures of MT, grit, resilience, and conscientiousness would be moderately correlated. These results provided evidence of convergent validity among the measures. However, the results also provided evidence of the jingle and jangle fallacy surrounding MT scores. That is, not only were scores on all MT, grit, resilience, and conscientiousness measures moderate-to-strongly correlated (i.e., the jangle fallacy), but scores on measures of MT did not consistently correlate stronger with other MT measures than they did with measures of grit, resilience, and conscientiousness (i.e., the jingle fallacy). For example, MTI scores correlated with Grit – Perseverance scores (r = .67) to a slightly stronger degree they did with overall SMTQ scores (r = .66) and much stronger than with MTSJT scores (r = .42). In sum, Hypothesis 1 was supported.

Predictive Analyses

Hypothesis 2

Hypothesis 2 proposed that measures of MT, grit, and resilience would demonstrate incremental validity in predicting job performance over cognitive ability and conscientiousness.

As mentioned above, unexpected negative relationships were found between job performance and cognitive ability/conscientiousness. However, these scores were retained in the predictive analyses for several reasons.

First, the correlations were weak overall and both correlations failed to reach significance at the .05 level. Second, in the current sample, conscientiousness and cognitive ability scores may be related to job performance ratings in a nonlinear fashion. This relationship would not be captured by the correlation but could be captured by a random forest algorithm. Previous research has suggested that conscientiousness may be related to job performance in a nonlinear fashion in certain occupations (LaHuis, Martin, & Avis, 2005; Le et al., 2010), and that nonlinear relationships between cognitive ability and job performance can occur on occasion (Hawke, 1970). Third, the correlations were based on scale scores, and it could be the case that items were more strongly related to job performance (either in linear or nonlinear fashion). Therefore, the control models remained unchanged for examining Hypothesis 2.

Tables 8 and 9 summarizes the average changes in R^2 and *RMSE* values obtained for each measure across 100 iterations of training the models on a development sample and testing them on a holdout sample, broken down by both random forest and elastic net approaches. Table 8 presents the scale-level analyses, and Table 9 presents the item-level analyses. The "Control" row for each of these tables represents the initial R^2 and *RMSE* values when predicting job performance first using cognitive ability and conscientiousness scores. For each subsequent row, the change in R^2 and *RMSE* values upon adding each model's respective predictors is shown.

Table 8

ΔR² and ΔRMSE in Test Data for Scale-level Models

	Elastic Net		Random Forest	
Model	*ΔR²*	*ΔRMSE*	*ΔR²*	*ΔRMSE*
Control	.02	.73	.01	.81
MTSJT	-.01	.00	.00	-.03*
MTI	.00	.00	.01	-.02
SMTQ	.00	.00	.00	-.03*
GRIT	.00	.00	.01	-.02
BRS	.00	.00	.00	-.02

Note: * significantly different than respective (random forest/elastic net) control model; MTSJT = Mental Toughness Situational Judgment Test; MTI = Mental Toughness Index; SMTQ = Sports Mental Toughness Questionnaire; BRS = Brief Resilience Scale; *N* = 122.

Table 9

ΔR² and ΔRMSE in Test Data for Item-level Model

	Elastic Net		Random Forest	
Model	*ΔR²*	*ΔRMSE*	*ΔR²*	*ΔRMSE*
Control	.01	.73	.02	.74
MTSJT	.00	.00	.01*	-.01
MTI	.00	.00	.01*	-.01
SMTQ	.00	.00	-.01	-.01
GRIT	.00	.00	-.01	.00
BRS	.00	.00	.00	-.01

Note: * significantly different than respective (random forest/elastic net) control model; MTSJT = Mental Toughness Situational Judgment Test; MTI = Mental Toughness Index; SMTQ = Sports Mental Toughness Questionnaire; BRS = Brief Resilience Scale; *N* = 122.

As can be seen in Table, 8 the scale-level MT, grit, and BRS models did not provide significant increases in R^2 values relative to the control models. However, the MTSJT and SMTQ random forest models significantly reduced the *RMSE*, although no other measures did. None of the scale-level elastic net models increased the R^2 values or reduced the *RMSE*. Overall, at the scale-level these results provided mixed support for Hypothesis 2 in that the magnitude of the errors produced by the scale-level random forest MTSJT and SMTQ models was smaller in

comparison to the control model (demonstrated by the decrease in *RMSE*), but these models did not account for significantly more variance than the control model.

As shown in Table 9, the item-level analyses also provided mixed support for Hypothesis 2, such that the MTSJT and MTI random forests provided a significant increase in R^2 relative to the control model, but no other measures did (regardless of using an elastic net or random forest approach). Further, none of the item-level MT, grit, or BRS models significantly reduced the *RMSE* relative to the control model. Overall, none of the item-level elastic net models significantly increased the R^2 or reducing the *RMSE*. Thus, although the item level MTSJT and MTI models significantly increased the R^2 values relative to the control, they did not reduce the *RMSE*.

Hypothesis 2a proposed the MTSJT models would be the best predictors among all measures of MT, grit, and resilience. As mentioned above, partial support for this hypothesis was found at the item-level such that the addition of MTSJT items accounted for a significant amount of incremental variance over the control items. Further, only one other model showed incremental variance over the control items — the MTI random forest model. At the scale level, both the MTSJT and SMTQ random forests significantly reduced *RMSE* relative to the control scale random forest, but the MTI, grit, and BRS scales did not. The MTSJT was the only scale to provide incremental prediction at both the item and scale level. Further, the item-level MTSJT random forest model was the top overall performing model (tied with the item-level MTI random forest) across all measures and algorithms, thereby supporting Hypothesis 2a.

Hypothesis 3

Partial support for Hypothesis 3 was found, as item-level models (regardless of the measures included or the algorithmic approach) had lower *RMSE* values (t (2397.2) = 6.74, p < .001) compared to the scale-level models but failed to account for significantly more variance in outcomes (t (2097.1) = 1.36, p = .17. Comparing Tables 8 and 9 shows that in nearly all cases, the average item-level R^2s were equal to or higher than the respective average scale-level R^2*s*, and the average item-level *RMSEs* were equal to or lower than respective average scale-level *RMSE* values. As mentioned previously, the two top-performing models overall were item-level MTI and MTSJT random forests. Ultimately, these results provided partial support for Hypothesis 3.

Random Forests vs. Elastic Nets

An additional aim of this study was to determine whether random forests or elastic nets provided the best prediction of job performance. Globally speaking, the mean R^2 values were .02 and .01 for the random forests and elastic nets, respectively. This difference was significant, t (2178.5) = 3.43, p < .01. At the item level, random forests had significantly higher R^2 values, but at the scale level, elastic nets had significantly higher R^2 values. Regarding the overall *RMSE*, elastic nets had significantly lower *RMSE* values t (2395.9) = 6.97, p < .01. At the item level, there was no difference between elastic nets and random forests regarding *RMSE* values, but at the scale level, elastic nets had significantly lower *RMSE* values. Thus, the random forest was superior at the item level, but the elastic net was superior at the scale level. However, the random forest produced the two best-performing models — the MTSJT item-level model and the MTI-item level model. Across all models (i.e., algorithm, construct, scale v. item level) these two

models had the absolute highest R^2 values and the lowest *RMSE* values, suggesting the random forest achieved the best prediction in these data.

Summary of Best Performing Models

Because both the MTI and the MTSJT item-level random forests produced the most accurate predictions, these models were explored further to examine the training performance, tuning parameters, and variable importance metrics associated with each model. Interpretations for all other models were not conducted — not only because of the computational burden, but also because attempting to interpret models that suffer in predictive accuracy can lead to erroneous conclusions (Breiman, 2001b).

Training Performance

As stated previously both the item-level MTI and MTSJT random forests accounted for 3% of the variance in job performance ratings and had *RMSE* values of .73 when used to predict values in the test set. For the training set — used to develop both of these models — both the item-level MTI and MTSJT random forests had average R^2 values of .19. Further, the item-level control random forest had an average R^2 value of .19. Similarly, each of these item-level random forests (control, MTI, and MTSJT) had *RMSE* values of .63 in the training sample. This suggests that the control model (with fewer items) performed identical to the item-level random forest MTSJT and MTI models in the training set but performed worse in the test set. These results indicated: a) the degree of attenuation in predictive accuracy when comparing training and test predictions, and b) the importance of tuning parameters for balancing the degree of underfit vs. overfit in the data. By properly tuning the model, an increase in the number of items was able to

discern systematic variance from noise when training the models, which allowed for better performance (relative to the control model, with fewer items) upon validation in the test set.

Tuning Parameter Values

As mentioned previously, tuning parameters are important model features that allow the algorithms to balance underfitting vs. overfitting the training data to provide more accurate predictions in future data. Tuning parameters for the top-performing models, the item-level MTI and MTSJT random forests were examined, along with the control item-level random forest. The item-level control elastic net tuning parameters were also examined, as doing so could provide insight into why the model was unable to predict as well as the random forests.

As mentioned previously, the tuning parameter for a random forest, *m*, indicates the number of variables subsampled at each node. Among the subsampled variables, the variable that produces the greatest reduction in the sum of squared errors is used in the model. For each model, *m* ranged from two to the number of items in the model. For the item-level MTSJT model, the ten-fold cross-validation process selected $m = 2$ in each of the 100 iterations. Similarly, for the item-level MTI model, the process selected $m = 2$ in 99 of the 100 iterations; $m = 6$ was selected in one iteration. Finally, for the control item-level random forest, the process also selected $m = 2$ in 99 of the 100 iterations, with one iteration selecting $m = 4$. In sum, these results suggest that $m = 2$ was the optimal tuning parameter for the control, MTSJT, and MTI random forests to balance under and overfit in training data and ultimately provide more accurate predictions in future data.

Because of the importance of tuning parameters to model performance, the tuning parameters of the item-level control elastic net were examined to determine how these

parameters were influencing the model. The first parameter, α, determines the extent to which the elastic net functions like a ridge regression (when $\alpha = 0$) or a LASSO (when $\alpha = 1$), while the second tuning parameter, λ, penalizes coefficients based on the degree of multicollinearity among the predictors. In every iteration, ten-fold cross-validation selected alpha = .01, which was the lower bound. This suggests that the algorithm was striving to perform more like a ridge regression (addressing multicollinearity among variables) than a LASSO (by conducting variable selection). Further, the average λ value was 9.87, which neared the upper bound of 10. This suggests there was a great degree of multicollinearity in the model which the elastic net attempted to mitigate by increasing the lambda penalty. Because this parameter frequently reached the maximum bound, the results may differ if a higher bound were set to allow the algorithm to further address multicollinearity.

Variable Importance for Top-performing Models

Tables 10 and 11 show the average variable importance metrics for the item-level random forest MTSJT and MTI models, respectively. As mentioned above, variable importance for the random forest is calculated by permuting the variable of interest (i.e., shuffling values across observations), then determining the change in predictive accuracy for the model. Large changes in predictive accuracy upon permuting suggest the variable made an important contribution to the model performance. These metrics are scaled from 0-100, such that 100 indicates maximum influence and 0 indicates no influence. Variable importance metrics were calculated independently during each of the 100 iterations, then averaged to determine which variables had the greatest degree of influence.

