Running Head: STEREOTYPE THREAT EFFECTS DURING TRAINING

Brain Drain? An Examination of Stereotype Threat Effects during Training on Knowledge Acquisition and Organizational Effectiveness

James A. Grand
University of Maryland

Supplemental materials: http://dx.doi.org/10.1037/apl0000171.supp

Author's note:
This manuscript was based on the author's doctoral dissertation completed at Michigan State University under the supervision of Ann Marie Ryan. The author gratefully acknowledges Ann Marie Ryan, Tim Pleskac, Neal Schmitt, Steve W.J. Kozlowski, Georgia T. Chao, Paul Hanges, Jennifer Wessel, and Goran Kuljanin for their helpful comments, feedback, and insightful discussions on previous versions of this work. Portions of this work were previously presented at the 30th annual meeting of the Society for Industrial and Organizational Psychology (2015).

Citation:
Grand, J.A. (2017). Brain Drain? An examination of stereotype threat effects during training on knowledge acquisition and organizational effectiveness. *Journal of Applied Psychology, 102*, 115-150.

Abstract

Stereotype threat describes a situation in which individuals are faced with the risk of upholding a negative stereotype about their subgroup based on their actions. Empirical work in this area has primarily examined the impact of negative stereotypes on *performance* for threatened individuals. However, this body of research seldom acknowledges that performance is a function of *learning*—which may also be impaired by pervasive group stereotypes. This study presents evidence from a three-day self-guided training program demonstrating that stereotype threat impairs acquisition of cognitive learning outcomes for females facing a negative group stereotype. Using hierarchical Bayesian modeling, results revealed that stereotyped females demonstrated poorer declarative knowledge acquisition, spent less time reflecting on learning activities, and developed less efficiently organized knowledge structures compared to females in a control condition. Findings from a Bayesian mediation model also suggested that despite stereotyped individuals "working harder" to perform well, their underachievement was largely attributable to failures in learning to "work smarter." Building upon these empirical results, a computational model and computer simulation is also presented to demonstrate the practical significance of stereotype-induced impairments to learning on the development of an organization's human capital resources and capabilities. The simulation results show that even the presence of small effects of stereotype threat during learning/training have the potential to exert a significant negative impact on an organization's performance potential. Implications for future research and practice examining stereotype threat during learning are discussed.

Keywords: stereotype threat, learning, training, knowledge acquisition, knowledge structures, Bayesian analyses, computational modeling

> *In one physics class, the teacher announced that the boys would be graded on the 'boy curve,' while the one girl would be graded on the 'girl curve'; when asked why, the teacher explained that he couldn't reasonably expect a girl to compete in physics on equal terms with boys.* (Pollack, 2013)

The impact of group stereotypes and discriminatory attitudes towards targeted individuals has been of interest to researchers and policy makers for decades (Marx, Brown, & Steele, 1999). In recent years, this research has been heavily influenced by Steele and Aronson's (1995) conceptualization of *stereotype threat* (ST), a situation in which individuals are faced with "the risk of confirming, as self-characteristic, a negative stereotype about one's group" based on their actions (p. 797). ST theory posits that the presence of culturally recognized stereotypes coupled with environmental factors that make salient or provide individuals opportunities to validate the personal relevance and accuracy of such stereotypes gives rise to a host of maladaptive psychological responses (Schmader, Johns, & Forbes, 2008; Steele, 1997; Walton, Murphy, & Ryan, 2015). The greatest proportion of empirical work on ST has examined *performance/evaluation outcomes* that assess stereotyped group members' ability to apply existing domain-relevant knowledge, skills, or abilities (KSAs) to demonstrate competence or expertise. For example, meta-analytic estimates of ST effects on cognitive test performance from more than 100 studies across race, gender, and domain report evidence for small-to-moderate effect sizes in support of this predicted pattern of results (Cohen's d ~ .2 to .3, Nguyen & Ryan, 2008; Walton & Cohen, 2003; Walton & Spencer, 2009). In the organizational psychology literature, the potential implications of ST have tended to emphasize its potential to create inequities in personnel selection, assessment, and performance evaluation practices (Roberson & Kulik, 2007; Steele & Davies, 2003; Walton et al., 2015). However, many have argued that these consequences are not likely to be of concern (e.g., Cullen, Hardison, & Sackett, 2004; Cullen, Waters, & Sackett, 2006; Sackett, Hardison, & Cullen, 2004; Sackett, Schmitt, Ellingson, & Kabin, 2001; Sackett & Ryan, 2012; Stricker & Ward, 2004, 2008). Such critiques frequently raise statistical (e.g., small empirical effect sizes, dearth of evidence of ST effects from archival data or field research), methodological (lack of fidelity between ST in the lab versus the "real world"), and practical (existence of policies, standards, and legal regulations designed to minimize discriminatory assessment practices, e.g., Grand, Gulobovich, Ryan, & Schmitt, 2013; Gulobovich, Grand, Ryan, & Schmitt, 2014) concerns as evidence against the relevance of ST for organizations and applied psychological research.

These are valid critiques; however, concluding that ST has little practical significance for organizations may be premature. First, there has been relatively little research on ST examining organizationally relevant outcomes other than those associated with the demonstration of individual competence or expertise (e.g., cognitive test performance, task performance, etc.). As alluded to by the opening quotation of this paper, one important domain in

which ST effects have been largely unexplored are *learning/training outcomes* in which individuals participate in non-evaluative activities intended to develop (rather than demonstrate) competency or expertise through acquisition and retention of KSAs. Learning, training, and the accrual of domain expertise impact both *what* performance outcomes individuals can achieve as well as *how* those outcomes are attained; that is, learning/training helps individuals "work smarter, not harder." A common theme of the modern organization is the importance of lean and flexible workforces capable of fulfilling multiple roles (Cascio, 1995; Kozlowski, 2012; Kraiger, Ford, & Salas, 1993; Ployhart & Moliterno, 2011); consequently, understanding impediments to expertise development represents a critical determinant of the success and effectiveness of an organization's human resource strategy. The potential for disruptive ST effects in learning/training contexts is also noteworthy given that many critiques emphasize the strict regulations present in high stakes performance/evaluation contexts as reasons why ST is unlikely to be concerning for organizations in hiring, assessment, and performance review practices. Though there are reasons to question whether such policies are effectively at removing conditions conducive to ST (Walton et al., 2015), learning/training practices in organizations already tend to be much less structured (e.g., peer learning, on-the-job training, etc.), standardized, monitored, or regulated by organizational and legal precedents. In sum, learning/training contexts are central to organizational effectiveness, but the acquisition of task- and job-relevant KSAs could suffer if detrimental effects engendered by ST are capable of impeding the development of knowledge and expertise.

An additional reason to question the conclusion that ST is irrelevant to organizations is that research has not been successful at conveying the significance of "small effects" within a broader organizational system. The modal experimental methodology for ST research has been single time-point, cross-sectional study designs examining differences between individuals experiencing versus not experiencing ST. These observed mean differences are then condensed into a summary effect size statistic (e.g., Cohen's *d*) and used to support or refute the presence and practical significance of ST effects (e.g., Stricker & Ward, 2004). However, such methodological and inferential practices are neither sufficient nor effective at conveying the practical impact of "real-world" phenomena that are inherently dynamic and whose consequences are emergent (Grand, Braun, Kuljanin, Kozlowski, & Chao, in press; Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013). For example, when describing the societal implications of empirical research on small and subtle sex biases, Eagly (1995) noted:

> The evaluation of the…importance of sex-related differences should not end with the translation of them into metrics that are easily understood. In practical terms, the importance of a difference depends on the *consequences of the behavior in natural settings*. (p. 152, italics added)

Questions of "practical significance" are poorly captured by a simple effect size statistic; instead, they must be examined with respect to how such differences may impact the critical functions and outcomes of the system in which those differences reside. In the context of ST, it is thus inadvisable to draw conclusions about the implications of ST for organizational effectiveness based solely on the magnitude of cross-sectional meta-analytic effect sizes. Adequately evaluating the impact of ST requires efforts to understand where and how such effects unfold and accumulate *over time* and *across levels* to influence criteria that are important to employees and organizations.

In light of these limitations, the present study contributes to the ST literature in three key ways. First, this research demonstrates that the presence of negative domain stereotypes are capable of adversely influencing threatened individuals' capacity to effectively and efficiently acquire knowledge needed to perform a task. Second, a hierarchical Bayesian regression approach is used that permits a direct and straightforward interpretation of the extent to which ST influences differences in learning outcomes across people and over time. Lastly, a computational model and computer simulation are presented demonstrating that even small-to-moderate effects of ST experienced during learning/training can result in substantial consequences for an organization's human capital. From a research perspective, this work extends existing ST literature in the organizational sciences by focusing on factors that contribute to subgroup differences in learning/training—as opposed to performance—outcomes. Additionally, the use of Bayesian analyses provides a flexible and highly adaptable framework for replicating and accumulating results related to ST effects during learning in future research. From a practical perspective, the computational model and computer simulation provides the first known attempt to situate and quantify the consequences of ST effects on organizational effectiveness. This work highlights the importance of evaluating training practices so that all individuals have equal opportunity to benefit from learning/training and to increase the likelihood that organizations maximize investments in employee development programs.

Stereotype Threat Mechanisms and Relation to Learning Outcomes

Stereotype threat is proposed to exist when a person faces a negative expectation about the relationship between their group membership and an ability domain that is irreconcilable with positive expectations about their self-and-group and their self-and-ability (i.e., "My group does not have this ability; I am like my group, but I have this ability." Schmader & Beilock, 2012; Schmader et al., 2008). A critical assumption of ST theory is that the effects of the threat are capable of extending to multiple different types of tasks and outcomes beyond just task performance. For example, the stereotype "women are less proficient at mathematics" has been shown to elicit effects consistent with ST when women complete a difficult mathematical ability test (performance directly evaluated, Spencer, Steele,

& Quinn, 1999), teach students whose mathematical ability is later assessed (performance not directly evaluated, Beilock, Gunderson, Ramirez, & Levine, 2010), or attempt to memorize novel mathematical operations (performance not primary goal, Rydell, Rydell, & Boucher, 2010). In all cases, the discrepancy between the negative group stereotype and one's perceptions of self elicits a state of cognitive dissonance and unresolved tension for the stereotyped individual that he/she is motivated, yet struggles, to dispel. In response, a variety of psychological resources and processes are engaged to aid the resolution effort.

Of relevance to learning/training domains is research suggesting that the added regulatory strain introduced by ST taxes working memory (Schmader et al., 2008). Working memory characterizes an individual's limited capacity "cognitive workspace" that coordinates attention among immediately relevant thoughts, operations, and information (Baddely & Hitch, 1974; Kane, Bleckley, Conway, & Engle, 2001). Working memory is central to numerous intellective activities relevant to learning, including reading comprehension, information synthesis, and advanced reasoning (see Feldman Barrett, Tugade, & Engle, 2004). Consequently, a primary mechanism through which ST is proposed to impact knowledge acquisition is through elicitation of irrelevant affective and cognitive stimuli that "hijack" portions of working memory, thereby leaving fewer cognitive resources for encoding domain-relevant KSAs and procedures during training (Beilock, Rydell, & McConnell, 2007; Schmader & Johns, 2003; Wraga, Helt, Jacobs, & Sullivan, 2006).

The motivations and strategies employed by learners during training are also key predictors of learning and training effectiveness (e.g., Colquitt, LePine, & Noe, 2000; Keith & Frese, 2005). Existing research offers mixed conclusions regarding the impact of ST on such processes, with some studies reporting reduced task-directed efforts (as a result of diminished performance expectations or self-handicapping, e.g., Schimel, Arndt, Banko, & Cook, 2004; Stone, 2002) and others increased task-directed efforts (to avoid fulfilling the negative stereotype, e.g., Jamieson & Harkins, 2007; Oswald & Harvey, 2000). The underlying mechanism relating ST to motivation appears to involve how individuals interpret and react to negative feedback following an error. In general, errors tend to focus attention towards identifying what led to a negative outcome and how it can be prevented in the future. However, this focus is only adaptive when the increased arousal elicited by negative feedback can be targeted towards task improvement and away from affective self-evaluations (Harkins, 2006; Kluger & DeNisi, 1996). Some research suggests that this cognitive activity is more difficult for and less readily accomplished by individuals experiencing ST (Keller & Dauenheimer, 2003; Krendl, Richeson, Kelley, & Heatherton, 2009; Wraga et al., 2006). As a result, stereotyped

individuals may be more likely to disengage prematurely from learning-related activities that involve acknowledging and rectifying errors (Mangels, Good, Whiteman, Maniscalo, & Dweck, 2012).

Whether disengagement is more or less likely to occur is also related to the nature of the task environment. When task activities are difficult, novel, or ambiguous, a narrowing of attention often inhibits one's ability to identify complex relationships and problem solutions (Schneider & Shiffrin, 1977). These circumstances should further exacerbate the frustrations and anxieties that accompany ST, and are consistent with evidence that ST effects are most pronounced on difficult tasks (Nguyen & Ryan, 2008; Steele & Aronson, 1995). This is particularly problematic in the context of learning/training environments, where material is often experienced for the first time and attending to and building upon errors is a hallmark of learning (Bell & Kozlowski, 2008). Thus, ST may undermine targeted individuals' ability to effectively regulate and sustain learning-related efforts during training.