Table 10

Average Variable Importance for Item-level Random Forest MTSJT Model

Item	*M*	*SD*
Con 7	76.43	23.08
TP 8	65.00	18.37
TP 1	64.51	20.33
TP 6	63.04	17.34
UF 2	62.88	13.39
TP 4	60.28	23.02
EC 2	57.80	16.76
TP 7	56.20	17.91
TP 2	54.38	12.93
Con 9	53.95	17.44
G7	53.31	27.21
Con 8	49.71	13.22
EC 1	48.69	16.23
EC 3	47.66	19.23
UF 5	46.37	28.88
TP 10	43.63	17.50
Con 5	43.17	17.41
Con 3	43.00	16.54
Con 4	40.94	20.19
G 9	39.45	16.02
Con 1	39.26	15.05
Con 6	37.52	12.46
UF 3	37.26	34.70
G 8	33.95	14.02
TP 5	33.54	10.88
G 10	32.12	9.70
Con 10	28.10	11.16
TP 3	26.97	11.42
G 11	24.25	12.47
UF 6	24.20	10.41
UF 7	23.87	9.66
G 6	22.77	9.55
TP 9	22.67	9.83
Con 2	21.80	10.06
UF 4	21.10	9.02
G 3	19.41	8.60
UF 8	18.42	9.39

UF 1	18.07	8.62
G 4	16.54	11.31
G 5	13.96	8.08
G 2	12.50	8.12
Student Status	11.63	13.49
G 1	6.51	6.96

Note: *N* =122; Con = conscientiousness, TP = Task Persistence, EC = Emotional Control, UF = Utilization of Feedback, G = cognitive ability.

Table 11

Average Variable Importance for Item-level Random Forest MTI Model

Item	*M*	*SD*
Con 7	76.49	22.14
MTI 2	74.01	23.78
MTI 5	62.19	23.01
MTI 1	59.67	18.58
MTI 3	57.81	19.02
G 7	54.16	28.05
MTI 6	53.03	21.96
Con 9	51.14	18.82
MTI 8	50.25	18.1
Con 8	47.6	12.07
Con 3	47.26	21.26
Con 5	43.19	16.04
Con 6	38.78	13.99
Con 1	38.66	16.5
G 9	37.73	14.65
G 8	37.26	18.23
MTI 7	37.11	15.82
Con 4	36.93	16.77
G 10	30.52	10.54
Con 10	29.2	11.31
G 6	23.5	12.18
MTI 4	22.53	8.43
Con 2	21.11	10.82
G 11	19.76	10.75
G 3	16.38	8.23
G 4	14.88	12.16
G 5	14.45	8.91
Student Status	10.98	12.36
G 2	10.72	9.61
G 1	6.1	7.86

Note: N =122; Con = conscientiousness, MTI = Mental Toughness Index, G = cognitive ability.

Examining Table 10 shows the MTSJT (particularly TP items) and conscientiousness items were the most predictive variables in the model. While item Con 7 ("I find it difficult to get down to work") was the top overall predictor, eight of the top ten variables were MTSJT items. Examining the three most predictive items in the model, item-level correlations suggested that although not significant, job performance was negatively correlated with Con 7 (r = -.14, p = .13), unrelated to TP 8 (r = -.01, p = .92), and positively correlated with TP 1 (r = .15, p = .11). These results, in conjunction with the finding that the item-level MTSJT random forest consistently outperformed the item-level MTSJT elastic net, suggest that although these items were not related to job performance ratings in a linear fashion, there were systematic nonlinear relationships with job performance and important interactions that influenced prediction. These relationships are explored further below. Importantly, no cognitive ability items ranked within the top ten items in terms of variable importance. Thus, it appears that MTSJT and conscientiousness items were driving the predictions within the item-level MTSJT random forest.

Similarly, Table 11 shows that conscientiousness and MTI items were the most predictive in the item-level MTI model, highlighting the incremental validity of MT in prediction. Again, cognitive ability items were not predictive and typically ranked low in variable importance relative to MTI and conscientiousness items. Con 7 again emerged as the top predictor, followed by several MTI items. Six of the eight MTI items ranked in the top ten, along with three conscientiousness items and one cognitive ability item. Item-level correlations suggested that job performance was uncorrelated with MTI 2 (r = .14, p = .12) and MTI 5 (r = -.13, p = .15), the most predictive MTI items. Similar to the MTSJT results discussed above, these findings (in conjunction with the fact that the item-level MTI random forest outperformed the item-level MTI

elastic net) suggest that these items were capturing variance in job performance through systematic nonlinear relationships and interactions.

Of note, student status did not serve as an influential variable with an average variable importance of 11.63 in the MTSJT model and 10.98 in the MTI model, suggesting that no meaningful differences in job performance ratings existed between student and non-student employees. In sum, across both models, the variable importance metrics suggest that TP, MTI, and conscientiousness items had more influence on job performance in a nonlinear fashion relative to cognitive ability and student status.

Interpreting Models

Random forests are known for their predictive accuracy in part because they capture nonlinear relationships between predictors and the outcome, and account for any meaningful interactions among predictor variables (Putka et al., 2018). However, these qualities also make random forests more difficult to interpret and explain. Despite this, recent statistical advances have provided procedures for interpreting the main effects and interaction effects from these models (Molnar, 2020).

One method of interpreting the main effects of random forests is using a partial dependence plot (PDP). PDPs show the marginal effect of a given predictor on the outcome. Examining PDPs allows researchers to examine, at a high level, the nature of the relationship between the predictor and the outcome. In brief, PDPs work by assigning each observation every possible value of the predictor variable, and averaging the model predictions for each given value, then plotting the results. While interpreting the PDPs for each variable in the item-level random forest MTSJT and MTI models is computationally expensive and beyond the scope of

the current research, three examples using the three variables with the highest variable importance from the MTSJT model are provided in Figures 5, 6, and 7.

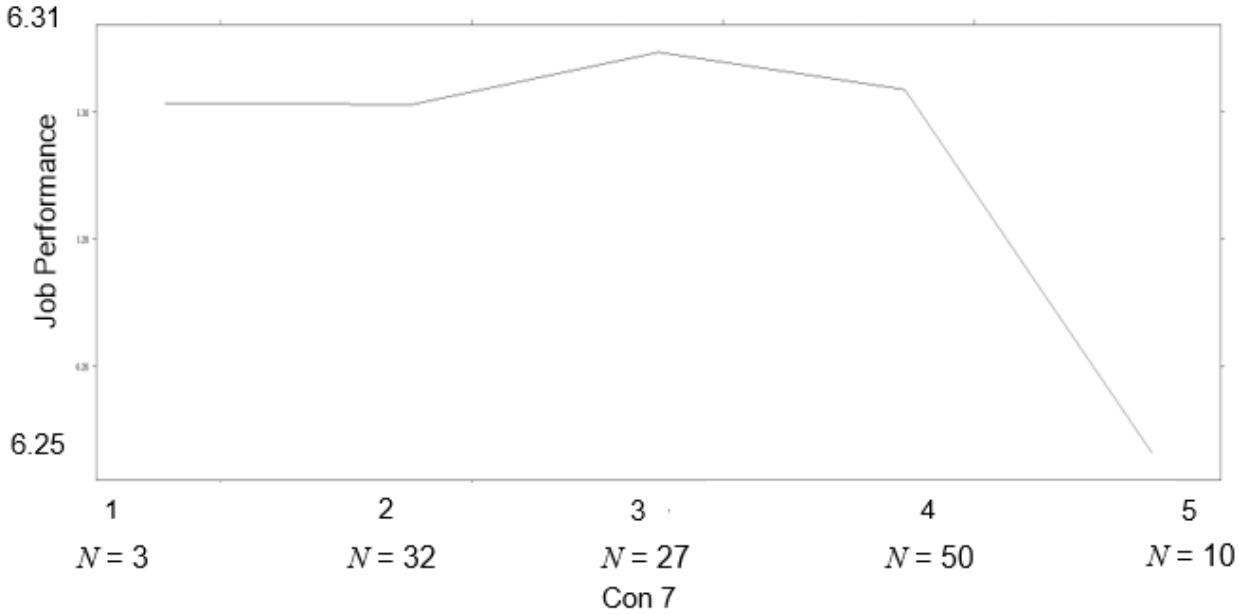

Figure 5. Partial dependence plot demonstrating the main effect of item Conscientiousness 7 on job performance scores. Note that the *N* for each response option is included only for informational purposes – partial dependence plots are *not* calculated solely by averaging the predicted outcome for each observed response at each scale point.

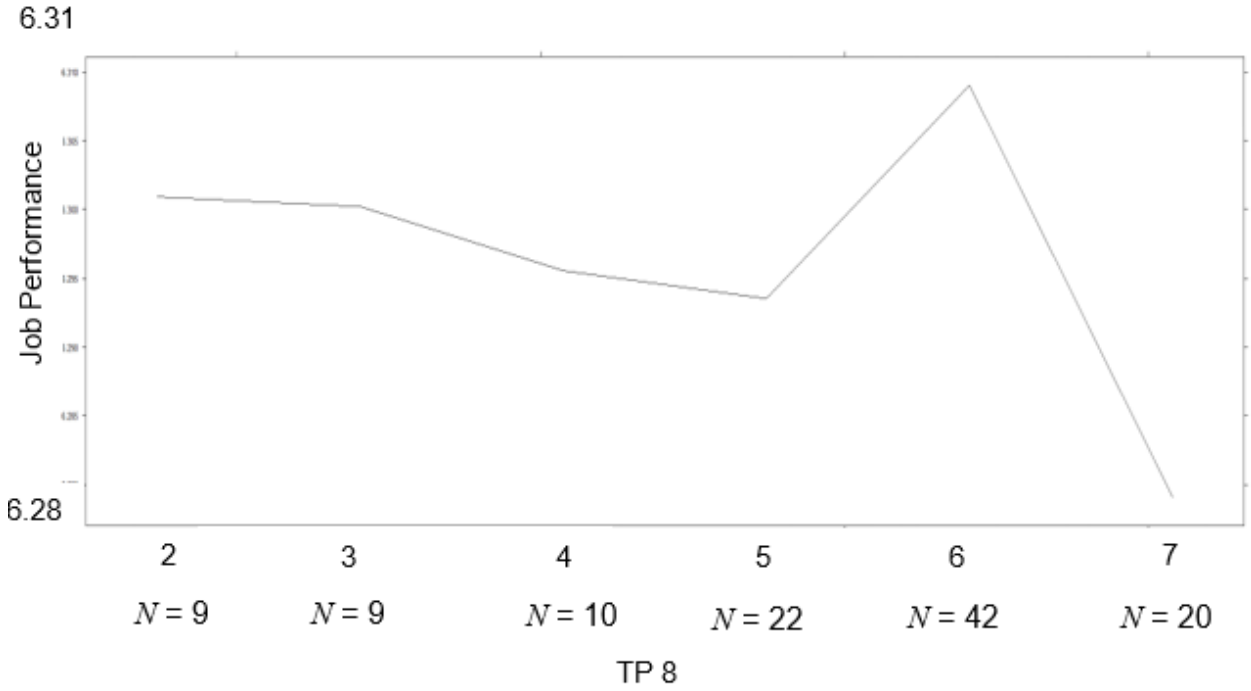

Figure 6. Partial dependence plot demonstrating the main effect of item Task Persistence 8 on job performance scores. The x-axis begins at 2 because no individuals endorsed 1 as a response option. Note that the *N* for each response option is included only for informational purposes – partial dependence plots are *not* calculated solely by averaging the predicted outcome for each observed response at each scale point.

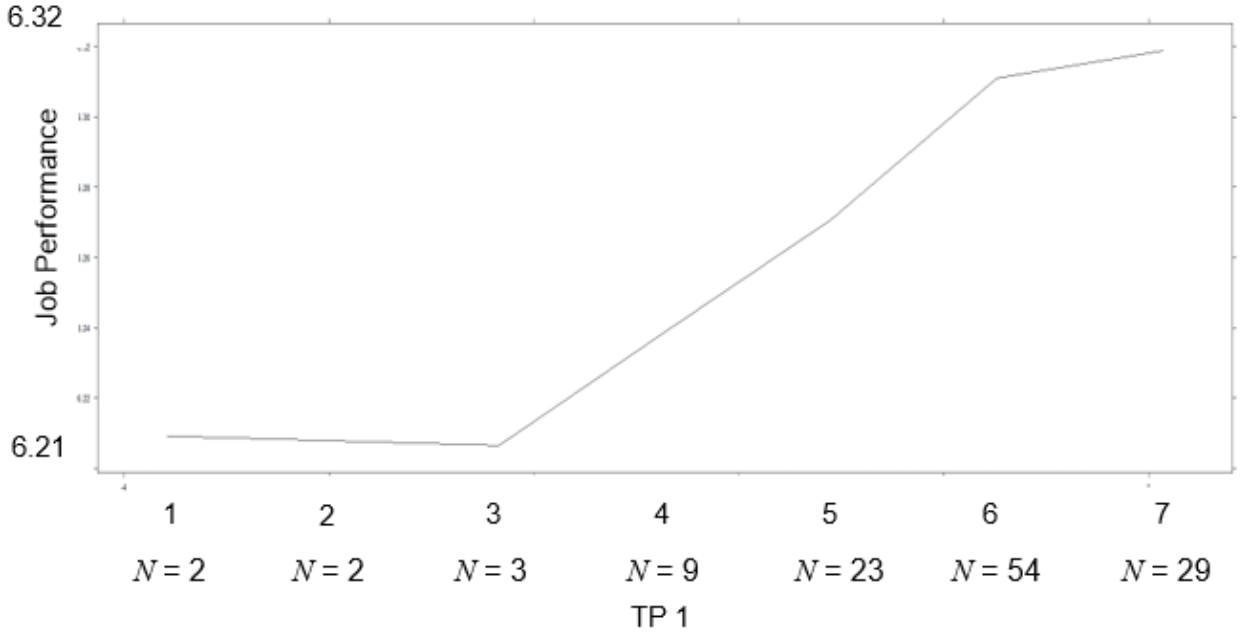

Figure 7. Partial dependence plot demonstrating the main effect of item Task Persistence 1 on job performance scores. Note that the *N* for each response option is included only for informational purposes – partial dependence plots are *not* calculated solely by averaging the predicted outcome for each observed response at each scale point.

When examining these figures, it is important to keep in mind that the y-axis for these figures does not extend below 6.21. If these plots were recreated with a lower bound of 1 (or even 3.14, the lowest observed job performance rating) the nonlinear relationship would not be visually observable; the variables would look unrelated (i.e., a flat horizontal line). However, given that most job performance ratings ranged between six and seven, the random forest model was able to systematically identify nonlinear trends within this small range. Thus, these figures represent small, yet systematic, trends in the data.

As can be seen from Figure 5, the average job performance level remains high and constant across the first several values of Con 7 but declines sharply when an individual endorses the last response option. Although most research has suggested conscientiousness to be positively related to job performance (e.g., Barrick et al., 2001), a body of research has found the relationship between job performance to weaken and eventually disappear beyond a certain point

(LaHuis et al., 2005; Le et al., 2010). This trend appears evident in the relationship between Con 7 and job performance ratings.

Showing a different pattern, the PDP for TP 8 (Figure 6) shows subtle declines in job performance across response options one through five. Job performance levels spiked for response option six, but then sharply declined for response option seven. This finding suggests there may be a "sweet spot" for this item, such that an above-average level of task persistence (within this work scenario) allows individuals to overcome obstacles. However, too high a level may be detrimental to performance, as individuals may persist beyond a point that is beneficial to their job performance and/or neglect other duties.

Finally, an examination of Figure 7 indicates TP 1 showed no increase in job performance from response options one through three, then steadily increased linearly between three and six before leveling off between response options six and seven. This figure suggests that a relatively high score on task persistence is again desirable when relating scores to job performance ratings, but beyond a certain point, increases in task persistence do not associate with further increases in job performance. While these are just a few examples from the top-performing model, the main finding is that the items contained in the model are related to the outcome in a nonlinear fashion, not only supporting the use of a random forest for generating predictions at the item-level in these data but also suggesting there may be a "sweet spot" level of task persistence.

To explore this further, PDPs were examined for all ten task persistence items in the model. Seven of the ten items demonstrated one of the two patterns mentioned above, such that increases in TP scores were associated with increases in job performance to a point beyond which job performance ratings either stabilized (e.g., TP 1) or decreased (e.g., TP 8). Two items

(TP 3 and TP 4) showed generally linear trends, while one item (TP 2) showed a unique jigsaw pattern. Thus, while the PDPs for several of the most predictive TP items suggest a "sweet spot," these results emphasize the importance of contextual assessment – the nature of the relationship between task persistence and job performance may look different depending on the contextual factors present.

Overall interaction strength. An additional strength of random forests is accounting for interactions among predictors. Such interactions further allow random forests to capture systematic relationships in the data. Although the interactions can grow to a great depth and become very complex, an overall score — Friedman's *H* — can be computed for each predictor that measures the strength of all possible interactions between that predictor and every other predictor in the model. Tables 12 and 13 present Friedman's *H* statistic for each predictor in both the MTSJT and MTI models respectively. Friedman's *H* ranges from zero to one and is interpreted as the proportion of variance accounted for by the interaction[3].

Table 12

Overall Interaction Strength among Predictors in Item-level Random Forest MTSJT Model

Item	Friedman's *H*
TP 6	0.08
TP 1	0.07
EC 2	0.06
Con 4	0.06
G 7	0.06
Con 8	0.06
TP 4	0.06
TP 3	0.06
TP 7	0.05
TP 8	0.05
EC 3	0.05

[3] This percentage of variance is out of the total amount of variance captured by the model, *not* the total amount of variance in job performance scores.

G 9	0.05
UF 2	0.05
TP 2	0.05
Con 10	0.05
TP 10	0.05
UF 8	0.05
Con 6	0.04
G 1	0.04
Con 1	0.04
Con 7	0.04
G 10	0.04
EC 1	0.04
G 4	0.04
UF 6	0.04
Con 5	0.04
G11	0.04
TP 5	0.04
G 5	0.04
Con 2	0.04
G 6	0.04
UF 1	0.04
G 8	0.03
TP 9	0.03
UF 7	0.03
Con 9	0.03
UF 4	0.03
UF 5	0.03
G 2	0.03
G 3	0.03
Con 3	0.02
UF 3	0.02
Student Status	0.01

Note: *N* =122; Con = conscientiousness, TP = Task Persistence, EC = Emotional Control, UF = Utilization of Feedback, G = cognitive ability.

Table 13

Overall Interaction Strength among Predictors in Item-level Random Forest MTI Model

Item	Friedman's *H*
G 7	0.11
MTI 6	0.11
Con 5	0.11
MTI 5	0.09

Con 4	0.08
MTI 1	0.08
G 6	0.08
Con 1	0.07
MTI 3	0.07
Con 10	0.07
MTI 7	0.07
Con 8	0.07
MTI 2	0.07
MTI 4	0.06
Con 3	0.06
Con 6	0.06
G 8	0.06
Con 7	0.06
G 9	0.06
G 4	0.05
Con 9	0.05
G 10	0.05
MTI 8	0.05
G 3	0.05
G 5	0.04
G 11	0.04
Con 2	0.04
G 2	0.04
G 1	0.03
Student Status	0.01

Note: N =122; Con = conscientiousness, MTI = Mental Toughness Index, G = cognitive ability.

As can be seen, MTSJT items consistently had stronger interactions than conscientiousness and cognitive ability items. Items TP 6 and TP 1 had the highest *H* values at .08 and .07. Thus, the two-way interactions involving TP 6 and all other predictors accounted for 8% of the variance accounted for by the entire model, and the interactions involving TP 1 and all other predictors accounted for 7% of the variance accounted for by the entire model. No other variables demonstrated the same interaction strength, as Con 4, G 7, TP 4, and TP 3 all had *H* values of .06. Overall, seven of the highest ten *H* statistics were MTSJT items (six were TP items), with two conscientiousness items and one cognitive ability item included.

Similarly, MTI items showed consistent levels of interaction as well, although the results suggest that interactions among conscientiousness items and cognitive ability items influenced the predictions more strongly in this model compared to the MTSJT model. Specifically, items G 7, MTI 6, and Con 5 tied with the highest *H* value of .11. Overall, four of the top ten highest *H* values corresponded to MTI items, four corresponded to conscientiousness items, and two corresponded to cognitive ability items. While the absolute values of the top *H* statistics are higher in the item-level MTI random forest than the item-level MTSJT, it is important to remember that there are fewer items in the MTI model and therefore each interaction can claim a larger proportion of the variance accounted for by the overall predictive model.

Ultimately, these results suggested random forest captured variance due to meaningful interactions among MT items — an advantage not inherently available in linear models such as the elastic net. Notably, interaction statistics across both models further supported the finding that student status did not influence job performance, suggesting the prediction of job performance scores was driven primarily by MT items across both models.

Specific interactions: TP 6. In addition to examining the overall interaction strength for a given predictor, interactions within the random forest can be explored for each specific predictor variable as well. While doing so for every variable in the item-level MTSJT and MTI random forests is beyond the scope of the current research (given the number of predictors in the models), Table 14 provides the specific interaction strengths for TP 6, which had the strongest overall interaction strength in the item-level random forest MTSJT. As can be seen, this item interacts strongest with other MTSJT items (i.e., TP 8, TP 9) to drive prediction.

Table 14

Interaction Strength between TP 6 and All Items in Item-level Random Forest MTSJT Model

Item	Friedman's *H*
TP 8:TP 6	0.10
TP 9:TP 6	0.08
G 9:TP 6	0.06
UF 7:TP 6	0.05
UF 2:TP 6	0.05
EC 2:TP 6	0.05
G 3:TP 6	0.05
Con 7:TP 6	0.05
TP 2:TP 6	0.04
G 1:TP 6	0.04
TP 10:TP 6	0.04
G 11:TP 6	0.04
Con 8:TP 6	0.04
TP 1:TP 6	0.03
G 4:TP 6	0.03
UF 8:TP 6	0.03
Con 1:TP 6	0.03
G 2:TP 6	0.03
UF 6:TP 6	0.03
Con 9:TP 6	0.03
Con 6:TP 6	0.03
G 10:TP 6	0.03
EC 3:TP 6	0.02
UF 4:TP 6	0.02
G 6:TP 6	0.02
UF 1:TP 6	0.02
TP 4:TP 6	0.02
TP 3:TP 6	0.02
EC 1:TP 6	0.02
Con 4:TP 6	0.02
TP 7:TP 6	0.02
UF 3:TP 6	0.02
Student Status:TP 6	0.01
Con 3:TP 6	0.01
Con 10:TP 6	0.01
G 8:TP 6	0.01
G 7:TP 6	0.01
UF 5:TP 6	0.01
TP 5:TP 6	0.01
G 5:TP 6	0.01
Con 5:TP 6	0.01

Con 2:TP 6	0.01

Note: *N* =122; Con = conscientiousness, TP = Task Persistence, EC = Emotional Control, UF = Utilization of Feedback, G = cognitive ability. Variable left of the colon indicates the specific variable interacting with TP 6 in the two-way interaction.