Although few in number, recent research lends support to these processes and the likelihood that negative stereotypes can impair the learning process. Taylor and Walton (2011) and Rydell, Rydell, and Boucher (2010) found that individuals attempting to acquire basic declarative knowledge (e.g., definitions of rare words, unfamiliar mathematical rules, patterns of abstract symbols) under ST were less proficient at later recalling that information compared to individuals learning these same facts in the absence of ST. Rydell, Shiffrin, Boucher, Van Loo, and Rydell (2010) also demonstrated that ST impaired skill-based learning on simple perceptual/visual search tasks. Lastly, Mangels et al. (2012) used EEG measures of individuals' emotional reactions to feedback during a learning activity to examine motivational changes in threatened learners. Their results revealed that females under ST tended to react more strongly and were poorer at regulating arousal towards negative feedback received during practice, which contributed to earlier disengagement during subsequent learning opportunities relative to non-threatened females. This pattern of results is further consistent with findings that individuals experiencing ST tend to focus more attention on performance errors and perceive negative feedback as validation of group stereotypes (Forbes, Schmader, & Allen, 2008; Grimm, Markman, Maddox, & Baldwin, 2009; Jamieson & Harkins, 2007).

The emerging empirical evidence thus appears to support the adverse effects of ST on knowledge acquisition. However, this research possesses a number of limitations that limit its generalizability to organizational learning/training contexts. First, the most common operationalization of learning in this research has involved evaluating the number of items individuals are able to correctly recall on a single follow-up assessment. Relatedly, participants were usually tasked with learning nonsensical or decontextualized information (e.g., unfamiliar words and definitions, abstract mathematical rules or symbols) that was not intended to transfer or generalize to future task

performance. Such measures are consistent with examinations of declarative knowledge acquisition. Though attaining declarative knowledge is a foundational element of the learning process (Anderson et al., 2004), training in organizations is seldom solely concerned with the acquisition and recall of disjointed facts. Instead, the goal of most training programs is the development of more advanced learning outcomes (knowledge organization, strategic knowledge, etc.) that facilitate an individual's ability to generalize learning to performance-relevant activities (Kraiger et al., 1993). In other words, organizational training is most commonly concerned with improving individuals' *expertise* (an understanding of task procedures and "how things work"), not simply declarative knowledge.

A third limitation of the existing research is that learning often occurred in "single-shot episodes" in which participants were exposed to the stimulus material once. The development of knowledge and expertise is an iterative process that requires time, exposure, and rehearsal (Chase & Simon, 1979; Anderson et al., 2004). The use of single exposure/measurement designs does not allow observations about rates of change in knowledge acquisition as a result of practice and experience and thus fail to provide a complete view of the learning process. A final limitation of previous research is that many organizational training environments integrate principles of instructional design (e.g., learning objectives, directive feedback, etc.) to provide learners a framework for sense-making (Goldstein & Ford, 2002). Few such provisions have been incorporated or made explicit in existing research on ST during learning, thus making it difficult to generalize these results to common organizational training experiences. In sum, the current evidence for ST effects on learning outcomes extends to the immediate acquisition of decontextualized and goal-irrelevant declarative knowledge following one exposure in an unstructured learning environment.

Research Hypotheses

The present study advances research in this domain by examining knowledge acquisition and learning engagement for female learners facing ST during a three-day training program designed to teach a complex decision-making task. Cognitive learning outcomes were the primary dependent variables of interest. As opposed to skill- or affective-based outcomes, cognitive learning outcomes reflect the procurement, organization, and synthesis of information that directly contributes to an individual's capacity to interpret, reason, or draw inferences (Kraiger et al., 1993). Cognitive learning outcomes are common criteria for many organizational training settings as well as the most common performance domain tapped by existing ST research on group differences on evaluative assessments (Nguyen & Ryan, 2008). Drawing from Kraiger et al.'s (1993) training evaluation typology, differences in three different cognitive learning outcomes are examined: declarative knowledge (i.e., acquisition of facts and definitions),

knowledge structures (i.e., how information is organized), and metacognitive activity (i.e., critically interpreting what/how knowledge is acquired to better "learn how to learn").

The allocation of working memory capacity to regulating affective self-appraisals elicited by ST should make it more difficult for threatened individuals to attend to and encode new information. As in previous research, this should be observable in measures of threatened participants' declarative knowledge (Taylor & Walton, 2012; Rydell, Rydell, & Boucher, 2010). However, decrements in working memory should also impair the ability for learners to maintain information in an activated state long enough to be integrated and organized into efficient and effective mental models that facilitate future task-relevant performance (cf., MacDonald, Just, & Carpenter, 1992; Whitney, 1991). Lastly, working memory deficits should also leave fewer cognitive resources to devote towards thinking critically about training activities and how/where to focus attention to maximize learning outcomes—an activity critical to developing more advanced comprehension of a task domain (Bell & Kozlowski, 2008). Thus, it is predicted that:

Hypothesis 1: Females who participate in training under conditions of ST will acquire less declarative knowledge than females who participate in training under control conditions.

Hypothesis 2: Females who participate in training under conditions of ST will develop less efficiently and effectively organized knowledge structures than females who participate in training under control conditions.

Hypothesis 3: Females who participate in training under conditions of ST will report less metacognitive activity directed towards learning than females who participate in training under control conditions.

The influence of ST on learning engagement is also examined. Although individual differences play a role in a trainee's motivation to learn, learning engagement can also be influenced by factors in the training environment (Goldstein & Ford, 2002). One particularly important characteristic of training environments is the degree to which learners are passively (lecture, reading, watching videos, etc.) versus actively (role play, simulation, behavioral modeling, etc.) involved in learning. Many training programs blend these design elements, often utilizing passive techniques to facilitate acquisition of basic declarative knowledge and active techniques to develop a deeper understanding of task principles and procedures. In the present study, a similarly integrated training program is employed that permits participants to both passively (through reading and study) and actively (through practice and exploration) engage in learning. Consistent with the broader ST research, it is predicted that learners facing ST during training should be more likely to disengage from and exert less effort during both modes of learning:

Hypothesis 4: Females who participate in training under conditions of ST will demonstrate less effort engaging in passive learning activities than females who participate in training under control conditions.

Hypothesis 5: Females who participate in training under conditions of ST will demonstrate less effort engaging in active learning activities than females who participate in training under control conditions.

Lastly, the generalization of knowledge acquired during training to task-relevant performance is examined. The transfer of learning to task achievement and job performance is arguably the single-most important indicator of training effectiveness and often the criteria of greatest interest to organizations. While there are numerous determinants of whether training experiences will positively influence subsequent task or job performance (Blume, Ford, Baldwin, & Huang, 2010), the completeness and quality with which material is learned should be a critical factor. Consequently, deficiencies in the learning process attributable to ST should be observable as diminished task performance (Taylor & Walton, 2011):

Hypothesis 6: Females who participate in training under conditions of ST will demonstrate poorer task performance than females who participate in training under control conditions.

To summarize, this study investigates the degree to which cognitive learning outcomes during a three-day training experience are influenced by the presence of a negative domain stereotype for female learners. Drawing from research on the psychological mechanisms of ST, it is predicted that detriments to working memory capacity related to ST impair the acquisition of both basic and advanced learning outcomes for threatened individuals. Additionally, the nature of the learning environment (novel/difficult material, error focus) is predicted to increase the likelihood that threatened females will disengage and exert less effort during learning opportunities. Finally, disparities in these learning outcomes should translate into demonstrable performance differences on a task requiring knowledge that could have only been acquired from training to effectively complete.

METHODS

To generate conditions suitable for ST, females served as the subgroup of interest and mathematical/analytical reasoning as the ability domain. This combination was chosen as stereotypes favoring men in mathematical/analytical domains have been frequently documented and are commonly recognized in Western cultures (Halpern et al., 2007). Furthermore, research indicates that females' cognitive and behavioral performance outcomes are reactive to negative stereotypes about mathematical/analytical aptitude even in instances where a task does not explicitly involve such operations (e.g., Jamieson & Harkins, 2007; Rydell, Shiffrin, et al., 2010).

Participants

Participants were 198 undergraduate students recruited from introductory and upper-level psychology courses. Given the focus on negative stereotypes towards female achievement on mathematical tasks and the

within-group nature of ST effects (i.e., females under ST versus females not under ST), women were the primary group of analytic interest and were therefore over-represented in the sample relative to men (total N_{female} = 145).[1] A total of 158 participants (79.8%) completed all three days of the experiment, 114 of whom were female (N_{ST} = 58, $N_{control}$ = 56). Slightly higher attrition rates were observed between Days 2 and 3 (12.22%) relative to Days 1 and 2 (9.09%) in the sample. However, a two-way ANOVA on number of days attended revealed no main effects for sex (F(1,194) = .015, *ns*) or condition (F(1,194) = .004, *ns*), nor a sex by condition interaction (F(1,194) = .820, *ns*), indicating that attrition was not differentially influenced by the ST manipulation or sex.

All individuals were compensated with course credit for participation in the study. As additional incentive, participants whose cumulative scores from three end-of-day performance trials were among the top 10% of performers received a $60 cash prize. Because women facing ST were expected to do more poorly, cash prizes were awarded to the top 10% of male and female performers within each condition.

Experimental Task

A modified version of the computer-based Tactical Naval Decision Making task (TANDEM, Weaver, Bowers, Salas, Cannon-Bowers, 1995) was used as the training stimulus and experimental platform. TANDEM is a dynamic information-processing and decision-making task set in the context of a low-fidelity radar tracking simulation and has been used to study a variety of outcomes related to individual-level learning and training outcomes (e.g., Bell & Kozlowski, 2002; Bell & Kozlowski, 2008; Kozlowski et al., 2001). Simulation-based training methods are increasingly common learning tools in both education and industry (Bell, Kanar, & Kozlowski, 2008). Some reports indicate that upwards of 97.5% of business schools use simulation gaming as part of their standard curricula and at least 75% of organizations in the United States with more than 1,000 employees use some form of business simulation for hiring/training purposes (Faria & Nulsen, 1996; Faria, 1998). The TANDEM task environment presents participants with an interface resembling a radar screen used by an air traffic controller. A number of unidentified “contacts” are displayed and move about as though they are vehicles tracked on the radar. The goal for participants is to correctly process as many contacts as possible during the allotted time period. To process a contact in this study, individuals were required to evaluate diagnostic cues (e.g., speed, heading, radio signature) about each contact and use that information to classify its Type (Air, Surface, or Submarine), Class (Civilian or Military), and Intent (Peaceful or Hostile). Based on those classification, participants then made a single final engagement decision (Clear, Warn, or Mark) that signaled a contact had been processed and removed it from the screen. Correctly processing a contact earns points during each trial, while incorrectly processing a contact results in a loss of points.

Making accurate classification decisions (Type, Class, and Intent) necessitated learning 27 declarative facts/definitions related to interpreting the diagnostic cues available for each contact (e.g., "a contact with speed greater than 35 knots is an Air Type," "a contact with identification Tango has Hostile Intent," etc.). To make accurate final engagement decisions, participants also needed to learn rules of engagement dictating how contacts should be processed once classified. The rules of engagement encompassed 12 associative rules that linked a contact's unique classification to a specific final engagement action (e.g., "if a contact is Air, Civilian, and Peaceful, then it should be Warned"). Both the diagnostic information and rules of engagement were held constant across all trials. More details on the classification and engagement decision rules used in the present study can be found in the Supplemental Materials accompanying this manuscript.

Training Environment

The organization, presentation, and experience of information by learners occurred in an exploratory learning environment. Exploratory learning involves self-directed active learning that encourages leaners to experiment with instructional content to infer the principles, rules, and mechanisms of a given operational domain (Kamouri, Kamouri, & Smith, 1986). Exploratory learning was implemented by providing participants with multiple opportunities to engage in practice trials with TANDEM during each day's training session coupled with feedback about the activities they practiced. Additionally, individuals had access to an operations manual containing all the relevant information needed to successfully operate TANDEM prior to every trial. The manual included information about basic task functions (e.g., how to operate the task, scoring rules, overall objectives), processing contacts (e.g., declarative information related to classification decisions and the rules of engagement), and task strategies (e.g., tactics for monitoring defensive perimeters, prioritizing targets). Participants were permitted to use the practice trials and study time to engage in self-directed task learning however they saw most beneficial.

A downside to purely exploratory learning environments is that learners can become overwhelmed and fail to grasp the desired instructional material if no guidance is provided. Learning recommendations were thus provided to participants during all training sessions at set points throughout the experiment. The recommendations were presented as reflective questions that encouraged participants to explore different ways of completing the task and to develop their own personal understanding of the knowledge/procedures needed to perform effectively. The questions focused on areas directly relevant to task completion and served as self-assessment anchors to guide learning during the practice trials (e.g., "Have you learned how to interpret the cues needed to make accurate decisions about

the target's Type?"). Previous research has shown that similar prompts are effective at improving knowledge acquisition in exploratory learning environments (Bell & Kozlowski, 2008; Debowski, Wood, & Bandura, 2001).

Procedure

Participants enrolled in the simulated training program through an online scheduling system. Double-blind assignment was used to assign participants to either the control or ST condition. Each simulated training program took place over three consecutive days. Participants attended the training in mixed-sex sessions of 8-12 people seated at personal computers in a large room. At the beginning of Day 1, individuals provided informed consent and completed a pre-trial assessment of working memory.[2] Participants then watched an introductory video outlining the sequence of events for the simulated training program and basic information about the TANDEM task interface. The experimental manipulation instructions were also presented for the first time during the video. Lastly, the video instructed participants that only points earned during the final performance trials of each day would be taken into consideration when determining the winners of the monetary prizes and that all other trials should be used to learn how best to complete the task.

The working memory assessment and introductory video were only administered on Day 1 of the experiment; the remainder of training was identical across all three days. Prior to engaging in the first practice trial, participants completed a short acclimation trial. During this period, individuals had access to the operations manual for 30 seconds and then completed a one minute trial in TANDEM. The purpose of this trial was to (re)orient participants to the computer equipment and how to perform common task operations. No feedback was given for activities during the acclimation trial. Following the acclimation period, participants completed six practice trials. All practice trials followed a standard sequence of two minutes for studying the operations manual, five minutes of hands-on practice with the radar interface, and one minute to review descriptive feedback about what they accomplished during that trial (i.e., points scored, number of classification/engagement decisions correct, number of targets processed, etc.). Each practice trial consisted of 21 targets comprising a task scenario. The same task scenario was used for all practice trials in a given day, with different task scenarios used each day. As part of the instructions provided during the practice trials, participants were once again reminded that they should use their time and the feedback they received to direct their learning activities as they saw best and that scores were not important.

After the sixth practice trial, participants completed a single performance trial. The performance trial was similar to the practice trials, though a number of changes were introduced to make the task more difficult and cognitively demanding (changed scoring algorithm, increased number of targets to process, and introduced more

"high priority" targets). The length of the performance trial was increased from five minutes to eight minutes and participants received only one minute to view the operations manual prior to the scenario. All participants were informed that these trials were more challenging and received instructions describing these critical differences before beginning the performance trial. The complexity of the performance trials was increased so that even if individuals who experienced ST during learning were capable of doing well on the practice trials by simply memorizing the correct decisions for contacts in a given task scenario, they would experience difficulties during the performance trial because they had not acquired the proficiency needed to generalize their knowledge. Following the final performance trial of the day, participants completed a series of post-trial questionnaires.

Experimental manipulation. The ST manipulation used was adapted from previous research (Beilock et al., 2007; Rydell, Shiffrin, et al., 2010; Spencer et al., 1999) and delivered via text presented on an individual's computer screen as well as audibly through headphones worn by participants during the experiment. The manipulation instructions for participants in the ST condition indicated that the purpose of the study was to examine possible explanations for why women tend to perform more poorly than men on tasks involving mathematical and analytical aptitude. Individuals were further informed that one reason for this finding may be that women have more difficulty distinguishing relevant information needed to solve a problem from irrelevant or distracting information, and that the TANDEM simulation was designed to examine differences in these skills. Instructions for control condition participants noted that the purpose of the experiment was to examine individual differences in learning and problem-solving skills; no mention of sex, sex differences, or any diagnostic considerations were made. In addition to the instructional prompts, all individuals were asked to report their sex for identification purposes prior to beginning the first practice trial each day. Participants in both conditions received a long version of these instructions prior to the first practice trial each day and a shorter version prior to the third and fifth practice trials.

The manipulation instructions generated controlled conditions under which ST effects could be observed. However, the detriments of ST arise from individuals' heightened vigilance to and interpretations of information from the environment about whether they are valued and/or able to succeed in a given situation (Schmader et al., 2008; Steele & Aronson, 1995). In their review of ST in organizational settings, Walton et al. (2015) characterize such environmental information as *identity contingency cues* and describe a number of such cues commonly found in organizational settings. Two of these cues—critical feedback and fixed-ability messages—are particularly relevant to organizational learning/training contexts and were reflected in the simulated training program of this study as well. Nearly all organizational training programs provide *critical feedback* that convey areas where individuals are and are

not proficient. The feedback following each practice trial provided similar information to learners in the present study. Many researchers have noted that such critical feedback can be anxiety-provoking for learners (Bell & Kozlowski, 2008), but in the case of learners sensitized to perceiving threatening identity contingency cues, such messages can be interpreted as evidence that they are failing to grasp the training material and are thus fulfilling the negative group stereotype. *Fixed-ability messages* convey beliefs that individuals' capabilities are fixed ("you either have it or you don't") and that those without the requisite KSAs cannot be successful. In organizations, such beliefs may be reflected in cultural norms and/or performance management systems that emphasize current achievement over growth and development (Roberson & Kulik, 2007; Walton et al., 2015). In the simulated training program, the instructional text provided to learners suggested that those who did not possess the ability to accurately distinguish important problem-relevant from problem-irrelevant information would not perform well on the performance trials and earn the cash prize. No indications were conveyed to participants that these skills could increase through practice with TANDEM; furthermore, and as part of the experimental manipulation, individuals in the ST condition were informed that females tended to be less proficient at this skill in general. In sum, the ST manipulation employed in this experiment highlighted a negative expectation for females in the domain, while features of the simulated learning/training environment with fidelity to those found in organizations (e.g., critical feedback, fixed-ability messages) afforded opportunities for targeted females to evaluate their standing against the group stereotype.

Measures

Manipulation check. A 7-item self-report measure of perceived ST (α = .67) was administered at the end of Day 3 to assess the efficacy of the ST manipulation. Participants were asked to rate the extent to which they believed negative expectations existed regarding their gender's performance in the experimental task (e.g., "A negative opinion exists about how members of my gender should perform on this type of task."). Although Steele (1997) contends that individuals need not be consciously aware of a negative stereotype to experience ST, previous research has found that threatened individuals often do perceive such threats and that self-report measures can be useful for assessing the saliency of ST manipulations (Grand, Ryan, Schmitt, & Hmurovic, 2011). Responses to the measure were provided on a 5-point Likert-type scale (1—*Strongly disagree* to 5—*Strongly agree*).

Declarative knowledge. An 11-item, multiple-choice test of declarative knowledge pertaining to TANDEM was completed by participants during the post-trial measurement period each day. The test questions focused on accurately identifying the diagnostic information needed to make classification decisions as well as the rules of engagement ("If a target's characteristics are Speed = 35 knots and Altitude/Depth = 15 feet, which of the following

actions should you take?"). To minimize reliance on memory of past administrations when answering the questions, a different set of items were administered each day (Cronbach's alpha: Day 1, α = .60; Day 2, α = .65; Day 3, α = .69; test-retest reliabilities: Day 1-2, *r* = .47, Day 2-3, *r* = .64, Day 1-3, *r* = .46).

Knowledge structures. Synonymous with mental models, cognitive maps, or schema, knowledge structures capture how individuals make associations among concepts, facts, functions, and other knowledge objects (Glaser, 1990; Schoenfeld & Herrmann, 1982). Qualitative differences in knowledge structure organization are useful for examining how individuals "make sense of" a content domain. For example, knowledge structure differences have been used to distinguish between domain experts and novices (Chi, Glaser, & Farr, 1988) as well as identify differences in concept learning and categorization based on group membership (Medin et al., 2006).

Observations of participants' knowledge structures were collected each day during the post-trial measurement period. Participants provided ratings of perceived similarity among 16 concepts critical to processing contacts in TANDEM (Table 1). The list of concepts was adapted from previous research examining knowledge structures in the TANDEM task environment (Kozlowski et al., 2001). For each rating, individuals were presented with two concepts and asked to indicate how related they were to one another using a 9-point scale (1—*not at all related* to 9—*highly related*). Similarity ratings were provided for every unique pairwise combination of concepts, resulting in 120 ratings per knowledge structure. Similarity matrices were analyzed using the Pathfinder software program and algorithm to construct visual representations of the associative networks (see Dearholt & Schvaneveldt, 1990, and Interlink, 2011, for more information about Pathfinder and its algorithms).

Metacognitive activity. Metacognitive activity reflects the extent to which individuals "think about their thinking" and is considered an advanced cognitive learning outcome that manifests as learners gain a greater appreciation of the content domain (Kraiger et al., 1993). Self-reported metacognition was assessed during the post-trial measurement period each day using a 12-item measure adapted from Ford, Smith, Weissbein, Gully, and Salas (1998). Each question asked participants to indicate the degree to which they consciously reflected on their learning and performance activities during the task (e.g., "I carefully determined what to study and practice in order to improve weaknesses identified in previous trials"), with responses given on a 5-point scale (*1—Never* to *5—Constantly*). Cronbach's alpha for the measure was *α* = .86, .91, and .95 on Days 1, 2, and 3, respectively.

Learning engagement. Trace measures of learning engagement were extracted from behavioral data automatically recorded by TANDEM during each experimental trial. Two indicators were employed to examine how participants made use of their discretionary practice time during each trial's learning opportunities. To examine

engagement during passive learning activities, the amount of time participants spent reviewing the manual sections related to processing contacts and strategy information during the study phase prior to each practice trial was recorded. To examine engagement during active learning activities, the total number of contacts processed by a participant during each practice trial was evaluated.

Task performance. The total number of points earned on the final performance trial each day was computed using a scoring algorithm made available to all participants. During all performance trials, 100 points could be earned for every target correctly processed (i.e., all three classification decisions and engagement decision correct), while 150 points were lost for every contact incorrectly processed (i.e., any classification decision or engagement decision incorrect) or that crossed through a defensive perimeter. Additionally, the number of contacts processed during the performance trials was also recorded.

Analysis Plan

Statistical analyses. Bayesian parameter estimation methods were used for all inferential statistical tests. Although Bayesian statistics have not been readily adopted in the organizational sciences, they offer a number of advantages over conventional null hypothesis testing (cf., Kruschke, 2015; Kruschke, Aguinis, & Joo, 2012; Zyphur & Oswald, 2015). First, Bayesian parameter estimation allows direct interpretation of a finding based on the observed data; that is, Bayesian methods allow researchers to directly evaluate the "believability" of a parameter estimate based on the data (i.e., p(estimate|data)) rather than evaluate whether the data were likely to have been observed if the estimated relationship is zero (i.e., p(data|estimate)). Second, Bayesian estimation relies on interpretation of credibility intervals (i.e., the range of parameter estimates with the highest degree of believability given a prediction and observed data) rather than confidence intervals (i.e., the range of parameter estimates within which the true population estimate should reside). While confidence intervals can be useful for evaluating estimates when the number of observations is very large, they are dependent on the sample size and stopping rule used (i.e., the point at which the researcher intended to stop collecting sample data) and are less interpretable under typical experimental circumstances. Finally, the results obtained through Bayesian methods are conducive to empirical replication as the estimates obtained from a research study can be explicitly and quantitatively integrated into future research as prior beliefs. Consequently, no p-values or tests of statistical significance are reported for any regression coefficients. Instead, Bayesian analyses rely on interpreting posterior distributions reflecting the believability of values for a given parameter estimate (e.g., mean, regression coefficient, etc.) based on the observed data and predictions about the

values of those parameter estimates. Appendix A provides additional details on conducting and interpreting Bayesian analyses.

Bayesian estimation of group differences was used to evaluate between-condition differences in the manipulation check measure (Kruschke, 2013). For the hypotheses examining differences in declarative knowledge, metacognitive activity, the learning engagement measures, and task performance over time, a two-level hierarchical Bayesian regression model that allowed for varying slopes and intercepts was used (Gelman & Hill, 2004). These models are identical to conventional random coefficient/multilevel models common in organizational research with the exception that they formally integrate predictions about the predicted coefficients into their computation (see Appendix A). Unless otherwise specified, the regression model used for all analyses was:

$$\text{Level 1: } DV_{it} = \pi_{0i} + \pi_{1i}(\text{Time}_{it}) + e_{it} \quad (1)$$

$$\text{Level 2: } \pi_{0i} = \beta_{00} + \beta_{01}(\text{Condition}_i) + e_i$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{Condition}_i) + e_i,$$

where DV_{it} indicates the dependent variable measured at time t for person i, Time_{it} reflects either Day or Trial depending on the level of measurement for the dependent variable, and Condition_i is a dummy-coded variable indicating the experimental condition of the participant (Control = 0, ST = 1). Time was coded so that zero reflected the first time period of measurement; thus, all coefficients are interpreted in relation to the intercept at Day 1 or Trial 1. The result of primary interest is the Time*Condition interaction (β_{11}), which reflects slope differences in the dependent variable between participants experiencing ST versus control participants during training.

Unstandardized coefficients and their 95% credibility intervals are reported for all regression parameters. Additionally, the proportion of the estimated posterior distribution above (or below) zero is given for each coefficient. Although it is inappropriate to interpret this index in the same manner as a conventional *p*-value, it can be useful as an indicator of the degree to which zero is a believable estimate. When zero is highly believable (i.e., zero falls well within the 95% credibility interval), this proportion will be closer to 50%; when zero is not believable (i.e., zero is not included in or is near the extremes of the 95% credibility interval), this proportion will be near 100%.

Given that nearly all previous research examining ST has employed cross-sectional/single time point designs in performance/evaluation contexts, there was no existing empirical research available to inform the prior distributions used in the regression models. Consequently, the prior distributions placed over the Level-2 regression coefficients were uninformative uniform distributions (Appendix A provides a full explanation and graphical representation of the models used in all analyses). Such a diffuse prior distribution corresponds with the belief that all

values within the range of the uniform distribution are equally likely. This ensures that the posterior distributions for the final parameter estimates are primarily driven by the observed empirical data rather than any values represented in the prior distributions. Markov Chain Monte Carlo (MCMC) sampling methods (Gibbs sampling) were used to generate no fewer than 30,000 representative values for each parameter's posterior distribution. Following the recommendations of Gelman et al. (2013) and Kruschke (2015), a total of three MCMC chains were burned in, checked for convergence (Gelman-Rubin statistic = 1), and run long enough to achieve an effective sample size of at least 10,000 for all pertinent parameters. Analyses were conducted in R (R Core Team, 2015) using the *runjags* package (Denwood, in press) and JAGS software (Plummer, 2013) for MCMC sampling.[3]

Knowledge structure analyses. A qualitative approach was used to explore differences in knowledge structure development across experimental conditions. The primary motivation for examining knowledge structures in this study was to identify between-condition differences in categorical/thematic associations attributable to ST (see Medin et al., 2006, for a similar analytic approach). Aggregated knowledge structures for each condition were computed by averaging the similarity ratings provided by female learners in the control and ST conditions separately at each day. The Pathfinder network algorithm was then applied to create the final knowledge structures following each day's training containing the minimum number of observed links (PFnet(n-1, ∞), Interlink, 2011).