Figures 8 and 9 are heatmaps (with TP 6 along the x-axis and either TP 8/TP 9 along the y-axis) that show how these interactions affect the prediction of job performance — bright yellow areas indicate high levels of predicted job performance, while dark blue areas represent low levels of predicted job performance. The relationship is nonlinear and the highest levels of job performance are predicted for individuals scoring in the mid-to-high range on each predictor, and the lowest predicted job performance levels are for individuals scoring low on TP 6 and high on TP 8 or TP 9. The results suggest that low scores on TP 6 are detrimental to job performance. However, these results again suggest there is a "sweet spot" for TP item scores, such that an above-average score on most TP items is desirable but too high of a response may be detrimental (i.e., employees may be overcommitting to tasks and neglecting other responsibilities).

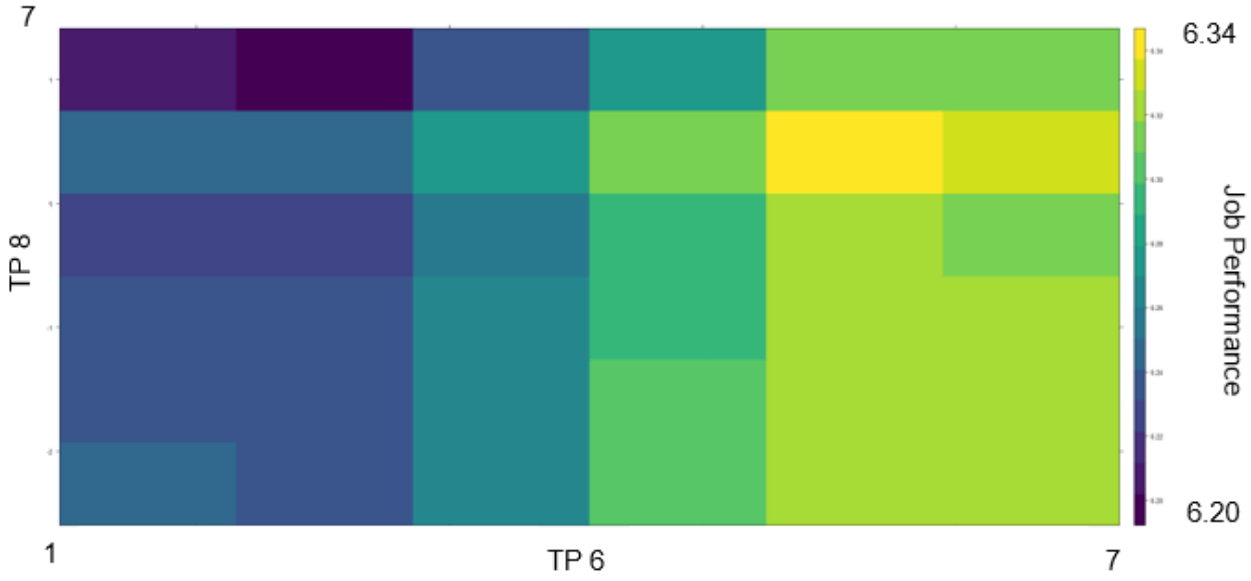

Figure 8. Interactive effect of items Task Persistence 6 and Task Persistence 8 on job performance scores. Bright areas represent high job performance scores and dark areas represent low job performance scores.

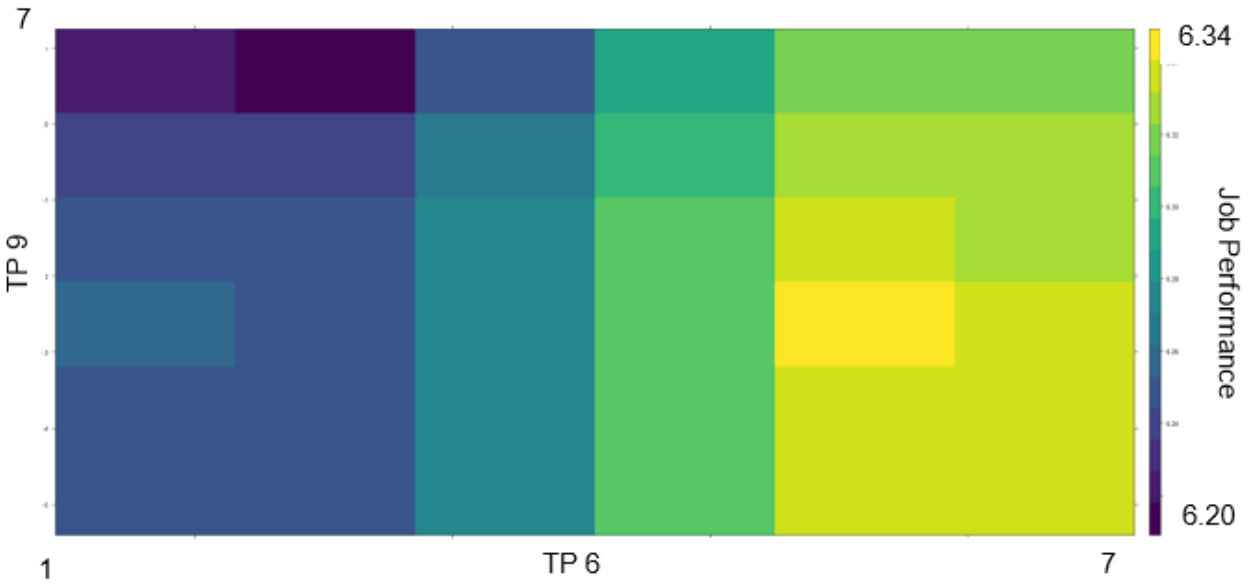

Figure 9. Interactive effect of items Task Persistence 6 and Task Persistence 9 on job performance scores. Bright areas represent high job performance scores and dark areas represent low job performance scores.

In summary, the variable importance metrics, PDP plots, and interaction metrics suggest: 1) MTSJT and MTI items provide value in predicting job performance and are stronger determinants compared to conscientiousness and cognitive ability items, and 2) the relationship between MTSJT/MTI items and job performance was nonlinear and dependent on interactions among predictors, explaining why random forests produced superior predictions compared to elastic nets.

Results Summary

In sum, support was found for Hypothesis 1, such that most measures of MT, grit, and resilience were moderately correlated at the scale-level. Partial support was found for Hypothesis 2, in that the item-level random forest MTI and MTSJT models provided incremental validity in predicting job performance over the control model. At the scale-level, the SMTQ and MTSJT random forests significantly reduced the *RMSE* values relative to the control model. Regardless

of the scale- or item-level approach, or algorithmic model (i.e., elastic net or random forest), the BRS and grit scales did not provide incremental validity.

These findings also provided partial support for Hypothesis 2a, which predicted that the MTSJT would be the strongest predictor among measures of MT, grit, and resilience. The MTSJT was the only measure that provided incremental validity at both the item and scale level. Further, the results suggested that the item-level MTSJT random forest was tied with the item-level MTI random forest for the top-performing model overall. Finally, the variable importance metrics suggested that in general, MTSJT items had more influence on the model than most conscientiousness and cognitive ability items.

In support of Hypothesis 3, item-level models had significantly lower *RMSE* values, and the two best performing models overall were the item-level random forest MTSJT and MTI models. However, no difference was found between item and scale-level models regarding their R^2 values. Finally, an exploratory aim of this study was to determine whether the elastic net or random forest approach produced superior predictions. The findings suggested that random forests were superior at the item level and elastic nets were superior at the scale level, but that the random forest produced the two top-performing models overall — the item level MTSJST and MTI models.

Chapter 5

Discussion

The current study investigated whether the MTSJT, a contextualized assessment of MT at work, predicts supervisor evaluations of employee job performance (when cross-validating predictive models). Further, the study served as an empirical investigation of the jingle-jangle fallacy surrounding the measurement of MT and like constructs. To that end, the specific aims of the study were to determine whether: 1) the MTSJT could account for incremental variance over cognitive ability and conscientiousness; 2) the MTSJT was a better predictor of job performance than non-contextualized MT measures; 3) and whether measures of MT, grit, and resilience equally predict job performance.

In brief, the results suggested that the MTSJT consistently provided incremental variance in predicting job performance above cognitive ability and conscientiousness when using a random forest model at either the item or scale level. Measures of grit and resilience did not account for incremental variance, regardless of the model used; while two non-contextualized assessments of MT (the MTI and SMTQ) showed an inconsistent pattern of results depending on

the predictor type (item v. scale) used. Bottom line: the MTSJT was the assessment tool best suited to consistently predict supervisor evaluations.

An additional aim of this study was to use machine learning algorithms to investigate: 1) whether item-level or scale-level predictive models performed best on test data; and 2) whether the elastic net or random forest algorithm performed best. Overall, the results suggested that the top-performing overall models were item-level MTI and MTSJT random forests, which had the highest R^2 and lowest *RMSE* values in the test data.

Internal Structure of the MTSJT

Before conducting the predictive analyses, the factor structure of the MTSJT was examined. While the results for the correlated three-factor model suggested a good overall fit and superior fit relative to a unidimensional model, several items had very low loadings. For the reasons mentioned above, all items were retained for the predictive analyses, yet it is necessary to critically examine the need to retain these items in the future. The task persistence factor did not contain any items with extremely low loadings. However, two of the EC items and two of the UF items had loadings near or below .20. Further, the TP items were generally superior to EC and UF items with regard to predictive validity. Although research in a larger and more representative sample is needed across differing criteria, the relatively low variable importance scores in conjunction with the low factor loadings may justify eliminating the EC factor and dropping UF 3 in future administrations of the MTSJT.

Further, factor correlations among the TP, UF, and EC dimensions were high (greater than .70). While these results may suggest a unidimensional factor structure, the unidimensional model-fit was inferior to the three-factor fit. Further, the predictive analyses suggested that the

TP items were most effective in predicting job performance ratings in comparison to UF and EC items. Thus, it appears that the subfacets (specifically TP and UF) are capturing distinct-yet-related aspects of MT. In sum, the results support the conceptualization of the MTSJT as a multidimensional, rather than unidimensional, assessment. These results contribute to the current debate in the literature regarding the dimensionality of MT (Gucciardi et al., 2015) providing support for a multidimensional conceptualization.

The Jingle-Jangle Fallacy

As previously discussed, measures of MT and related constructs have contributed to the jingle-jangle fallacy, wherein moderate-to-high correlations have been observed between MT measures and measures of grit, conscientiousness, and resilience (among others), while measures of MT correlate with other measures of MT to a lesser extent than expected. As such, it is difficult to ascertain what is being captured by measures of MT and whether such measures are simply "old wine in new bottles." To that end, Hypothesis 1 was that all measures of MT, grit, resilience, & conscientiousness would correlate moderately. Generally speaking, support was found for this hypothesis, as correlations among the total scores of these measures ranged from .19 to .66 and were significant at the .05 level. In support of the jingle fallacy, the SMTQ correlated at .66 with the MTI and .19 with the MTSJT, while the MTI and MTSJT correlated at .42, showing that measures intended to capture the same construct were not highly related.