Prior to evaluating the knowledge structure data, the number of links produced in each participant's knowledge structure was examined to identify individuals who were likely not attending to the similarity rating task. Specifically, any individual network containing 120 links was excluded from the aggregate knowledge structure for that day as this reflected a participant who provided the same numeric rating for all 120 pairwise comparisons during the rating task. This procedure resulted in removal of 6 female participants (4 from the ST condition, 2 from the control condition). A computer error also resulted in loss of knowledge structure data at Day 2 for 8 female participants in the ST condition, though these participants' data were still included in the analyses for Days 1 and 3.

RESULTS

Table 2 provides a summary of the descriptive statistics and correlations among the quantitative study variables. Results from the Bayesian estimated group differences test of the manipulation check measure revealed that control condition participants (M = 2.88 [95% credibility interval = 2.75, 3.03], SD = .51 [.41, .62]) reported less perceived ST than participants exposed to the experimental ST manipulation (M = 3.29 [3.15, 3.42], SD = .50 [.40, .60]). This difference reflects a moderate-to-large effect size (d_a = -.79, [-1.20, -.41], 100% of posterior distribution

below zero[4]) and suggests that the manipulation was successful in generating the experience of ST for females in the experimental condition.

Declarative Knowledge

The percentage of items answered correctly on the declarative knowledge test each day (number of items correct / 11) was analyzed using multilevel logistic regression assuming a binomial likelihood distribution for the observed data. The top half of Table 3 presents the coefficient estimates for this model reported in log-odds. Both control and ST condition females achieved nearly equivalent levels of declarative knowledge by the end of Day 1 (β_{01} = .01, [-.28, .31], 54% of posterior distribution above zero). However, the rate of declarative knowledge acquisition was notably slower for females in the ST condition compared to those in the control condition (Figure 1). For control condition females, the probability of achieving a perfect score on the declarative knowledge measure increased by 62% each day (β_{10} = .47, [.31, .63], 100% of posterior distribution above zero); by comparison, the probability of achieving a perfect score for ST condition females increased by only 56% each day (β_{11} = -.25, [-.46, -.03], 99% of posterior distribution above zero). In other words, the rate of declarative knowledge acquisition for female learners in the control condition was nearly 11% greater than for females in the ST condition, thus supporting Hypothesis 1.

Knowledge Structures

Figure 2 presents the aggregate knowledge structures for females in the control and ST conditions at each day and Table 4 provides a summary of network descriptive statistics. A qualitative comparison of between-condition differences in the organization and development of knowledge structures over time revealed three key observations. First, a clear separation between the decision-making and task operation concepts (see Table 1) was observed in the averaged knowledge structures of both conditions. With the exception of the aggregate Day 1 knowledge structure for ST females, gaining/losing points (i.e., “Points” in Figure 2) also tended to serve as a bridge linking these two concept domains together. This was suggestive that, overall, female learners in all conditions appeared to associate performing well in the task with both accurate decision-making and attending to operational/strategic tactics.

Second, Table 4 shows that gaining/losing points was always rated among the most centralized task concepts in the aggregate knowledge structures of female learners in the control condition across all days. Additionally, the Points concept emerged as the singularly most interconnected concept on Days 2 and 3 for this group. These two characteristics (high centrality and high connectedness) account for the distinctive hub-like pattern of associations around the Points concept observable in the aggregate knowledge structures of female learners in the control condition seen in Figure 2. In contrast, this organizational structure did not manifest in the aggregate

knowledge structure for females in the ST condition. Gaining/losing points was the most interconnected concept on Day 2 for ST females; however, four other concepts shared only one fewer link on this day and produced the "clumpier" aggregate knowledge structure observed at this time point. Further, three other concepts shared the same number of links as the Points concept by Day 3. One possible interpretation for these structural discrepancies is that they reflect differences in learners' functional interpretation of goals in the task environment (Medin et al., 2006). The nature of the simulated task environment was such that gaining/losing points was the primary performance objective for participants, and virtually all decisions and activities contributed directly to achieving this goal. Consequently, configuration of the aggregate knowledge structures for control condition females—in which Points emerged as a centralized and highly interconnected hub—is consistent with associative learning that was strongly goal-oriented. In contrast, the knowledge structure for ST females tended to possess multiple concepts with a high degree of connectedness and appeared less representative of goal-driven learning.

A final observation concerning knowledge structure differences involves the pattern of interrelations among decision-making concepts and gaining/losing points. One hallmark of domain expertise is development of knowledge representations that allow individuals to encode, retrieve, and apply learned knowledge efficiently and accurately (Chase & Simon, 1973). In the present task, a critical component of the knowledge acquisition process involved learning to apply the rules of engagement in order to process contacts and earn points. This procedure required making three distinct classification decisions (Type, Class, Intent) for a contact and then matching that contact's unique classification profile (e.g., Air/Military/Hostile; Surface/Civilian/Peaceful, etc.) to a particular final engagement decision (Clear, Warn, or Mark). One method for learning this procedure would be to memorize all 12 unique classification profiles and their associated final engagement decisions (i.e., "all Air/Civilian/Peaceful contacts should be Warned"). However, a more efficient heuristic approach to organizing this information would be to learn which engagement decisions are most likely given a particular classification (i.e., "Peaceful contacts tend to be Cleared"). Most notably, this knowledge reduces the complexity of the decision-making task without greatly reducing accuracy.[5] Thus, a knowledge structure in which each final engagement decision is associated with its most diagnostic classification (Clear ↔ Peaceful; Warn ↔ Civilian; Mark ↔ Hostile, see Supplemental Materials) would be suggestive of learning efficient and effective heuristics for carrying out performance activities within the task domain.

To facilitate comparison of this type of organizational structure, the links corresponding to both heuristic and performance-oriented associations are highlighted with bolded lines in Figure 2. Although the aggregate knowledge structures of female learners in both conditions contained a similar number of heuristic links at each day, the

structural configuration of these associative relationships was notably different. By Day 2 and continuing to Day 3, the aggregate knowledge structure for female learners in the control condition was configured in a manner consistent with *both* a goal-oriented and heuristic organization. That is, gaining/losing points was directly associated with all three final engagement concepts (Points ↔ Clear, Warn, and Mark), and each final engagement concept was directly associated with its most probabilistic classification outcome (Clear ↔ Peaceful; Warn ↔ Civilian; Mark ↔ Hostile). In contrast, the knowledge structure of female learners in the ST condition never exhibited this pattern.

Taken together, the knowledge structure data suggests that ST during training influenced how female participants inferred associative relationships among important task concepts. Of greatest significance, ST appeared to impede female learners' ability to draw critical inferences among task-relevant concepts that were both goal-oriented and heuristically efficient. This pattern lends support to Hypothesis 2.

Metacognitive Activity

The bottom half of Table 3 presents coefficient estimates from the regression model examining metacognitive activity assuming a normal likelihood distribution over the observed data. Female participants in the ST condition reported engaging in slightly less reflection on their learning than control condition females at Day 1 (β_{01} = -.09, [-.27, .09], 84% of posterior distribution below zero) and this difference tended to increase over time (Figure 3). Specifically, a small but consistent increase in self-reported metacognitive activity was reported by females in the control condition each day (β_{10} = .06, [-.02, .14], 94% of posterior distribution above zero); in contrast, self-reported metacognitive activity tended to decrease by a similarly small margin for ST females over time (β_{11} = -.11, [-.23, .00], 97% of posterior distribution below zero). The pattern of results indicates that female learners in the ST condition reported directing less effort to self-reflection about their learning compared to control condition female learners during training, thus supporting Hypothesis 3.

Learning Engagement

Passive learning activities. To evaluate how female learners in each condition allocated their self-directed study time during the practice trials, the proportion of total time spent viewing the manual sections related to processing contacts and task strategies was computed (time spent / 120 seconds). As is common with many measures of elapsed time, the observed data for both study time variables were not normally distributed; as such, beta and exponential likelihood distributions were used to model the study time data for processing contacts and task strategies, respectively. Visual examination of the observed time spent reviewing the task strategy portion of the

manual also strongly suggested the presence of nonlinear change in this variable over time. As such, a quadratic time term was included in the regression model for this dependent measure.

The top and middle portions of Table 5 present the coefficient estimates (reported in log-odds) for the analyses examining time spent learning to process contacts and task strategies, respectively. Figures 4a and 4b provide a visual depiction of these relationships over time. Female learners in both conditions spent less time studying how to process contacts each trial; however, the proportion of time females in the ST condition spent learning to process contacts (β_{11} = -.07, [-.10, -.04], 100% of posterior distribution below zero) decreased at a faster rate than control condition females (β_{10} = -.03, [-.06, -.01], 100% of posterior distribution below zero). With respect to learning task strategies, both control and ST female learners tended to spend similar proportions of time studying the task strategy section of the manual during the first day's practice trials. However, female learners in the ST condition spent increasingly less time studying this section during the remaining trials until the final day when learners in both groups spent little time reviewing this material (β_{11} = -.14, [-.28, .00], 98% of posterior distribution below zero; β_{21} = .03, [.01, .04], 100% of posterior distribution above zero). Overall, this pattern of results supports Hypothesis 4 and indicates that females in the ST condition tended to spend increasingly less of their discretionary study time learning information critical to task performance compared to control condition females.

Active learning activities. The bottom portion of Table 5 presents the coefficient estimates for the regression model examining the total number of contacts processed during each practice trial assuming a normal likelihood distribution for the observed data. Female participants in the ST condition tended to engage nearly one more contact than control condition females (β_{01} = .83, [.11, 1.54], 99% of posterior distribution below zero) during the first practice trial. The number of contacts processed each trial tended to increase at a roughly similar rate for females in both the control (β_{10} = .30, [.25, .34], 100% of posterior distribution above zero) and ST (β_{11} = -.03, [-.10, .03], 83% of posterior distribution below zero) conditions, with both groups tending to process one more contact every 3-4 trials. Overall, these results do not lend strong support to Hypothesis 5 and suggest that ST had relatively little impact on the amount of effort female learners directed towards engaging in active practice.

Task Performance

Table 6 summarizes the regression estimates for the total number of points earned as well as the total number of contacts processed by participants during the end of day performance trials. Normal likelihood distributions were specified for both measures. Female learners in both the control and ST conditions earned nearly the same number of points on the first performance trial (β_{01} = 57.8, [-216, 326], 66% of posterior distribution above

zero). However, the rate of improvement on subsequent trials for control condition females (β_{10} = 764, [598, 930], 100% of posterior distribution above zero) was nearly 1.5 times greater than that of females in the ST condition (β_{11} = -246, [-487, -14.9], 98% of posterior distribution below zero; Figure 5a). Notably, this performance discrepancy was observed despite female participants in the ST condition tending to process one more contact each performance trial than females in the control condition (bottom of Table 6, Figure 5b). Taken together, these results indicate that female participants who experienced ST exerted a significant degree of effort to complete the more complex and demanding performance trials, though they failed to reach levels of task performance comparable to females in the control condition. This pattern of findings supports Hypothesis 6.

Additional Analyses

The results for Hypothesis 6 indicated that female participants in the ST condition tended to show poorer performance improvements over time than participants in the control condition, the rationale being that observed differences in performance trajectories should be attributable to differences in knowledge acquisition caused by ST. To formally test this mediation, additional analyses were conducted to evaluate the indirect effect of ST on performance through knowledge acquisition and training engagement. Following the recommendations of Yuan and MacKinnon (2009), a series of Bayesian regression models were used to estimate the posterior distributions for the relationships between ST, the measures of learning/engagement, and performance. Figure 6 provides a summary of the overall mediation model. To capitalize on the longitudinal nature of the data as well as maintain consistency with the previous hypotheses which examined the influence of ST on knowledge acquisition/engagement over time (rather than between-group cross-sectional differences), the mediators and dependent variable for these analyses were individual's estimated rates of change in learning/engagement and performance, respectively, across the three-day training session. These variables were obtained for each individual using the following regression model:

$$\text{Level 1: } V_{it} = \pi_{0i} + \pi_{1i}(\text{Time}_{it}) + e_{it} \quad (2)$$

$$\text{Level 2: } \pi_{0i} = \beta_{00} + e_i$$

$$\pi_{1i} = \beta_{10} + e_i,$$

in which π_{1i} = the estimated change in variable V for person *i*. Equation 2 was thus evaluated for each learning/engagement mediator and the dependent performance variable, and the subsequent π_{1i}'s used as the mediating and outcome variables in the mediation analyses. The total indirect effect of ST on performance improvement through learning/engagement was assessed by summing together the specific indirect effects for each mediating variable (e.g., specific indirect effect = $\pi_{\text{ST}\rightarrow\text{Mediator}} * \pi_{\text{Mediator}\rightarrow\text{Outcome}}$). Because each regression coefficient

from the mediation analysis is a posterior distribution, a 95% credibility interval for the final total indirect effect can easily be computed to provide the range of most credible values for the total indirect effect.

Figure 6 summarizes the estimated regression weights for each of the paths specified in the mediation analyses. Overall, changes in the knowledge acquisition and engagement measures over time did mediate the relationship between ST and performance improvement (total indirect effect = -73.9, [-159, 12.8], 96% below zero). However, examination of the specific indirect effects showed that the primary mediating mechanisms were the rate of declarative knowledge acquisition (indirect effect = -73.7, [-143, -10.9], 99% below zero) and rate of change in metacognitive activity (indirect effect = -25.7, [-64.6, 5.86], 96% below), such that ST led to poorer rates of change in both of these variables which led to poorer performance trajectories. Trajectories in the learning engagement variables (amount of time spent studying the task manual and task strategies, number of contacts engaged during practice) were not as consistently related to performance development over time, suggesting that improvement in task performance was not generally influenced by differences in the rate of effort/engagement given by participants. In sum, the additional analyses support the conclusion that acquiring knowledge was important to performance improvement, and that the acquisition of task-relevant knowledge outcomes was impeded by the presence of ST.

COMPUTER SIMULATION OF STEREOTYPE THREAT EFFECTS DURING TRAINING ON ORGANIZATIONAL EFFECTIVENESS

The results from the empirical study reveal that ST effects can have a demonstrable negative impact on the learning outcomes of targeted groups (Figures 1-4) and that those differences can manifest in deficient transfer of training to task performance (Figures 5 and 6). This pattern was consistently observed across multiple measures of knowledge acquisition and learning engagement, though the effects appeared to be generally small-to-moderate in size—a finding similar to much of the research examining ST (Nguyen & Ryan, 2008). So what should be concluded about the practical significance of these findings? As noted previously, simple regression coefficients and statistical indicators of effect size do not adequately convey the practical consequences of an observed effect. Instead, evaluating these impacts requires considering the broader organizational context in which the effects occur (Eagly, 1995). Given the methodological challenges with conducting comprehensive longitudinal research of a complete organizational system, computational modeling techniques that can simulate how effects permeate through a system are a potent tool for contextualizing practical consequences (Ilgen & Hulin, 2000). A computational model is “a precise formulation of the processes through which the values of variables change over time based on theoretical reasoning” (Harrison, Lin, Carroll, & Carly, 2007, p. 1232). The formalism of computational models involves

explicating a set of algorithms (e.g., if X, then Y) and/or mathematical equations that reflect a justifiable and plausible account of events or interactions within a system (Law & Kelton, 1991; Miller & Page 2007). A significant advantage of computational models is that they can be used to conduct computer simulations (i.e., "virtual experiments") in which key parameters or relationships can be manipulated to explore how dynamic events unfold and emerge over time (Grand et al., in press; Kozlowski et al., 2013; Vancouver & Weinhardt, 2012). In the present paper, a simulation is developed to complement the experimental data by examining the extent to which small-to-moderate ST effects experienced during training could potentially impact an organization's effectiveness.

Description of Computational Model

Theories of human resource management note that a key purpose of learning/training is to facilitate the acquisition and improvement of employee KSAs that can be put towards meeting organizational demands (Cascio, 1995; Delaney & Huselid, 1996; Huselid, 1995; Kraiger et al., 1993). Factors that impede the development of these competencies should subsequently influence organizational effectiveness by preventing a firm from realizing its maximum human capital potential. The primary outcome variables of the computational model were thus conceptualized as employee and organization *performance potential*. Employee performance potential reflected an individual-level variable representing the "measure of a worker on the full range of knowledge, skills, abilities, and personality or other characteristics that influence the level at which that individual will perform on the job" (Scullen, Bergey, & Aiman-Smith, 2005, p.7). This construct is not the same as an individual's actual job performance/productivity, which may be influenced by numerous factors across multiple levels of analysis (e.g., motivation, team processes, leadership, economic conditions, etc., Campbell, McCloy, Oppler, & Sager 1993). Instead, performance potential represents an overall set of competencies that an employee has available to use when performing job duties. Importantly, aspects of an employee's performance potential (i.e., KSAs) are capable of changing over time as a result of job-relevant learning/training (Kraiger et al., 1993). For purposes of the present model, an organization's performance potential (i.e., human capital) was represented as the average performance potential among its employees. This treatment is synonymous with compositional models of a firm's human capital resources and is a common operationalization in the strategic human resource management literature (Ployhart, 2006; Ployhart & Moliterno, 2011).[6]

Table 7 outlines the pseudocode (i.e., the algorithmic steps/logical flow) of the final computational model and was adapted from a previous simulation of organizational capital by Scullen et al (2005). A detailed description of the steps in Table 7 as well as the full model specifications is provided in Appendix B, with a general summary

provided here. The overall goal of the simulation was to examine the extent to which the adverse effects of ST on knowledge acquisition similar to those observed in the previously discussed experimental data could negatively impact organizational-level outcomes. As such, modeling the actual process of learning was not of primary interest; instead, the model assumed that all individuals in an organization develop and learn to varying degrees and that impediments to this process can influence the development of an organization's performance potential. The primary sequence of events in the simulation involved the growth of employee and organizational performance potential due to learning/training (Steps 5-6 in Table 7) followed by the departure/replacement of employees due to voluntary turnover (Step 7) over time. Including turnover in the model was critical given that the cycle of employees departing and being replaced is an important and regular source of turbulence in an organization's human capital resources (Ployhart & Moliterno, 2011). When employees leave or retire from a company, they "remove" a portion of the organization's performance potential that must be recouped through new hires and/or developing existing employees (i.e., learning/training). Consequently, obstacles which impair these recuperative mechanisms—such as ST experienced during learning/training—should make it difficult for organizations to develop and sustain high levels of organizational capital and effectiveness. In the present model, voluntary turnover was modeled as a Poisson process in which some number of randomly selected employees left each organization every time period at a rate controlled by a free parameter to the simulation.

The amount by which simulated employees' performance potential due to learning/training changed over time point was constructed to follow the well-documented asymptotic relationship between learning and task performance in which the returns of continued learning for performance improvement become increasingly less impactful over time (e.g., Anderson, 2002). The specific nature of this relationship follows a decreasing power law:

$$y_{it} = a_i t^{b_i} \quad (3)$$

Described in terms of the present simulation, y_{it} represents the change in performance potential for individual *i* at time point *t*, a_i reflects the initial change in performance potential for individual *i* following the first time period (i.e., intercept at time *t* = 1), and b_i equals the rate of change in performance potential as a result of learning/training for individual *i*. The *b* parameter is always a negative value indicating that the rate of improvement in performance potential decreases over time. The result of this parameterization is that performance potential does not grow linearly forever, but will eventually reach an asymptote corresponding to an individual's maximum performance potential. Every agent in the model was thus given unique *a* and *b* parameters that defined that agent's "learning curve" and cumulative performance potential trajectory over time. To model the effect of ST experienced during learning/training,

the rate at which the performance potential of simulated employees' experiencing ST changed over time was computed using the following equation:

$$b_{ST} = \theta b_{NoST}, \quad (4)$$

where b_{ST} is the mean of the sampling distribution used to generate *b* parameters for agents experiencing ST, b_{NoST} is the mean of the sampling distribution used to generate *b* parameters for agents not experiencing ST, and θ is a free parameter on the interval [0, ∞] reflecting the relative difference in the rates of change in performance potential between agents not experiencing ST compared to those experiencing ST. Note that when θ = 1 in Equation 4, $b_{ST} = b_{NoST}$ and there will be no expected differences in the rate of performance potential change between agents experiencing versus not experiencing ST. When θ > 1 though, agents experiencing ST will have a larger negative *b* value and thus a poorer performance potential trajectory.

In sum, the primary process dynamics represented in the computational model simulate employees entering into an organization and developing their performance potential over time as a function of learning. Similar to the previous experimental data, this developmental trajectory was impeded by the presence of ST for some employees by a small-to-moderate degree. Lastly, a proportion of employees left as a result of voluntary turnover at every time point. This process reflects a simplified yet plausible account of the emergence of organizational capital and thus provides greater context for examining the potential "practical" consequences that even small perturbations experienced at the individual-level could exert within an organizational system.

Simulation Design and Analysis Plan

Two free parameters were manipulated in the simulation: voluntary turnover rate (0%, 5%, 10%, 15%, 20%) and the relative impact of ST on the accumulation of performance potential over time (θ = 1, 1.1, 1.25, 1.5). The values for voluntary turnover rate were chosen to represent realistic levels of annual turnover that might be experienced within an organization, while the values for θ were selected to reflect relatively small-to-moderate differences in performance potential improvement between conditions of ST. Note that θ = 1.5 roughly corresponds to the difference in task performance trajectories observed between ST and control participants in the empirical data reported in Table 6. Both of these factors were fully crossed, resulting in 20 simulation conditions. A total of 500 organizations each comprising 100 employees were simulated over 30 time periods within each condition. One half of the organizations in each condition (*k* = 250) contained ST effects in its learning/training practices. The computational model and computer simulation were programmed and run in R (R Core Team, 2015).[7]

Given the large number of simulated observations and the fact that the two free parameters will account for nearly all of the variance in the dependent outcomes across conditions, analyzing between-condition differences using inferential statistics is not informative. Instead, the analysis plan for the simulations focuses on the patterns of organizational performance potential observed and the role that ST experienced during learning/training played in shaping these patterns. The following research questions thus guided the interpretation of the simulated data:

RQ1: Does ST experienced during learning/training influence the development of an organization's performance potential?

RQ2: Does the impact of a small-to-moderate ST effect experienced during learning/training translate into "practically consequential" effects for organizations?

Simulation Findings

Figure 7 plots the average organizational performance potential of simulated organizations from 7 of the 20 simulated conditions. The dotted lines were selected to provide a baseline comparison and show the development of organizational performance potential in simulated organizations without ST during learning/training with turnover rates from 5% to 20%. In contrast, the solid lines show organizations in which turnover was held constant at 10% and the effect of ST on employees' performance potential improvement varied from $\theta = 1.1$ to 1.5. Overall, the pattern of results shows that the performance potential of all organizations tends to increase quickly within the first few time periods but eventually reaches a stable asymptote. The asymptotic trajectory is primarily a function of voluntary turnover and the countervailing impact that experienced employees leaving (thus removing their performance potential from the organization) and being replaced by less experienced employees (thus bringing new, but initially less, performance potential to the organization) have on an organization's human capital resources. As exemplified by the dotted lines shown in Figure 7, the size of the voluntary turnover rate had a notable effect on the overall performance potential achieved by simulated organizations. When turnover was low, organizations were able to benefit from employees who had developed expertise for a longer period of time, thus enabling the firm to reach and sustain higher levels of human resource capital. As turnover increased though, employees with high performance potential were less likely to stay in the organization for long periods of time and thus the organization's overall pool of employee potential tended to be lower. In sum, this pattern of findings reflects a more generalizable conclusion that organizations which are more successful at retaining their workforce will see a greater return on investment on their human resource practices (i.e., learning/training, e.g., Huselid, 1995; Ployhart & Moliterno, 2011).

This same conclusion and pattern of results also held true for organizations that experienced ST during learning/training. However, ST effects also resulted in a demonstrably lower ceiling for organizational performance potential. The most direct evidence of this effect in Figure 7 can be observed by comparing the single dotted line showing the results from organizations with a 10% voluntary turnover rate/ST θ = 1 against the four solid lines showing the results from organizations with a 10% voluntary/ST θ > 1. These comparisons demonstrate that as the impact of ST during learning/training increases for a given turnover rate, the accumulation of organizational performance potential decreases substantially. An even more striking interpretation of this impact can be gleaned by considering the effect of ST in terms of comparable turnover rates. For example, the comparisons shown in Figure 7 reveal that simulated organizations with turnover = 15%/ST θ = 1 achieved roughly the same performance potential as organizations with turnover = 10%/ST θ = 1.1. In other words, the presence of a relatively small effect of ST during learning/training in the simulation had the same relative impact on the development of an organization's human capital as an organization whose turnover rate was 5% higher. Though not shown in Figure 7, additional simulations revealed that the performance potential for an organization with a 0% turnover rate and ST θ = 1.5—the size of the ST effect most comparable to that observed in the experimental data (see Table 6)—was roughly equivalent to that found for organizations with no ST and a turnover rate = 30%. In other words, an organization in which the impact of ST on learning was similar to that found in the present study that never lost its most experienced and competent employees would *still* fail to achieve greater levels of aggregate employee performance potential than an organization that turned over nearly a third of its workforce every year. In sum, these comparisons suggest that even small-to-moderate ST effects at the individual-level appear capable of exerting a significant impact at the organizational-level. If one factors in the challenge and cost associated with needing to retain anywhere from 5-30% of an organization's workforce to even come close to counteracting the adverse effects of ST, the simulation results demonstrate that ST during learning/training could have substantial practical consequences.

To provide further context of the emergent impact that small-to-moderate ST effects on individual-level performance improvement rates exert at the organizational-level, Figure 8 plots the change in observed effect size (Cohen's *d*) in organizational performance potential differences calculated at each time point for simulated organizations with ST versus those without ST when holding turnover rate constant at 10%. Similarly, Figure 9 plots changes in the observed effect size for improvements in individual-level performance potential (i.e., y_{it} from Equation 3) among simulated employee's experiencing versus not experiencing ST at these time points for these organizations.[8] A comparison of these figures clearly show that although differences in the *rates* of simulated agents'

performance improvement at each time point were generally small-to-moderate, the *cumulative* effect of these discrepancies rapidly led to substantial differences at the organizational-level.

Overall, the simulation findings supplement the results from the empirical study by providing a more contextualized perspective on the "practical" consequences of ST effects experienced during learning/training. The results showed that the experience of ST effects during learning/training could significantly impede the development of an organization's performance potential (RQ1; Figure 7). Furthermore, they revealed that even if the effect of ST on the rate of performance improvement attributable to learning was relatively small, those effects could manifest as large disparities at the organizational-level with potential for significant practical consequences (RQ2; Figure 8).