Further, these measures demonstrated correlations with grit, conscientiousness, and resilience that were equal to or higher than their correlations with other measures of MT, demonstrating the jangle fallacy. One important exception to this was the fact that the BRS scores were unrelated to the MTSJT total score, suggesting these measures may be capturing distinct constructs. Generally, he results support the conclusions of previous researchers (e.g.,

Credé et al., 2017) that moderate-to-high levels of overlap exist among measures of MT and related constructs.

Notably, the correlations among these self-report measures support Flannery et al.'s (2019) claim that the MTSJT avoids common-method bias in responding due to the differing response format. As shown in Table 7, MTSJT scores demonstrated the weakest correlations with all other self-report measures in the assessment battery. The fact that the MTSJT is a contextually driven measure that uses an SJT format suggests that it captures distinct variance from typically self-report measures of MT, grit, resilience, and conscientiousness.

Predictive Validity

The central aim of this study was to examine the predictive validity for several measures of MT, grit, and resilience in the context of supervisor evaluations of job performance using machine learning models (i.e., random forests and elastic nets). Further, the research was designed to determine if such measures could account for more variance than traditional predictors of job performance and whether the predictions were more accurate when modeled at the item- or scale-level. Importantly, conclusions were drawn from the cross-validated results, *not* the results obtained from the data used to train the models. Therefore, predictive accuracy was expected to attenuate when comparing the training and test results.

Importance of Cross-validated R^2

While the absolute value of the R^2 values produced by the models in the current study are relatively small, it is important to consider that the values are cross-validated R^2 values, which are expected to be smaller. Therefore, the relatively small boost in predictive validity when using the item-level MTI or MTSJT random forests is not negligible. The models' training

performance (which is analogous to the R^2 values typically reported in psychological research) suggests the models were capturing an average of 19% of the variance in job performance. However, the large degree of attenuation suggests that a substantial portion of the variance captured during model development was noise — certainly an issue not idiosyncratic to the current study (Putka et al., 2018). Furthermore, the fact that the models were trained and tested 100 times suggests that these MT models were able to consistently capture systematic variance in job performance.

Predictive Validity of MT and Related Constructs

As Johnson et al. (2008) indicated, one source of evidence for determining whether multiple measures capture the same or distinct constructs is to compare their predictive validity. To that end, the incremental validity of measures of MT, grit, and resilience above cognitive ability and conscientiousness were compared. The results indicated that measures of MT accounted for incremental variance in job performance while measures of grit and resilience did not. Specifically, the MTSJT and MTI item-level random forests had significantly higher R^2 values compared to the control model, and the SMTQ and MTSJT scale-level random forests had significantly lower *RMSE* values compared to the control model. These results suggest MT may be a more useful construct for predicting job performance compared to grit and resilience. Thus, Hypothesis 2 was supported for MT measures but not for measures of grit and resilience. Interestingly, this research contributes to a growing body of literature that suggests grit scores do not provide incremental prediction above conscientiousness and cognitive ability (Credé et al., 2017).

However, the pattern of results regarding MT measures suggests that not all measures of MT are created equal. That is, the MTSJT was the only measure capable of improving the

predictive accuracy at both the scale and item level. The MTI was only successful in capturing incremental variance at the item level, and the SMTQ was only successful in reducing *RMSE* at the scale level. Thus, the MSJT provided the most consistent results. Further, in terms of absolute values, the overall best models were the item-level MTI and MTSJT random forests. As such, support was found for Hypothesis 2a. Although the MTSJT was not the most predictive model in all cases; it was the only model to improve prediction over the control model at both the item and scale level, and was tied with the item-level MTI random forest for the top-performing overall.

Predictive validity of the task persistence subscale. Importantly, the task persistence subscale scores of the MTSJT were the most predictive. As evident by Tables 10, 12, and 13, task persistence items scored high in variable importance and consistently interacted with each other and other predictors to account for additional variance in job performance. While the interactive, nonlinear nature of these relationships makes it difficult to make straightforward claims about the nature of the relationship (e.g., "Individuals scoring high on task persistence are better performers), the results speak to the importance of this dimension in general and suggest there may be a "sweet spot" regarding task persistence scores. This "sweet spot" suggests that individuals who score above average on task persistence, but not at the maximum of the scale, perform best. That is, a certain degree of task persistence is needed for employees to overcome obstacles and perform well on the job, but apparently, too much task persistence can be detrimental to job performance. Speculation as to why this could be includes a lack of flexibility and a neglect of other job responsibilities, or overwork or fatigue.

In further support of the importance of measuring MT with a contextually-bound assessment at the item-level, task persistence items did not influence job performance equally.

Broadly speaking, this result suggests that certain contexts and behaviors (as described in each TP item) are more intimately related to job performance than other contexts/behaviors. For example, compare the trends exhibited in Figures 6 and 7. Although both measures assess task persistence, each item has a unique relationship to job performance, emphasizing the importance of contextual information included in the item. Thus, while TP items were most important for predicting job performance and the results suggested a "sweet spot" may be ideal, idiosyncrasies were observed in the relationship between TP items and job performance.

Item vs. Scale Models

An additional aim of this study was to compare item- vs. scale-level models as performance predictors. Grounded in prior research (e.g., Putka et al., 2018), Hypothesis 3 proposed that item-level models would provide significantly more accurate predictions in test data than scale-level models. Comparing all models, item-level models had significantly lower *RMSE* values, but the R^2 values were not significantly different. However, as mentioned previously, top-performing models overall were the item-level MTI and MTSJT random forests. These models produced both the highest R^2 values and the lowest *RMSE* values. Thus, it appears that predictive validity was maximized at the item-level.

Random Forests vs. Elastic Nets

A final exploratory aim of the current study was to determine whether predictions were best modeled using a random forest or an elastic net. The overall results suggested that random forests produced significantly higher R^2 values, and elastic nets produced significantly lower RMSE values. Further inspection revealed that random forests performed better than elastic nets at the item-level, and elastic nets performed better than random forests at the scale-level. Again,

the top-performing overall models were random forest models, indicating that random forests provided the best predictions overall.

It is not surprising that random forests were superior at the item level and elastic nets were superior at the scale level. Why? Because random forests perform best when the predictors are related to the outcome in a nonlinear fashion and when there is a multitude of valuable interactions among the predictors. These conditions are evident in the item-level models, as the number of predictors ranged from 27-42. In contrast, elastic nets cannot naturally account for interactions (i.e., they must be specified by the researcher) and perform well when predictors are linearly related to the outcome. To that end, the scale-level models contained three to seven predictors, providing fewer opportunities for interactions for the random forests to use for capturing unique variance. Further, fewer predictors introduce fewer nonlinear relationships to the model. Thus, it is unsurprising that the elastic net performed better at the scale level, but ultimately item-level MTI and MTSJT random forests consistently provided the most accurate predictions.

Predictive Accuracy of Cognitive Ability and Conscientiousness

Although the cognitive ability and conscientiousness scale scores were weakly and negatively related to job performance scales, these predictors were included in the control level model per Hypothesis 2. As discussed, the item-level random forest MTSJT model emerged as the top-overall model and was further explored. Within this context, it appeared that Con 7 was the most influential predictor in the model. The relationship between Con 7 and job performance is depicted in Figure 5 and demonstrates that conscientiousness scores are slightly positively related to job performance ratings through response options one to four, before declining sharply at the fifth response option (consider that this is a small yet systematic effect).

As mentioned, conscientiousness scores have traditionally been positive, linear predictors of job performance (e.g., Barrick et al., 2001), yet subsequent research has found that the relationship between conscientiousness and job performance attenuates at high levels of conscientiousness (LaHuis et al., 2005, Le et al., 2010). Thus, this finding supports the "too much of a good thing" hypothesis regarding conscientiousness and job performance. However, it is important to keep in mind that these are results for just one item (albeit the most predictive item). Although it would be too cumbersome to examine the relationships with job performance for all conscientiousness items, different items may exhibit different patterns. Further, these results may have been influenced to some extent by the four prominent outliers identified previously.

The model-interpretation results generally suggested that cognitive ability items were unrelated to job performance, regardless of linear or nonlinear relationships. This finding contradicts decades of research (e.g., Hunter, 1986, Hunter & Hunter, 1984) suggesting not only that cognitive ability is positively related to job performance, but that it is the strongest predictor. Similar to the conscientiousness results, these findings could be driven by the four outliers, although the scale-level correlation between job performance and cognitive ability was still negative upon the removal of these individuals. As discussed further in the limitations below, this may be a methodological limitation involving the validity of the cognitive ability and or job performance measures.

Interpretation vs. Prediction

Overall, the top-performing models were the item-level MTSJT and MTI random forests. These models had the highest cross-validated R^2 values and the lowest *RMSE* values, suggesting they provided the most accurate predictions in the data. As mentioned, the accuracy of these

predictions was developed by accounting for systematic nonlinear relationships and interactions among predictors. Although several methods for interpreting the random forests were used, interpreting random forests is still difficult. While some speculative conclusions can be drawn, such as the "sweet spot" for task persistence items, it is important to remember that conclusions about the model's performance are based on the entire model. Therefore, recommendations for future researchers and practitioners are complex. For instance, recommending to hire individuals "high, but not too high" on task persistence would be a leap from the evidence provided here. Rather, the results suggest that all of an individual's scores on the entire assessment battery (conscientiousness, MTSJT, and cognitive ability assessments) should be modeled via an item-level random forest to provide the most accurate predictions of job performance ratings.

Practitioners interested in applying such techniques should consider that the ultimate prediction matters most in machine learning models — attempting to work backward and identify key drivers in the model to focus on could weaken the validity and accuracy of the predictions. However, one challenge presented by the use of machine learning techniques is that hiring managers who aren't well-versed in statistics/machine learning may struggle to interpret or apply such models.

One solution to this would be for the machine learning model to be developed by a team of I/O psychologists using the battery of assessments selected by the hiring manager. This development phase would proceed just as any other criterion study. Once the criterion-related validity of the model is established for the position/organization of interest, the hiring manager can be trained to apply the scores of new applicants to the model. This simple process would result in a job performance prediction for each individual, and the hiring manager could then use a top-down approach. Thus, it is ultimately the predicted values for job performance that would

matter most to the hiring manager, not the scores/relationships among the assessments in the selection battery (placing legal issues aside).

While some researchers may favor models that are more interpretable (e.g., linear), such models must be interpreted with caution (Breiman, 2001b). Clearly, researchers prefer models that maximize both predictive accuracy and interpretability. However, the same characteristics that make models explainable (e.g., linear assumptions, few predictors) are the same characteristics that impede those models from making accurate predictions. Thus, prediction vs. explanation is often seen as opposite ends of a continuum.

While machine learning scientists and I/O psychologists are both working to make machine learning models more interpretable (Molnar, 2020; Putka et al., 2018), researchers have cautioned against favoring more interpretable models with lower predictive validity over complex models with higher predictive accuracy (Breiman, 2001b). Ultimately, predictive accuracy should be the foremost criterion for judging models. Given the fact that machine learning models will continue to permeate into I/O psychology in both research and practice (Oswald, Behren, Putka, & Sinar, 2020; Putka et al., 2018), such issues will be prevalent to I/O psychologists soon.

Limitations

Although this study used supervisor evaluations and novel machine learning techniques in addition to controlling for conscientiousness and cognitive ability, several methodological limitations must be considered.