DISCUSSION

Using a longitudinal experimental design, the results of this study revealed that the presence of ST across a three-day self-directed training session impaired females' acquisition of both basic and more advanced knowledge outcomes. Females faced with a negative stereotype about their gender achieved lower levels of declarative knowledge, developed less efficiently organized knowledge structures, and reported engaging in less metacognitive activity about their learning relative to females who did not face these conditions by the end of training. Female learners experiencing ST also tended to spend less time studying task-relevant material and were less effective at performing the trained task. These findings provide converging evidence that ST can significantly impair knowledge acquisition and performance outcomes for learners facing a negative group stereotype. Building from these conclusions, results from a computational model and computer simulation were presented to demonstrate the "practical" consequences that ST effects experienced during learning/training could hold for an organization. The model instantiated a set of plausible assumptions about the role of learning/training practices in an organization's human capital development. Findings from the model simulations suggest that even small-to-moderate ST effects experienced during learning/training could significantly impair the development of an organization's workforce and substantially reduce the effectiveness and efficiency of its human resource practices.

Theoretical Implications

Rydell, Rydell, and Boucher (2010) posit that ST theory has seen little attention in the domain of knowledge acquisition because "learning is difficult to distinguish from performance" (p. 885). This suggests that an important consideration for continued research in this domain is the need to accurately distinguish performance and learning outcomes. In the present study, this was accomplished through the use of Kraiger et al.'s (1993) taxonomy of learning outcomes to operationalize knowledge acquisition from task performance. Although other similar taxonomies

exist (e.g., Kirkpatrick, 1976; Gagne, 1984), the Kraiger et al. (1993) framework is particularly useful as it describes multiple types of learning outcomes that span both basic and advanced indicators of knowledge acquisition which may be targeted by organizational training interventions. The taxonomy also describes a variety of non-cognitive training outcomes that could prove fruitful for research on ST effects during training. For example, future work could expand the consideration of learning outcomes to investigate ST effects on the development of skill- (e.g., does ST impair behavioral adaptation or automatization of task behaviors?) and affective-based (e.g., does ST promote greater resistance towards error-based training or result in smaller improvements in self-efficacy following training?) learning outcomes. Careful operationalization of learning outcomes marks a critical need for future research on ST effects experienced during learning/training.

An added benefit of studying ST effects across a broader range of learning outcomes is that it provides greater understanding of the impact that negative stereotypes exert on knowledge acquisition. Research on ST effects during learning in previous research have primarily examined declarative knowledge using measures identical to those used to demonstrate ST effects on performance outcomes (multiple-choice, true-false, or free recall tests). Such operationalizations are problematic because it is challenging—if not untenable—to disentangle the degree to which ST influences encoding, storage, and synthesis mechanisms (characteristics most often associated with learning and training, Kraiger et al., 1993) relative to information retrieval and manipulation (characteristics associated with task performance, Schmader et al., 2008). The use of multiple metrics of learning and learning engagement in the present study helped to address these methodological concerns. The examination of knowledge structures and metacognitive activity, which do not require recall of declarative knowledge, lessens concerns of "double-dipping" (i.e., ST impacting both encoding and recall of declarative knowledge) in dependent measures of learning that may muddy interpretations of ST effects on knowledge acquisition. Furthermore, the acquisition of associative knowledge representing comprehension of task procedures/contingencies (i.e., knowledge structures) as well as "learning how to learn" (i.e., metacognitive activity) are both advanced cognitive learning outcomes that tend to hold more direct implications for task performance. Nearly all considerations of the relationship between knowledge acquisition and performance recognize that basic declarative knowledge is an important factor (e.g., Campell et al., 1993; Anderson et al., 2004). However, the translation of knowledge about "what" (declarative) into knowledge about "how, when, and why" (procedural/associative, strategic) is more critical to developing task expertise—the typical goal of organizational training (e.g., Day, Winfred, & Gettman, 2001; Kozlowski et al., 2001; Medin et al., 2006; Schuelke et al., 2009).

An additional unique aspect of the empirical study was the evaluation of changes in knowledge structures over time. Longitudinal measures that track the development and maturity of associative knowledge networks provide a powerful but rarely used technique for examining learning outcomes (Ifenthaler, Masduki, & Seel, 2011). Many studies which assess learners' knowledge structures do so for the purpose of evaluating similarity between novice and expert mental models (e.g., Chi et al., 1988; Day et al., 2001) and/or to correlate descriptive network metrics with a related performance outcome at a single time point (e.g., Kozlowski et al., 2001; Schuelke et al., 2009). Such studies assume that the pattern of relations among knowledge concepts conveys meaning about the way in which learners make sense of a given domain space, but deducing that meaning is not of central importance or is too speculative. In contrast, the theoretical rationale proposed in the present study suggested that impediments to working memory caused by ST would make it more difficult for affected learners to integrate task-critical information into efficient/effective task heuristics. It was possible to evaluate this prediction by identifying a heuristic effective for making performance-relevant decisions and examining whether the pattern of relations among knowledge concepts directly relevant to making these decisions emerged over time. Interpreting the meaning of associative knowledge structures and their longitudinal development were thus critical to evaluating knowledge acquisition during training and potential obstacles to related outcomes.

A useful direction for future research in this area would be efforts to further investigate the unique patterns of learning engagement observed across conditions of ST and whether such patterns could be used to identify threatened individuals in a particular learning environment. It was noteworthy that although female learners in the ST condition tended to disengage more quickly and completely from studying the task manual relative to female learners in the control condition (passive learning), there were no such differences in the level of engagement observed during practice opportunities in the task environment (active learning). In fact, female learners in the ST condition tended to consistently process more contacts than their control condition counterparts each practice trial, a finding also replicated during the end of day performance trials. However, and consistent with the poorer developmental trajectories in declarative knowledge, knowledge structures, and metacognitive activity, this additional effort did not translate into improvements in the final performance outcome for females in the ST condition (Figure 6). This pattern of results suggests that learners influenced by ST may be attempting to "work harder" at the expense of learning to "work smarter." There are a variety of possible explanations that may account for this occurrence and which suggest possible methods for intervention. For example, it may be that the approach-oriented, error-promoting framing of training programs (i.e., focus on learning as much as possible, mistakes are okay) conflicts with the avoidance

orientation stimulated by ST (i.e., focus on not appearing incompetent, Grimm et al., 2009). Efforts to frame training as non-evaluative and include mechanisms that help learners regulate their cognitive and affective reactions during learning activities (e.g., Bell & Kozlowski, 2008) may thus be particularly critical in training environments where persistent group stereotypes are likely to be present.

Simon's (1956) theory of bounded rationality also posits that the observed differences between "working harder versus smarter" could be attributable to differences in how learners experiencing ST perceive the learning environment. Steele and Aronson (1995) postulated that a key motivation for threatened individuals is to avoid confirming the validity of negative group stereotypes through their actions. In the language of bounded rationality, individuals experiencing ST would consider this need an additional demand of the learning environment that must be addressed by avoiding deficiencies in the amount of knowledge possessed. This could encourage suboptimal learning strategies (i.e., rote memorization) rather than integrative approaches that facilitate more generalizable knowledge acquisition (i.e., experimentation, trial and error, Chase & Simon, 1973; Lipschitz, Levy, & Orchen, 2006). It may thus be possible to utilize information about specific patterns of learning engagement (e.g., less effort during passive learning activities coupled with high levels of effort during more active learning activities) to identify learners experiencing ST during training so that appropriate interventions could be delivered.

Lastly, and on a more general note, the use of Bayesian analytic approaches and computational modeling reflect a unique methodological contribution of this paper. Although neither technique has seen prevalent use in the organizational sciences, they were highly flexible and useful tools for examining and contextualizing the key findings from this study. With respect to Bayesian analyses, much has already been described of the shortcomings of frequentist and null-hypothesis significance testing (NHST) approaches to statistical inference (e.g., Kruschke et al., 2012; Kruschke, 2013; Zyphur & Oswald, 2015). To be sure, Bayesian statistics face their own challenges; nevertheless, they offer a number of advantages over frequentist approaches that were exemplified in the present study. Most notably, Bayesian models offer a direct means for interpreting the likelihood/strength of evidence for *any and all* potential values of a parameter. Although it is common to draw attention to the most probable value for a parameter, it is straightforward to determine the extent to which any specific value (such as zero) or even some range of values are likely to be believable estimates. This provides significantly greater potential for drawing inferences about an effect than whether a single point estimate or range of values are likely to be different from zero—the primary inference achievable with frequentist/NHST approaches. In addition to providing a richer means for interpreting results from a single study, reporting data about the posterior distribution of an effect facilitates the

accumulation of evidence through future research as these values can be used as the starting point (i.e., prior distributions) for examining future similar relationships. With respect to computational modeling, there are many potential uses of simulation techniques for generating insights into organizationally relevant phenomena (Harrison et al., 2007; Kozlowski et al., 2013). In the present study, the computational model was used for illustrative purposes in an attempt to provide a more nuanced perspective on the "practical significance" of small ST effects that cannot be adequately captured by simple effect sizes (Eagly, 1996). Furthermore, the model could be easily extended or adapted in a variety of ways to examine the effects of other individual- or organizational-level variables. In sum, there is growing recognition in the psychological sciences that moving beyond NHST and conventional inferential techniques is necessary for improving our understanding of phenomena (Cumming, 2014). The use of Bayesian statistics and computational modeling techniques are both powerful and flexible tools for accomplishing these goals and ultimately enhancing the accuracy, interpretability, and potency of empirical findings related to ST and beyond.

Practical Implications

Since its inception, ST theory has generated the most interest—and heated debate—in the context of high stakes educational testing and personnel assessment (e.g., Cullen et al., 2004; Cullen et al., 2006; Sackett et al., 2004; Sackett & Ryan, 2012; Steele & Davies, 2003; Stricker & Ward, 2004, 2008). Critiques of the practical implications for ST in these and other applied settings can be summarized into two related issues. The first concerns the likelihood that ST effects can manifest and/or be experienced outside of controlled laboratory conditions. While systematically conducting research specific to ST in real organizations would be desirable, it is important to note ST is a *specific instance* of the more general family of expectancy effects—for which ample support has been observed in field settings. For example, evidence of the Pygmalion (in which people live up to a positive expectation held by an authority figure, e.g., Eden, 1990; Eden & Shani, 1982) and Golem effects (in which people "sink down" to a negative expectation held by an authority figure, e.g., Babad, Inbar, & Rosenthal, 1982) on performance outcomes have been found in applied settings. Therefore it seems doubtful that the *psychological experiences* described in ST theory should operate differently in a natural versus laboratory setting.

A question of greater significance is whether conditions that may elicit ST in organizations may be likely to exist. To this end, Walton et al. (2015) have identified a number of cues, practices, and signals within organizations which empirical research suggests reinforce the adverse effects of ST. Two of these cues—messages implying fixed-ability beliefs and the receipt of critical feedback—are often highly salient in learning/training environments (e.g., Bell & Kozlowski, 2008; Keith & Frese, 2005; Kluger & DeNisi, 1996). Importantly, such cues need not explicitly state or

call attention to the presence or salience of a negative group stereotype; instead, these cues *provide the means* for individuals already concerned with the possibility of fulfilling a negative group stereotype to self-validate those beliefs. These self-fulfilling and self-defeating expectancies decrease the likelihood that threatened individuals will adopt the approach/mastery orientation critical to successful learning (Forbes et al., 2008; Mangels et al., 2012; Jamieson & Harkins, 2007). Furthermore, the norms, culture, and policies surrounding organizational training environments tend to be far less standardized, monitored, and regulated than high-stakes assessment contexts. For example, a large proportion of training and development in organizations occurs informally through "on-the-job" sources that typically fall outside the purview of organizational control (Loewenstein & Speltzer, 2000). Consequently, to the extent that such policies might be effective at mitigating conditions conducive to ST and suppressing its negative consequences (Sackett et al., 2001), these factors are less likely to be present in learning/training contexts. In sum, whether intentional or innocuous, there are both plausible reasons and empirical evidence to suspect that the psychological mechanisms and cues underlying ST could manifest in organizational learning/training environments (Roberson & Kulik, 2007; Walton et al., 2015). The present research highlights the need for organizations to direct as rigorous attention to evaluating and regulating learning/training practices as has been devoted to securing amenable and non-discriminatory selection/assessment practices.

A second critique frequently raised against the practical implications of ST is that, even if such effects were to exist in organizations, their impact is likely to be so small as to be inconsequential. However, concluding about the practical significance of potentially discriminatory effects on the basis of simple or even meta-analytic effect sizes is insufficient. Organizational scholars have long acknowledged that individual and organizational effectiveness should not solely be evaluated with respect to *what* outcomes are achieved, but also *how* those outcomes are achieved and unfold within the broader organization (Campbell et al., 1993; Ployhart, 2012; Sonnentag & Frese, 2012). The computational model and computer simulations presented in this study are highly suggestive that even small effects of ST can ripple through an organizational system to generate significant practical concerns for performance. It is important to recognize, however, that the model did not reflect a key observation from the experimental study: *how* individuals facing ST complete task-relevant activities differs from individuals not facing ST. Such process differences may manifest in many ways, such as taking longer to do the same tasks, expending greater cognitive resources to achieve a given performance level, or demonstrating poorer developmental trajectories. These consequences could potentially contribute to a host of undesirable outcomes, including burnout, stagnated growth, poorer job satisfaction, and slower opportunities for advancement/promotion in areas with prevalent group stereotypes and rapidly-paced

learning environments (i.e., STEM occupations). Both individual and organizational effectiveness emerges from complex interactions among individuals' affective, behavioral, and cognitive activities as well as situational/environmental demands. Together, the empirical and simulation results highlight a greater need to adopt a dynamic perspective when considering the practical implications of ST for employees and organizations that is informed by evidence beyond single time-point measures and simple between-group effect sizes.

Limitations and Future Directions

It is important to note some limitations of the present study that represent opportunities for future research. First, the structure of the learning environment plays a key role in the acquisition of learning outcomes (Bell & Kozlowski, 2008). In the empirical study, an active learning paradigm based on principles of exploratory learning was implemented. Future research should investigate whether ST effects influence knowledge acquisition similarly in other learning contexts (i.e., lecture/procedural instruction, behavioral modeling, etc.). Second, the trained task was complex, highly proceduralized, and time restricted; consequently, simple heuristics were strongly rewarded as a means for achieving task effectiveness. Additional research is needed to determine whether tasks or jobs that are simpler, more fluid, and/or possess less "concrete" demands (i.e., problem-solving, troubleshooting, etc.) are susceptible to similar learning deficiencies attributable to ST. Third, the use of psychology undergraduate students who may be familiar with ST could have influenced the observed results. However, Johns, Schmader, and Martens (2005) report that knowledge of ST tends to attenuate or even remove the negative effects of threat for at-risk individuals. Thus, the present results may actually be more conservative compared to if completely naïve individuals had participated in the study. Lastly, because female participants in the ST condition may have also been experiencing ST during the performance trials, it was not possible to attribute observed task performance differences solely to variability in learning between the two groups. In fact, the mediation analyses revealed that although rates of knowledge acquisition accounted for a sizable portion of the variation in performance improvement differences observed across the conditions, unexplained sources of variance in task performance not accounted for by differences in learning remained.

Recommendations for Examining the External Validity of Stereotype Threat

The preponderance of laboratory-based studies on ST—such as the present research—has caused many in the applied social sciences to question whether findings from controlled studies on ST generalize to more natural settings (i.e., organizations). While a justifiable concern, critiques of this issue are often ambiguous with respect to their primary concerns (e.g., Does the phenomenological experience of ST occur in a natural setting? Do observed

effect sizes replicate? Does ST hold for some groups/stereotypes but not others? Are conditions favorable to ST frequently encountered?) and thus seldom provide specific direction for how to adequately evaluate the external validity of ST theory and contextualize related findings. To this end, a brief discussion and suggestions for evaluating and improving our understanding of the generalizability of ST theory are presented.

In their widely cited text, Shadish, Cook, and Campbell (2002, p. 83) state that "external validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes." This definition draws attention to two critical considerations relevant to evaluating the generalizability of ST. The first concerns appropriately specifying the causal relationships proposed by ST theory that should be the focus of external validity studies. It is commonly assumed that the relation of interest in ST theory is "presence of stereotype → diminished performance." However, this is a significant oversimplification of the intra-individual processes and situational elements that are central to the claims of ST theory. The realization of detrimental outcomes attributable to a negative group expectancy/stereotype is proposed to arise through a series of psychological processes that subsume working memory capacity (e.g., increased anxiety, heightened monitoring of feedback, etc., Schmader et al., 2008) and that are sustained/reinforced by specific environmental features (e.g., identity contingency cues such as fixed ability belief messages, critical feedback, underrepresentation, etc. Walton et al., 2015). *The presence and interactions among these characteristics are the fundamental causal relationships of ST theory and the most important criteria for establishing its generalizability*. Attempting to infer the presence of ST effects by examining subgroup differences in stereotyped domains using archival performance data (e.g., Cullen et al., 2004; Cullen et al., 2006) or priming group identities without explicitly measuring any characteristics of individuals or the environment in which they operate (e.g., Stricker & Ward, 2004) are not assessments of whether the *phenomenological experience* of ST (i.e., the set of causal relationships constituting ST theory) is generalizable. Indeed, Stricker and Ward (2004)—whose study is often referred to as the best evidence that ST effects do not extend to natural settings—acknowledged this point in their own research: "A clear limitation of these studies was that data were only available about test performance and not about its possible causes (e.g., stereotype threat) or mediators (e.g., anxiety)" (p. 690). An additional point of caution in using such data is exemplified by the pattern of findings observed in the present study. Virtually no differences in any of the measured learning or performance outcomes were observed between females in the ST and control conditions at the end of Day 1; however, clear disparities manifested over time between these groups as they continued to operate in an environment that reinforced the negative stereotype. Such results demonstrate that the "failure" to detect between-group differences in

a criterion measure at a single time point does not mean that members of those groups were experiencing the same psychological processes. In sum, *a first critical goal of research designed to evaluate the external validity of ST should be to consider the core set of causal mechanisms proposed in ST theory rather than only outcome data at a single point in time*.

A second important consideration from Shadish et al.'s (2002) definition is the recognition that external validity involves generalization across samples, settings, treatments, *and* outcomes. Importantly, this does not mean that the only or even most diagnostic evaluation of the external validity of ST is to study the phenomenon in a "real world" population of working adults. Consistent with Shadish et al.'s (2002) definition, the current study actually contributes to the external validity of ST theory by demonstrating that ST effects were observed using a different outcome criteria (i.e., knowledge acquisition vs. cognitive test performance) and treatment design (longitudinal vs. cross-sectional) than has been conventionally examined in past research. Most critiques of ST theory, though, are singly concerned with demonstrating external validity across samples and settings. This is understandable given that such a large proportion of the empirical data on ST has been collected under controlled laboratory conditions with college student populations (e.g., Sackett et al., 2004; Sackett & Ryan, 2012). However, this critique implies that if the psychological mechanisms responsible for ST only operate in controlled laboratory experiments and not in "natural" settings, there should be moderating factors that account for these differences. Generating (and ideally testing) falsifiable predictions of these contingencies would bring much greater precision and clarity to the debate over the generalizability of ST theory, and would thus be a welcome addition to critiques of ST theory. Consequently, *a second critical goal of research designed to evaluate the external validity of ST should be to identify, measure, and evaluate moderating factors that distinguish when and why ST effects are (or are not) likely to be observed across different settings, samples, treatments, and/or outcomes*. In line with these overarching goals, two specific suggestions are provided for advancing future research on the external validity of ST.

Expand and improve the domain of measurement in ST research. A first recommendation concerns broadening the scope of measurements used to examine ST effects in natural settings. More specifically, *research must expand the focus and content of measures to consider alternative outcome criteria, more proximal/diagnostic indicators of the phenomenological experience of ST, and make more concerted efforts to evaluate ST effects longitudinally*. As was noted in the opening of this paper, performance/evaluative criteria are overwhelmingly the most popular dependent variable in studies of ST. However, performance is arguably the most operationally difficult, conceptually complex, multiply-determined, and multi-faceted constructs in the fields of organizational psychology

and behavior (e.g., Campbell et al., 1993; Sonnentag & Frese, 2012). This does not mean that researchers should cease efforts to evaluate the consequences of ST for performance in natural settings; however, it does suggest that research should also devote resources to assessing the effects of ST on other important and organizationally-impactful outcomes that are relevant to employee and organizational effectiveness. For example, the current study demonstrated the effect of ST on knowledge acquisition and learning outcomes in a simulated training environment, while the computational model and simulation results demonstrated how such learning detriments emerge to impact organizational-level human capital. More affective and attitudinal measures such as experiences of stress, burnout, job satisfaction, turnover/turnover intentions, and organizational commitment also represent potential and plausible outcomes that may be subject to ST effects and which hold clear implications for individuals and organizations.

In addition to evaluating more and different outcome criteria, greater effort is needed to assess indicators of the psychological experiences associated with ST. Evaluating the external validity of ST involves evaluating whether its proposed psychological mechanisms operate consistently across samples, settings, treatments, and outcomes—not simply whether between-group performance differences are observable. Measures of felt anxiety/negative affect, behaviors and cognitions related to seeking and interpreting evaluative information, and the regulation of negative thoughts and emotions are all factors posited as core to the phenomenological experience of ST. Collecting data on such variables would thus be highly diagnostic of whether ST may be operating for particular individuals in a given situation (Schmader et al., 2008). Examinations of this sort will also require more frequent use of within-person assessment strategies than are commonly pursued in the existing ST literature. Longitudinal designs that evaluate the experiences of employees potentially facing identity-threatening situations over both longer and shorter periods would be invaluable to interpreting the external validity of ST theory. The use of experience-sampling methodologies that gather data on both diagnostic indicators of ST as well as the corresponding situational elements in which those experiences are faced seems a particularly promising strategy. Unlike existing studies purporting to test the external validity of ST, such designs would provide a more precise and direct comparison of whether the critical interactions among person and situational characteristics observed in controlled studies of ST also emerge when experienced "naturally."

Direct more attention to the situation and less to stereotype priming. A second recommendation for external validity studies of ST concerns the importance of assessing situational characteristics and the misconception that field researchers must prime stereotypes in order to evaluate ST in natural settings. It is true that virtually all laboratory-based ST research has relied on paradigms that manipulate the salience of group stereotypes to create

conditions/environments in which to examine ST effects. However, the reason for this is one of experimental control rather than theoretical necessity. A central proposition of Steele and Aronson's (1995) theory is that ST occurs when individuals find themselves in *situations* where their actions could potentially validate an unfavorable stereotype as self-characteristic. Actively priming the salience or awareness of a stereotype is thus neither a requirement for eliciting ST nor examining its external validity. Despite this fact, the manipulation of stereotype salience has received a disproportionate degree of both empirical (e.g., ST studies examining differences between subtle versus overt stereotype primes on performance outcomes, e.g., Grand et al., 2011; Nguyen & Ryan, 2008) and critical (e.g., Sackett & Ryan, 2012; Stricker & Ward, 2004; Whaley, 1998) attention.

Given that manipulating group stereotypes in natural settings has legal, ethical, and practical limitations and is unnecessary for examining the generalizability of ST effects in more natural settings, *attempts to examine the external validity of ST should direct greater attention towards the situational features capable of eliciting and reinforcing ST*. In other words, the assumption of the ST generalizability researcher should be that a targeted individual in a stereotyped domain is already aware of the stereotypes they face.[9] In exchange, the situational elements of an environment capable of facilitating (or attenuating) the adverse experience of that stereotype should be more closely scrutinized to identify whether ST effects are likely to unfold. In conjunction with the suggestion to measure key mediating/process mechanisms of ST, this strategy would allow researchers to more accurately explore its phenomenological experience as well as categorize the prevalence of ST-provoking features in natural settings. As a point of departure, the framework of identity contingency cues proposed by Walton et al. (2015) offers a theoretically defensible and empirically supported taxonomy for future research to identify and systematically capture situational characteristics known to influence the affect, behavior, cognition, and outcomes of domain stereotyped individuals in ways consistent with ST. As this suggestion places emphasis on individuals' reactions to the environment, it will be prudent to obtain measures of these situational features through multiple sources and operationalizations. Furthermore, consideration must be given towards justifying whether such data should be aggregated across people to obtain a single/collective index of the environment (e.g., a compositional construct) or whether such data should be kept disaggregated and treated as an individual-level perception (e.g., a configural or "frog-pond" construct, Kozlowski & Klein, 2000).

Conclusion

This research demonstrated that the presence of negative domain stereotypes interfered with the acquisition of cognitive learning outcomes and engagement during training activities for threatened individuals. Results from a

computer simulation further revealed that such learning deficiencies can accumulate over time and across organizational levels to generate substantial practical concerns. Investigating the effects of ST on learning/training carries a number of conceptual and practical implications, including a better understanding of how ST experienced during performance differs from or is similar to ST experienced during learning, the need to evaluate training practices/procedures to ensure individuals learn and engage information in stereotyped content domains, and further explication of the boundary conditions, cognitive mechanisms, and consequences of ST beyond performance/evaluative testing contexts. Finally, this research reveals the importance of exploring *how* ST effects manifest over time and in context rather than simply *whether* such effects can be elicited. Investigating the impact of ST effects on both short- and long-term longitudinal outcomes is a critical need for continued research in this area.

REFERENCES

Aguinis, H., & O'Boyle, E. (2014). Star performers in twenty-first century organizations. *Personnel Psychology, 67*, 313-350.

Anderson, J.R. (2002). *Learning and memory* (2nd ed.). Wiley: New York, NY.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036-1060.

Babad, E.Y., Inbar, J., & Rosenthal, R. (1982). Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology, 74*, 459-474.

Baddely, A.D., & Hitch, G. (1974). Working memory. In G.A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). New York: Academic Press.

Beilock, S.L., Gunderson, E.A., Ramirez, G., & Levine, S.C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 1860-1863.

Beilock, S.L., Rydell, R.J., & McConnell, A.R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136*, 250-276.

Bell, B.S., Kanar, A.M., & Kozlowski, S.W.J. (2008). Current issues and future directions in simulation-based training in North America. *The International Journal of Human Resource Management, 19,* 1416-1434.

Bell, B.S., & Kozlowski, S.W.J. (2002). Adaptive guidance: Enhancing self-regulation, knowledge, and performance in technology-based training. *Personnel Psychology, 55*, 267-306.

Bell, B.S., & Kozlowski, S.W.J. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology, 93*, 296-316.

Blume, B.D., Ford, J.K., Baldwin, T.T., & Huang, J.L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*, 1065-1105.

Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1993). A theory of job performance. In N. Schmitt & W. Borman (Eds), *Personnel Selection in Organizations* (pp. 35-70). San Francisco, CA: Jossey-Bass.

Cascio, W.F. (1995). Whither industrial and organizational psychology in a changing world of work? *American Psychologist, 50*, 928-939.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Chi, M.T.H., Glaser, R., & Farr, M.J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Colquitt, J.A., LePine, J.A., & Noe, R.A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 5*, 678-707.

Cullen, M.J., Hardison, C.M., & Sackett, P.R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*, 220-230.