Sample Issues

First, the overall sample size for the predictive analyses (N =122) was relatively small. Therefore, for each iteration, 98 (80%) cases were used to train the model and 24 (20%) were used in the test set. Given the relatively small sample sizes, both of these samples may be influenced by sampling error per each iteration. Further, while research suggests that machine learning models outperform traditional models (e.g., OLS) in small sample-size-to-predictor ratios, the models ultimately perform best when they have more data from which to learn. Thus, although the machine learning techniques are relatively well equipped to handle small N/P ratios, the models may still generate more accurate predictions with a larger sample size. Moreover, this sample predominantly consisted of white female undergraduate students from a large Southeastern university, suggesting a more diverse sample is needed for external validity.

Occupations

As indicated in Table 5, the predictive sample included 110 employed students and 12 full-time employees. Thus, a wide range of occupations was represented, including but by no means limited to managers, analysts, researchers, servers, receptionists, lifeguards, cashiers, teachers, salespersons, and babysitters, among others. Because of this wide variety, systematic influences of certain occupations could not be analyzed. The degree of variability in occupational type was too large to conduct any moderating analyses, and therefore any potential systematic influences could not be captured.

Cognitive Ability and Conscientiousness Scores

As discussed, one surprising finding of the current study was that the scale-level cognitive ability and conscientiousness scores were negatively related to overall job

performance. Several potential explanations for this finding exist. First, it may be that there was a lack of motivation among participants to put forth effort when completing the cognitive ability assessment. Second, while progressive matrices are typically seen as culture-fair assessments of cognitive ability (e.g., Lewis, DeCamp-Fritson, Ramage, McFarland, Archwamety, 2007), more traditionally assessments such as the Wonderlic Personnel Test may have produced different results.

While questioning the validity of the cognitive ability assessment is fair, such issues do not account for why conscientiousness was negatively related to job performance. Given the convergent validity among the conscientiousness scores and measures of MT, grit, and resilience, the validity of the conscientiousness scale seems not to be the issue. Rather, the most likely explanation for the surprising relationships was perhaps influenced by the criterion itself.

Two issues relating to the criterion may have influenced the negative relationships found: 1) the validity of the assessment scores, and 2) range restriction. Although previous research has supported the use of this assessment tool (Gucciardi et al., 2015; Williams & Anderson, 1991), no measures were implemented in the current study to assess the quality of the data obtained from supervisors. Due to the short length of the supervisor evaluation (seven items), no attention checks were placed within the survey and analyses of response times could not provide conclusive evidence regarding the validity of the responses. Given the online nature of the study and the fact that there were no stakes involved for the supervisors, it is possible that many supervisors acquiesced or inflated their job performance ratings.

Further, there were two selection effects evident based on the methodology of the current study. Over half of the individuals who completed the employee assessment battery (excluding those who failed attention checks) did not provide usable emails to contact their supervisors.

Further, of those who did provide supervisor contact information, only half of the supervisors responded. Therefore, supervisor ratings were only obtained from 22% of the individuals who complete the assessment battery. One potential explanation is that individuals who feared negative evaluations from their supervisors did not provide their contact information, decreasing the range of the observed job performance scores.

Indeed, the variance in the job performance assessment was small, leaving little systematic variance to be captured by the predictive models. Job performance scores ranged from 3.14 to 7, despite having anchors that ranged from 1-7. Therefore, no employees were deemed to be "poor" performers — at the very worst, a select few were deemed to be "below average" (i.e., scoring a 3.14) while the majority were in the "excellent" range (i.e., 6-7). Such a drastic range restriction when combined with potential threats to the validity of the cognitive ability and job performance assessments are likely explanations for the negative and unexpected correlation observed.

Future Directions

These limitations notwithstanding, future research can build upon these findings in several key ways, including studying a larger sample, assessing different job performance criteria, investigating additional constructs included in the jangle fallacy network surrounding MT, and using different machine learning models. Foremost, researchers may consider replicating the current study in an actual selection setting. This would presumably increase respondents' motivation to perform well on the cognitive ability assessment while simultaneously providing a wider range of job performance scores. Beyond that, additional future research extensions are discussed below.

Future Samples

As mentioned in the limitations, the sample size in the current study was relatively small and not representative in terms of gender and ethnicity. Thus, future researchers should replicate this study in a larger, more representative sample of individuals to examine the generalizability of the results. Moreover, a systematic evaluation of occupations is necessary. That is, future researchers should consider studying a select few occupations or grouping occupations into occupational types to include occupation/occupation type as a moderator in the analysis.

Occupational Criteria

In the current study, job performance was measured via a seven-item assessment of in-role task behavior, completed by the employee's supervisor. Future research could expand upon the current study in two main ways — more thoroughly assessing in-role task behavior and assessing different workplace outcomes. As discussed in the limitations section, potential threats to the validity of the job performance ratings in the current study exist. Future researchers may consider assessing in-role task behavior using a more comprehensive assessment tool and by including attention checks to assess the quality of responding. Alternatively, researchers may be interested in obtaining genuine performance appraisals from an organization to serve as the criteria.

Further, subsequent research may examine how MT relates to other workplace criteria. Three potential criteria that are most relevant include counterproductive work behavior (CWB), burnout, and organizational citizenship behavior (OCB). CWBs and burnout are suitable outcomes because research suggests that perceived distress is a proximal antecedent of these outcomes (Demerouti, Bakker, Nachreiner, & Schaufeli, 2001; Penney & Spector, 2005). Given

the research suggesting that MT and like constructs buffer against the negative effects of distress (e.g., Gucciardi et al., 2015; Lin et al., 2017), assessments of these constructs may predict burnout and CWBs. Specifically, these outcomes pertain to the "surviving" aspect of Gucciardi's et al.'s (2015) conceptualization of MT as the ability to "strive, survive, and thrive."

Further, OCBs may be a relevant MT criterion because of the "strive" and "thrive" elements. Individuals striving to perform their very best may exert additional effort beyond that required by their job, while those thriving may be more likely to assist coworkers or take on additional responsibilities. In sum, researchers should consider more thoroughly assessing in-role task behavior while expanding the criteria used to assess job performance in the context of MT.

Additional Constructs

While three prominent constructs influencing the jangle fallacy surrounding MT — grit, resilience, and conscientiousness — were included in the study, several important constructs were omitted to make the assessment battery manageable for participants. Additional constructs that should be investigated include hardiness, optimism, and self-motivation. Also, other measures of MT, grit, resilience, and conscientiousness exist that were not included in the current study, including the MTQ48 (Clough et al., 2002), the Mental Toughness Inventory (Middleton et al., 2004), the Connor-Davidson Resilience Scale (Connor & Davidson, 2003), the Resilience Scale for Adults (Friborg, Hjemdal, & Rosenvinge, 2003), as well as variations of the grit scale used in the current research (e.g., Short Grit Scale; Duckworth & Quinn, 2009). In sum, future research should use additional measures within the nomological network of MT.

Other Machine Learning Algorithms

A final domain future researchers may wish to explore is the use of other machine learning algorithms to relate MT predictor scores to job performance outcomes. Given the fact that these models perform well with small N/P ratios, researchers could include many scales mentioned in the previous section to compare the predictive validity of these assessments. Moreover, additional algorithms, including support vector machines, gradient boosted trees, and neural networks seek to address the aforementioned bias-variance tradeoff in different manners using different tuning parameters. As a result, these algorithms may make different predictions and identify different predictor variables as important drivers of job performance. As indicated by prominent I/O psychologists (e.g., Oswald et al., 2020; Putka et al., 2018), such predictive models will continue to grow in popularity with IO psychology and therefore should be increasingly applied in selection research.

Conclusion

This study examined the validity of measures of MT and similar constructs, including grit and resilience, to predict job performance. It was hypothesized that the MTSJT — an assessment designed to measure MT directly in work the workplace — would emerge as the most accurate predictor of job performance among other self-report measures of MT, grit, and resilience while providing incremental variance beyond that accounted for by cognitive ability and conscientiousness. A unique element of this study was the use of machine learning models (i.e., elastic nets and random forests) to examine predictions at both the item and scale level, hypothesizing that such algorithms would result in item-level models producing the top-performing models. By examining the predictive validity of such measures, the study aimed to

determine: 1) whether measures of MT, grit, and resilience differentially predicted job performance, and 2) whether the context-bound MTSJT outperformed other measures.

Despite exhibiting moderate correlations among nearly all measures of MT, grit, and resilience, the MTSJT was the only measure capable of providing incremental validity over cognitive ability and conscientiousness at both the scale and item level. Moreover, the item-level MTSJT random forest was one of two top-performing models overall, suggesting that this model was capable of consistently and systematically capturing unique variance in job performance. The results provided initial support for the use of the MTSJT as a predictor of job performance and advance the current literature (e.g., Putka et al., 2018) that suggests that item-level predictions can outperform scale-level predictions upon cross-validation when using machine learning algorithms. The results emphasize not only the growing importance of MT in the workplace but also the growing importance of machine learning models in I/O psychology.

References

Andersen, M.B. (2011). Who's mental, who's tough, and who's both? Mutton constructs dressed up as lamb. In D. F. Gucciardi & S. Gordon (Eds.). (2011). *Mental Toughness in Sport: Developments in Theory and Research.* Abingdon, UK. Routledge.

Arthur, C.A., Fitzwater, J., Hardy, J. J., Beattie, S., & Bell, J. (2015). Development and validation of a Military Training Mental Toughness Inventory. *Military Psychology*, *27*(4), 232-241.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248-287.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology, 44*(1), 1-26.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9(*1-2), 9-30.

Bédard-Thom, C., & Guay, F. (2018). Mental toughness among high school students: a test of its multidimensionality and nomological validity with academic achievement and preference for difficult tasks. *Social Psychology of Education, 21*(4), 827-848.

Bell, J.J., Hardy, L., & Beattie, S. (2013). Enhancing mental toughness and performance under pressure in elite young cricketers: A two-year longitudinal intervention. *Sport, Exercise, and Performance Psychology, 2*(4), 281-297.

Birch, P.D.J., Crampton, S., Greenlees, I., Lowry, R., & Coffee, P. (2017). The Mental Toughness Questionnaire-48: A re-examination of factorial validity. *International Journal of Sport Psychology, 48*(3), 331-355.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*(3), 265-284.

Breiman, L. (2001a). Random forests. *Machine Learning, 45*(1), 5-32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199-231.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 161-168). ACM.

Chen, M. A., & Cheesman, D. J. (2013). Mental toughness of mixed martial arts athletes at different levels of competition. *Perceptual and Motor Skills*, *116*(3), 905–917.

Cheung, P., & Li, C. (2019). Physical activity and mental toughness as antecedents of academic burnout among school students: A latent profile approach. *International Journal of Environmental Research and Public Health*, *16*(11), 2024-2033.

Clough, P., Earle, K., & Sewell, D. (2002). Mental toughness: The concept and its measurement. *Solutions in Sport Psychology*, 32–43. London: Thomson Publishing.

Cohen, S. Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(1), 386-396.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43*(1), 52-64.

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depression and anxiety*, *18*(2), 76-82.

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270-283.

Credé, M. (2018). What shall we do about grit? A critical review of what we know and what we don't know. *Educational Researcher, 47*(9), 606-611.

Credé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, *36*(6), 845-872.

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492-511.

Crust, L., Earle, K., Perry, J., Earle, F., Clough, A., & Clough, P.J. (2014). Mental toughness in higher education: Relationships with achievement and progression in first-year university sports students. *Personality and Individual Differences*, *69*(1), 87-91.

Crust, L., & Swann, C. (2011). Comparing two measures of mental toughness. *Personality and Individual Differences, 50*(2), 217-221.

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied psychology*, *86*(3), 499-512.

Dewhurst, S.A., Anderson, R.J., Cotter, G., Crust, L., & Clough, P.J. (2012). Identifying the cognitive basis of mental toughness: Evidence from the directed forgetting paradigm. *Personality and Individual Differences, 53*(5), 587-590.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81-106.

Duckworth, A.L. (2013a). The key to success? Grit. Retrieved from https://www.ted.com/talks/angela_lee_duckworth_the_key_to_success_grit?language=en#t-9644

Duckworth, A.L, Kirby, T., Tsukayama, E., Bernstein, H., & Ericsson, K. (2011). Deliberate practice spells success: Why grittier competitors triumph at the national spelling bee. *Social Psychological and Personality Science, 2*(2), 174-181.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166-174.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101.

Dugan, R., Hochstein, B., Rouziou, M., & Britton, B. (2019). Gritting their teeth to close the sale: the positive effect of salesperson grit on job satisfaction and performance. *Journal of Personal Selling & Sales Management, 39*(1), 81-101.

Elliot, A.J., & Thrash, T.M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology, 82*(5), 804-818.

Eskreis-Winkler, L., Duckworth, A. L., Shulman, E. P., & Beal, S. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology*, *5*(1), 1-12.

Eysenck, H. J. (1994). Personality and intelligence: Psychometric and experimental approaches. *Personality and Intelligence*, *1*(1), 3-31.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research*, *15*(1), 3133-3181.

Flannery, N.M., Glasgow, T.E., Torian, B.P., DeVore, B.B., & Acton, B.A. (2017). *Predicting well-being and performance with measures of mental toughness, grit, and hardiness.* Paper presented at the 125th annual American Psychological Association Conference, Washington D.C.

Flannery, N.M., Hauenstein, N.M.A., & Geller, E.S. (2019). *Development of the Mental Toughness Situational Judgment Test: A novel approach to assessing mental toughness.* Poster presented at the 34th annual meeting of the Society of Industrial and Organizational Psychology, Washington, D.C.

Friborg, O., Hjemdal, O., Rosenvinge, J. H., & Martinussen, M. (2003). A new rating scale for adult resilience: What are the central protective resources behind healthy adjustment?. *International journal of methods in psychiatric research*, *12*(2), 65-76.

Godlewski, R., & Kline, T. (2012). A model of voluntary turnover in male Canadian Forces recruits. *Military Psychology*, *24*(3), 251-269.

Golby, J., Sheard, M., & van Wersch, A. (2007). Evaluating the factor structure of the Psychological Performance Inventory. *Perceptual and Motor Skills*, *105*(1), 309-325.

Goldberg, A.S. (1998) *Sports slump busting: Ten steps to mental toughness and peak performance.* Champaign, IL: Human Kinetics.

Goldberg, L.R. (1999). A broad-bandwidth, public domain, personality inventory measuring lower-level facets of several five-factor models. In I. Mervielde, I.J. Deary, F. De Fruyt, and F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, The Netherlends: Tilburg University Press.

Griffin, M. M., & Steinbrecher, T. D. (2013). Large-scale datasets in special education research. In *International Review of Research in Developmental Disabilities* (Vol. 45, pp. 155-183). Academic Press.

Gucciardi, D.F. (2012). Measuring mental toughness in sport: A psychometric evaluation of the Psychological Performance Inventory-A and its predecessor. *Journal of Personality Assessment, 94*(4), 393-403.

Gucciardi, D.F. (2017). Mental toughness: Progress and prospects. *Current Opinions in Psychology, 16(1),* 17-23.

Gucciardi, D.F., & Gordon, S. (2009). Development and preliminary validation of the Cricket Mental Toughness Inventory (CMTI). *Journal of Sports Sciences, 27*(12), 1293-1310.

Gucciardi, D.F., & Gordon, S. (Eds.). (2011). *Mental toughness in sport: Developments in theory and research.* Abingdon, UK. Routledge.

Gucciardi, D.F., Gordon, S., & Dimmock, J.A. (2009a). Development and preliminary validation of a mental toughness inventory for Australian football. *Psychology of Sport and Exercise, 10*(1), 201-209.

Gucciardi, D.F., Gordon, S., & Dimmock, J.A.(2009b). Evaluation of a mental toughness training program for youth-aged Australian footballers: I. A quantitative analysis. *Journal of Applied Sport Psychology, 21*(3), 307-323.

Gucciardi, D.F., Gordon, S., & Dimmock, J.A.(2009c). Evaluation of a mental toughness training program for youth-aged Australian footballers: II. A qualitative analysis. *Journal of Applied Sport Psychology, 21*(3), 324-339.

Gucciardi, D.F., Hanton, S., Gordon, S., Mallett, C.J., & Temby, P. (2015). The concept of mental toughness: Tests of dimensionality, nomological network, and traitness. *Journal of Personality, 83*(1), 26-44.

Gucciardi, D.F., Hanton, S., & Mallett, C.J. (2012). Progressing measurement in mental toughness: A case example of the Mental Toughness Questionnaire-48. *Sport, Exercise, and Performance Psychology, 1*(3), 194-214.

Gucciardi, D.F., Mallett, C.J., Hanrahan, S.J., & Gordon, S. (2011). Measuring mental toughness in sport: current status and future directions. In D.F. Gucciardi and S. Gordon (Eds.). (2011). *Mental toughness in sport: Developments in theory and research.* Abingdon, UK. Routledge.

Gucciardi, D.F., Peeling, P., Ducker, K.J., & Dawson, B. (2016). When the going gets tough: Mental toughness and its relationship with behavioral perseverance. *Journal of Science and Medicine in Sport, 19*(1), 81-86.

Hardy, L., Bell, J., & Beattie, S. (2014). A neuropsychological model of mentally tough behavior. *Journal of Personality, 82*(1), 69-81.

Hardy, J.H., Imose, R. A., & Day, E.A. (2014). Relating trait and domain mental toughness to complex task learning. *Personality and Individual Differences, 68*(1), 59-64.

Hawk, J. A. (1970). Linearity of criterion—GATB aptitude relationships. *Measurement and Evaluation in Guidance, 2*(4), 249-251.

Horsburgh, V. A., Schermer, J. A., Veselka, L., & Vernon, P. A. (2009). A behavioural genetic study of mental toughness and personality. *Personality and Individual Differences, 46*(2), 100–105.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. *Performance Measurement and Theory, 257*(1), 266.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*(3), 340-362.

Hunter, J. E., & Hunter, R. E (1984). Validity and utility of alternate predictions of job performance. *Psychological Bulletin, 96*(1), 72-98.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*(6), 869-879.

International Cognitive Ability Resource Team, 2014.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer

Johnson, R.E., Rosen, C.C., & Levy, P.E. (2008). Getting to the core of core self-evaluations: A review and recommendations. *Journal of Organizational Behavior, 29*(3), 391-413.

Jones, G., Hanton, S., & Connaughton, D. (2002). What is this thing called Mental Toughness? An investigation of elite sport performers. *Journal of Applied Sport Psychology*, *14*(3), 205–218.

Joseph, A. I. (2009). The role of grit in predicting performance in collegiate athletes. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 70,* 7255.

Kaiseler, M., Polman, R.C.J., & Nicholls, A. (2009). Mental toughness, stress, stress appraisal, coping and coping effectiveness in sport. *Personality and Individual Differences, 47*(7), 728-733.

Kelley, T. L. (1927). *Interpretation of educational measurement*. Yonkers, NY: World Book.

Kelly, D. R., Matthews, M. D., & Bartone, P. T. (2014). Grit and hardiness as predictors of performance among West Point cadets. *Military Psychology, 26*(4), 327-342.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1-26.

LaHuis, D. M., Martin, N. R., & Avis, J. M. (2005). Investigating nonlinear conscientiousness-job performance relations for clerical employees. *Human Performance*, *18*(3), 199-212.

Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E. & Westrick, P. (2010). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology*, *11*(1), 1–21.

Lewis, J. D., DeCamp-Fritson, S. S., Ramage, J. C., McFarland, M. A., & Archwamety, T. (2007). Selecting for ethnically diverse children who may be gifted using Raven's Standard Progressive Matrices and Naglieri Nonverbal Abilities Test. *Multicultural Education, 15*(1), 38-42.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*(1), 426-441.

Lin, Y., Clough, P.J., Welch, J., & Papageorgiou, K. (2017). Individual differences in mental toughness associate with academic performance and income. Personality and Individual Differences, *113*(1), 178-183.

Lin, Y., Mutz, J., Clough, P.J., & Papageorgiou, K. (2017). Mental toughness and individual differences in learning, educational and work performance, psychological well-being, and personality: A systematic review. *Frontiers in Psychology, 8*(1), 1-13.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*(3), 269-276.

Lockwood, P., Jordan, C.H., & Kunda, Z. (2002). Motivation by positive or negative role models: Regulatory focus determines who will best inspire us. *Journal of Personality and Social Psychology, 83*(4), 854-864.

Loehr, J. E. (1986). *Mental toughness training for sports: Achieving athletic excellence*. New York City, New York: Plume.

Mack, M.G., & Ragan, B.G. (2008). Development of the Mental, Emotional, and Bodily Toughness Inventory in college athletes and non-athletes. *Journal of Athletic Training, 43*(2), 125-132.

Marchant, D.C., Polman, R.C.J., Clough, P.J., Jackson, J.G., Levy, A.R., & Nicholls, A.R. (2009). Mental toughness: Managerial and age differences. *Journal of Managerial Psychology, 24*(1), 428-437.

Masten, A.S. (2014). Global perspectives on resilience in children and youth. *Child Development, 85*(1), 6-20.

Middleton, S.C., Harsh, H.W., Martin, A.J., Richards, G.E., Savis, J., Perry, C., & Brown, R. (2004). The Psychological Performance Inventory: Is this mental toughness test enough? *International Journal of Sport Psychology, 35*(1), 91-108.

Middleton, S. C., Martin, A. J., & Marsh, H. W. (2011). Development and validation of the mental toughness inventory (MTI). In D. F. Gucciardi and S. Gordon (Eds.). (2011). *Mental Toughness in Sport: Developments in Theory and Research.* Abingdon, UK. Routledge.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246-268.

Molnar, C. (2020). "Interpretable machine learning. A Guide for Making Black Box Models Explainable", https://christophm.github.io/interpretable-ml-book/.

Nicholls, A.R., Polman, R.C.J., Levy, A.R., & Backhouse, S.H. (2008). Mental toughness, optimism, pessimism, and coping among athletes. *Personality and Individual Differences, 44*(5), 1182-1192.

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Annual Review of Organizational Psychology and Organizational Behavior. *Annual Review of Organizational Psychology and Organizational Behavior, 7*(1), 505-533.

Penney, L. M., & Spector, P. E. (2005). Job stress, incivility, and counterproductive work behavior (CWB): The moderating role of negative affectivity. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *26*(7), 777-796.

Perry, J.L., Clough, P.J., Crust, L., Earle, K., & Nicholls, A.R. (2012). Factorial validity of the Mental Toughness Questionnaire-48. *Personality and Individual Differences, 54*(5), 587-592.

Petrie, T. A., Deiters, J., & Harmison, R. J. (2014). Mental toughness, social support, and athletic identity: Moderators of the life stress–injury relationship in collegiate football players. *Sport, Exercise, and Performance Psychology*, *3*(1), 13–27.

Ployhart, R.E., Ehrhart, M.G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, *11*(1), 1-16.

Procter, R. W., & Dutta, A. (1995). *Skill development and human performance*. Oakland, CA: Sage.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689-732.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rojon, C., McDowall, A., & Saunders, M. N. (2015). The relationships between traditional selection assessments and workplace performance criteria specificity: A comparative meta-analysis. *Human Performance*, *28*(1), 1-25.

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology*, *88*(4), 797-834.

Scelfo, J. (2016, April 8). Angela Duckworth on passion, grit, and success. *The New York Times.* Retrieved from https://www.nytimes.com/2016/04/10/education/edlife/passion-grit-success.html

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*(3), 445-494.

Sheard, M., & Golby, J. (2006). Effect of a psychological skills training program on swimming performance and positive psychological development. *International Journal of Sport and Exercise Psychology, 4*(2), 149–169.

Sheard, M., Golby, J., & van Wersch, A. (2009). Progress toward construct validation of the Sports Mental Toughness Questionnaire (SMTQ). *European Journal of Psychological Assessment, 25*(3), 186-193.

Shoenfelt, E.L. (2016). How much do we really know about employee resilience? More, if we include the sport psychology research. *Industrial and Organizational Psychology, 9*(2), 442-446.

Smith, B. W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The Brief Resilience Scale: Assessing the ability to bounce back. *International Journal of Behavioral Medicine, 15*(3), 194-200.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*(4), 718-737.

St. Clair-Thompson, H., Bugler, M., Robinson, J., Clough, P., McGeown, S. P., & Perry, J. (2015). Mental toughness in education: Exploring relationships with attainment, attendance, behaviour and peer relationships. *Educational Psychology, 35*(7), 886-907.

St. Clair-Thompson, H., Giles, R., McGeown, S.P., Putwain, D., Clough, P. J., & Perry, J. (2017). Mental toughness and transitions to high school and to undergraduate study. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *37*(7), 792-802.

Suzuki, Y., Tamesue, D., Asahi, K., & Ishikawa, Y. (2015). Grit and work engagement: A cross-sectional study. *PloS One*, *10*(9), e0137501.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Teachers College, Columbia University.

Van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *The Australian Journal of Advanced Nursing*, *25*(4), 40-48.

Van Iddekinge, C. H., Aguinis, H., Mackey, J. D., & DeOrtentiis, P. S. (2018). A meta-analysis of the interactive, additive, and relative effects of cognitive ability and motivation on performance. *Journal of Management, 44*(1), 249-279.

Vaughan, R., Hanna, D., & Breslin, G. (2018). Psychometric properties of the Mental Toughness Questionnaire 48 (MTQ48) in elite, amateur and nonathletes. *Sport, Exercise, and Performance Psychology*, *7*(2), 128-140.

Veselka, L., Schermer, J.A., Petrides, K., & Vernon, P.A. (2009). Evidence for a heritable general factor of personality in two studies. *Twin Research and Human Genetics, 12*(3), 254-260.

Ward, F., St Clair-Thompson, H., & Postlethwaite, A. (2018). Mental toughness and perceived stress in police and fire officers. *Policing: An International Journal, 41*(1), 674-686.

Whetzel, D.L., & McDaniel, M.A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*(3), 188-202.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17*(3), 601-617.

Youssef, C. M., & Luthans, F. (2007). Positive organizational behavior in the workplace: The impact of hope, optimism, and resilience. *Journal of Management, 33*(5), 774-800.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(5), 301-320.

Appendix A

MTSJT – Final Form

Note: *indicates the response option is not scored

1. A coworker is absent. Your supervisor asked you to perform some of your coworker's tasks for the day, but your day is planned with very little free time. What is the likelihood you would engage in each of the following behaviors?
    a. Stay at work longer than usual until you've completed your tasks and your coworker's tasks **(TP 1)**
    b. Focus your time on the most important tasks so you can leave work without working late*
    c. Spend time during lunch relieving stress so you don't feel overwhelmed by the added tasks **(EC 1)**
2. Your supervisor is disappointed with a project you completed with a coworker. You just didn't mesh with the coworker, but the blame has been placed entirely on you. What is the likelihood you would engage in each of the following behaviors?
    a. Ask not to work with this coworker in the future
    b. Complete extra work to make-up for the disappointing project **(TP 2)**
    c. Calmly inform your supervisor that the responsibility should be split between you and your coworker **(EC 2)**
3. You made a critical mistake on the job, and are required to discuss it tomorrow with your supervisor. What is the likelihood you would engage in each of the following behaviors?
    a. Spend some time before the meeting reviewing the details of the mistake to identify what may have caused it **(TP 3)**

b. Tell your supervisor that you've made it your personal goal not to make this mistake again **(EC 3)**

c. Ask your supervisor for suggestions about how to avoid making the mistake in the future **(UF 1)**

4. A coworker took credit for a task that you completed very well. You decide to address the issue with your coworker. What is the likelihood you would think and feel in each of the following manners?

    a. You want to address the problem head on to earn your recognition*

    b. You want to inform the coworker their behavior has caused you to avoid working with them in the future **(UF 2)**

    c. You want to inform the coworker that you are not upset*

5. Your company introduced new technology that you are now required to use. The technology is hard to use and has a steep learning curve. What is the likelihood you would engage in each of the following behaviors?

    a. Spend some extra time during your lunch break learning how to use the device **(TP 4)**

    b. Ask a proficient coworker to teach you how to use the device **(UF 3)**

    c. Remove distractions from your workplace to focus on mastering the device*

6. You have been working on an extremely difficult project that is due next week, but you are nowhere near finished. What is the likelihood you would engage in each of the following behaviors?

    a. Write down a list of the most crucial parts of the project to focus on those*

b. Listen to an experienced coworker's advice about how best to complete the project **(UF 4)**
c. Set aside a block of time on Saturday to complete the project **(TP 5)**

7. You've been recently hired and have been given plenty of new responsibilities. Your probationary period ends in just two weeks, at which point the company will decide whether to keep you or not. What is the likelihood you would engage in each of the following behaviors?
    a. Work during the weekends leading up to the end of the probationary **(TP 6)**
    b. Rewrite your daily schedule so that you can focus on the most relevant responsibilities first*
    c. Listen to suggestions from coworkers on how to improve your performance **(UF 5)**
8. You've been working on a project that you thought would be quick, but has become increasingly difficult as it has developed. Your supervisor has become impatient waiting for you to finish. What is the likelihood you would engage in each of the following behaviors?
    a. Have a working lunch every day this week to complete extra tasks **(TP 7)**
    b. Write down the most essential parts of the project to focus on those first*
    c. Listen to the advice of one of your successful coworkers regarding how best to complete the project **(UF 6)**
9. You feel you've been performing very poorly at your job as of late, and want to make some changes. A coworker recommends you talk to your supervisor about this issue. What is the likelihood you would engage in each of the following behaviors?

a. Remove distractions from your workplace environment so you can focus*
b. Incorporate any suggestions for improving performance that your supervisor might have **(UF 7)**
c. Attend extra training events to improve your abilities **(TP 8)**

10. You have heard from coworkers that your company is intending on laying off a vast number of workers from a variety of positions. You ignore the rumors. What is the likelihood you would think and feel in each of the following manners?
    a. You view the rumors as distractions from your job*
    b. You want to maintain the morale of the coworkers*
    c. You want to demonstrate your value to the company **(TP 9)**
11. You have received an exceptional performance review from your supervisor, indicating you should keep up the excellent work by creating a plan to maintain your level of performance. What is the likelihood you would think and feel in each of the following manners?
    a. You want to solidify yourself as one of the top employees*
    b. You take pride in the positive comments made by your supervisor **(UF 8)**
    c. You acknowledge the fact that you need to continue to put in long hours to perform well **(TP 10)**

Appendix B

Mental Toughness Index (MTI; Gucciardi et al., 2015)

1. I believe in my ability to achieve my goals
2. I am able to regulate my focus when performing tasks
3. I am able to use my emotions to perform the way I want to
4. I strive for continued success
5. I effectively execute my knowledge of what is required to achieve my goals.
6. I consistently overcome adversity
7. I am able to execute appropriate skills or knowledge when challenged.
8. I can find a positive in most situations.

Appendix C

Sports Mental Toughness Questionnaire (SMTQ; Sheard et al., 2009)

Item	Subscale
I interpret potential threats as positive opportunities	Confidence
I have an unshakeable confidence in my ability	Confidence
I have qualities that set me apart from other competitors	Confidence
I have what it takes to perform well under pressure	Confidence
Under pressure, I am able to make decisions with confidence and commitment	Confidence
I can regain my composure if I have momentarily lost it	Confidence
I am committed to completing the tasks I have to do	Constancy
I take responsibility for setting myself challenging targets	Constancy
I give up in difficult situations (R)	Constancy
I get distracted easily and lose my concentration (R)	Constancy
I worry about performing poorly (R)	Control
I am overcome by self-doubt (R)	Control
I get anxious by events I did not expect or cannot control (R)	Control
I get angry and frustrated when things do not go my way (R)	Control

Note: (R) = reverse coded.

Appendix D

12-item Grit Scale (Duckworth et al., 2007)

Item	Subscale
New ideas and projects sometimes distract me from previous ones (R)	Passion
I have been obsessed with a certain idea or project for a short time but later lost interest (R)	Passion
I often set a goal but later choose to pursue a different one (R)	Passion
I have difficulty maintaining my focus on projects that take more than a few months to complete (R)	Passion
My interests change from year to year (R)	Passion
I become interested in new pursuits every few months (R)	Passion
I have achieved a goal that took years of work.	Perseverance
I have overcome setbacks to conquer an important challenge.	Perseverance
Setbacks don't discourage me	Perseverance
I am a hard worker	Perseverance
I finish whatever I begin	Perseverance
I am diligent	Perseverance

Note: (R) = reverse coded.

Appendix E

Brief Resilience Scale (Smith et al., 2008)

1. I tend to bounce back quickly after hard times
2. I have a hard time making it through stressful events (R)
3. It does not take me long to recover from a stressful event
4. It is hard for me to snap back when something bad happens (R)
5. I usually come through difficult times with little or no trouble
6. I tend to take a long time to get over set-backs in my life (R)

Note: (R) = reverse coded.

Appendix F

10-item IPIP Conscientiousness Measure (Goldberg, 1999)

1. I am always prepared.
2. I pay attention to details.
3. I get chores done right away.
4. I carry out my plans.
5. I make plans and stick to them.
6. I waste my time. (R)
7. I find it difficult to get down to work. (R)
8. I do just enough work to get by. (R)
9. I don't see things through. (R)
10. I shirk my duties (R)

Note: (R) = reverse coded.

Appendix G

In-role Behavior (Williams & Anderson, 1991)

1. Adequately completes assigned duties
2. Fulfills responsibilities specified in job description
3. Performs tasks that are expected of him/her
4. Meets formal performance requirements of the job
5. Engages in activities that will directly affect his/her performance evaluation
6. Neglects aspects of the job he/she is obligated to perform (R)
7. Fails to perform essential duties (R)

Note: (R) = reverse coded.