Cullen, M.J., Waters, S.D., & Sackett, P.R. (2006). Testing stereotype threat theory predictions for math-identified and non-math-identified students by gender. *Human Performance, 19,* 421-440.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29.

Day, E., Winfred, A., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill, *Journal of Applied Psychology, 86*, 1022-1033.

Dearholt, D.W., & Schvaneveldt, R.W. (1990). Properties of pathfinder networks. In R.W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 1-30). Norwood, NJ: Ablex Publishing Corp.

Debowski, S., Wood, R.E., & Bandura, A. (2001). Impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic search. *Journal of Applied Psychology, 86*, 1129–1141.

Delaney, J.T., & Huselid, M.A. (1996). The impact of human resource management practices on perceptions of organizational performance. *Academy of Management Journal, 39*, 949-969.

Denwood, M.J. (in press). runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*.

Eagly, A. (1995). The science and politics of comparing women and men. *American Psychologist, 50*, 145-158.

Eden, D. (1990). Pygmalion without interpersonal contrast effects: Whole groups gain from raising manager expectations. *Journal of Applied Psychology, 75*, 394-398.

Eden, D., & Shani, A.B. (1982). Pygmalion goes to boot camp: Expectancy, leadership, and trainee performance. *Journal of Applied Psychology, 67*, 194-199.

Faria, A.J. (1998). Business simulation games: Current usage levels—an update. *Simulation & Gaming, 29,* 295-308.

Faria, A.J., & Nulsen, R. (1996). Business simulation games: Current usage levels a ten year update. *Developments in Business Simulation & Experiential Exercises, 23,* 22-28.

Feldman Barrett, L., Tugade, M.M., & Engle, R.W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin, 130*, 553-573.

Figueroa-Zúñiga, J.I., Arellano-Valle, R.B., & Ferrari, S.L.P. (2012). Mixed beta regression: A Bayesian perspective. *Computational Statistics & Data Analysis, 61*, 137-147.

Forbes, C.E., Schmader, T., & Allen, J.J.B. (2008). The role of devaluing and discounting in performance monitoring: A neurophysiological study of minorities under threat. *Social Cognition and Affective Neuroscience, 3*, 253-261.

Ford, J.K., Smith, E.M., Weissbein, D.A., Gully, S.M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83*, 218-233.

Gagne, R.M. (1984). Learning outcomes and their effects: Useful categories of human performance. *American Psychologist, 4*, 377-385.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall.

Gelman, A., & Hill, J. (2004). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Glaser, R. (1990). The reemergence of learning theory within instructional research. *American Psychologist, 45*, 29-39.

Goldstein, I.L., & Ford, J.K. (2002). *Training in organizations: Needs assessment, development, and evaluation* (4th ed.). Belmont, CA: Wadsworth.

Golubovich, J., Grand, J.A., Ryan, A.M., & Schmitt, N. (2014). An examination of common sensitivity review practices in test development. *International Journal of Selection and Assessment, 22*, 1-11.

Grand, J.A., Braun, M.T., Kuljanin, G., Kozlowski, S.W.J., & Chao, G.T. (in press). The dynamics of team cognition: A process-oriented theory of knowledge emergence in teams. *Journal of Applied Psychology* [Monograph].

Grand, J.A., Golubovich, J., Ryan, A.M., & Schmitt, N. (2013). The detection and influence of problematic item content in ability tests: An examination of sensitivity review practices for personnel selection test development. *Organizational Behavior and Human Decision Processes*, *121*, 158-173.

Grand, J.A., Ryan, A.M., Schmitt, N., & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24*, 1-28.

Grimm, L.R., Markman, A.B., Maddox, W.T., & Baldwin, G.C. (2009). Stereotype threat reinterpreted as a regulatory mismatch. *Journal of Personality and Social Psychology, 96*, 288-304.

Halpern, D.F., Benbow, C.P., Geary, D.C., Gur, R.C., Hyde, J.S., & Gernsbacher, M.A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*, 1-51.

Harkins, S. G. (2006). Mere effort as the mediator of the evaluation–performance relationship. *Journal of Personality and Social Psychology, 91*, 436–455.

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review, 32,* 1229-1245.

Huselid, M.A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal, 3*, 635-672.

Ifenthaler, D., Masduki, I., & Seel, N.M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. *Instructional Science, 39*, 41-61.

Ilgen, D.R., & Hulin, C.L. (2000). *Computational modeling of behavioral processes in organizations: The third scientific discipline*. Washington, DC: American Psychological Association.

Interlink. (2011). *FAQs*. Retrieved July 12, 2011 from http://interlinkinc.net/FAQ.html

Jamieson, J.P., & Harkins, S.G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology, 93*, 544–564.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16*, 175-179.

Kamouri, A.L., Kamouri, J., & Smith, K.H. (1986). Training by exploration: Facilitating the transfer of procedural knowledge through analogical reasoning. *International Journal of Man-Machine Studies, 24*, 171-192.

Kane, M.J., Bleckley, M.K., Conway, A.R.A., & Engle, R.W. (2001). A controlled-attention view of WM capacity. *Journal of Experimental Psychology: General, 130*, 169–183.

Keith, N., & Frese, M. (2005). Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology, 90*, 677-691.

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin, 29*, 371–381.

Kluger, A.N., DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Kozlowski, S.W.J. (2012). The nature of organizational psychology. In S.W.J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology* (Vol. 1, pp. 3-21). New York: Oxford University Press.

Kozlowski, S.W.J., Gully, S.M., Brown, K.G., Salas, E., Smith, E.M., & Nason, E.R. (2001). Effects of training goals and goal orientation traits on multidimensional training outcomes and performance adaptability. *Organizational Behavior and Human Decision Processes, 85*, 1-31.

Kozlowski, S.W.J., Chao, G.T., Grand, J.A., Braun, M.T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods, 16*, 581-615.

Kozlowski, S.W.J., & Klein, K.J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K.J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco, CA: Jossey-Bass.

Kraiger, K., Ford, J.K, & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328.

Krendl, A.C., Richeson, J.A., Kelley, W.M., & Heatherton, T F. (2008). The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math. *Psychological Science, 19*, 168-175.

Kruschke, J.K. (2015). *Doing Bayesian data analysis*. (2nd ed.). Burlington, MA: Academic Press.

Kruschke, J.K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142, 573-603.

Kruschke, J.K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722-752.

Law, A.M., & Kelton, D.W. (1991). *Simulation modeling and analysis* (2nd ed.). New York: McGraw-Hill.

Lipshitz, R., Levy, D.L., & Orchen, K. (2006). Is this problem to be solved? A cognitive schema of effective problem-solving. *Thinking and Reasoning, 12,* 413-430.

Loewenstein, M.A., & Speltzer, J.R. (2000). Formal and informal training: Evidence from the NLSY. *Research in Labor Economics, 18*, 403-438.

MacDonald, M.C., Just, M.A., & Carpenter, P.A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive psychology*, *24*, 56-98.

Mangels, J.A., Good, C., Whiteman, R.C., Maniscalo, B., & Dweck, C.S. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognition and Affective Neuroscience, 7*, 230-241.

Marx, D.M., Brown, J.L., & Steele, C.M. (1999). Allport's legacy and the situational press of stereotypes. *Journal of Social Issues, 55*, 491-502.

McKown, C., & Weinstein, R.S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development, 74*, 498-515.

Medin, D.L., Ross., N.O., Atran, S., Cox, D., Coley J., Proffitt, J.B., & Blok, S. (2006). Folkbiology of freshwater fish. *Cognition, 99*, 237-273.

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems*. Princeton, NJ: Princeton University Press.

Nguyen, H-.H., & Ryan, A.M. (2008). Does stereotype threat test affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*, 1314-1334.

Oswald, D.L., & Harvey, R.D. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology: Developmental, Learning, Personality, Social, 19*, 338-356.

Pitner, R.O., Astor, R.A., Benbenishty, R., Haj-Yahia, M.M., & Zeira, A. (2003). The effects of group stereotypes on adolescents' reasoning about peer retribution. *Child Development, 74*, 413-425.

Pollack, E. (2013, October 3). Why are there still so few women in science? *The New York Times Magazine*. Retrieved from http://www.nytimes.com/2013/10/06/magazine/why-are-there-still-so-few-women-in-science.html?_r=0.

Ployhart, R.E., & Moliterno, T.P. (2011). Emergence of the human capital resource: A multilevel model. *Academy of Management Review, 25*, 127-150.

Ployhart, R.E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management, 32*, 868-897.

Ployhart, R.E., Ziegert, J.C., & McFarland, L.A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*, 231-259.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

R Core Team (2015). *R: A language and environment for statistical computing*. [Software]. R Foundation for Statistical Computing, Vienna, Austria (https://www.R-project.org/).

Roberson, L, & Kulik, C.T. (2007). Stereotype threat at work. *Academy of Management Perspectives, 21*, 24-40.

Rydell, R.J., Rydell, M.T., & Boucher, K.L. (2010). The effect of negative performance stereotypes on learning. *Journal of Personality and Social Psychology, 99*, 883-896.

Rydell, R.J., Shiffrin, R.M., Boucher, K.L., Van Loo, K., & Rydell, M.T. (2010). Stereotype threat prevents perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 14042-14047.

Sackett, P.R., Hardison, C.M., & Cullen, M.J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59*, 7–13.

Sackett, P.R., & Ryan, A.M. (2012). Concerns about generalizing stereotype threat research findings to operational high stakes testing. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 249-263). New York, NY: Oxford Press.

Sackett, P.R., Schmitt, N., Ellingson, J.E., & Kabin, M.B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318.

Schimel, J., Arndt, J., Banko, K.M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition, 22*, 75-99.

Schmader, T., & Beilock, S. (2012). An integration of processes that underlie stereotype threat. In M. Inzlicht & T. Schmader (Eds), *Stereotype threat: Theory, process, and application* (pp. 34-50). New York, NY: Oxford Press.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*, 440–452.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115*, 336-356.

Schneider, W., & Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search, and Attention, *Psychological Review, 84*, 1-66.

Schoenfeld, A.H., & Herrmann, D.J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 484–494.

Schuelke, M.J., Day, E.A., McEntire, L.E., Boatman, P.R., Boatman, J.E., Kowollik, V., & Wang, X. (2009). Relating indices of knowledge structure coherence and accuracy to skill-based performance: Is there utility in using a combination of indices? *Journal of Applied Psychology, 94*, 1076-1085.

Scullen, S.E., Bergey, P.K., Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology, 58*, 1-32.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63,* 129-138.

Sonnentag, S., & Frese, M. (2012). Dynamic performance. In S.W.J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology* (Vol. 1, pp. 548-575). New York: Oxford University Press.

Spencer, S.J., Steele, C.M., & Quinn, D.M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4-28.

Steele, C.M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629.

Steele, C.M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

Steele, C.M., & Davies, P.G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance, 16*, 311-326.

Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in white athletes. *Personality and Social Psychology Bulletin*, *28*, 1667-1678.

Stricker, L.J., & Ward, W.C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665-693.

Stricker, L.J., & Ward, W.C. (2008). Stereotype threat in applied settings re-examined: A reply. *Journal of Applied Social Psychology, 38*, 1656-1663.

Taylor, V.J., & Walton, G.M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin, 37*, 1055-1067.

Vancouver, J.B., & Weinhardt, J.M. (2012). Modeling the mind and the milieu: Computational modeling for micro-level organizational researchers. *Organizational Research Methods, 15*, 602-623.

Walton, G.M., & Cohen, G.L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39*, 456-467.

Walton, G.M., & Spencer, S.J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*, 1132-1139.

Walton, G.M., Murphy, M.C., & Ryan, A.M. (2015). Stereotype threat in organizations: Implications for equity and performance. *Annual Review of Organizational Psychology and Organizational Behavior, 2,* 523-550.

Weaver, J.L., Bowers, C.A., Salas, E., & Cannon-Bowers, J.A. (1995). Networked simulations: New paradigms for team performance research. *Behavioral Research Methods, Instruments, & Computers, 27*, 12–24.

Whaley, A.L. (1998). Issues of validity in empirical tests of stereotype threat theory. *American Psychologist, 53*, 679-680.

Whitney, P., Ritchie, B.G., & Clark, M.B. (1991). Working-memory capacity and the use of elaborative inferences in text comprehension. *Discourse Processes*, *14*, 133-145.

Wraga, M., Helt, M., Jacobs, E., & Sullivan, K. (2006). Neural basis of stereotype-induced shifts in women's mental rotation performance. *Social Cognition and Affective Neuroscience, 2*, 12-19.

Yuan, Y., & MacKinnon, D.P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*, 301-322.

Zyphur, M.J., & Oswald, F.L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41*, 390-420.

Footnotes

[1] The analysis of data from the male participants was not necessary or relevant to testing any of the proposed hypotheses. Consequently, no data from the male participants is included in any of the reported analyses. Men were recruited into the study to improve the fidelity of the training sessions (i.e., organizational training is not typically split into male- and female-only trainee groups) and heighten the relevance and salience of the gender stereotype for females (Walton et al., 2015). For purposes of full reporting, 44 males (N_{ST} = 22, $N_{control}$ = 22) completed all three days of the training.

[2] The working memory assessment was collected as a potential control variable. Inclusion of this variable did not change the substantive conclusions of the analyses and so is not considered further.

[3] The R code and datasets for conducting all the Bayesian analyses in this paper are available in the Supplemental Materials accompanying this manuscript.

[4] Formula for effect size computation: $d = (\mu_{\text{Control}} - \mu_{\text{ST}}) \,/\, \sqrt{(\sigma^2_{\text{Control}} + \sigma^2_{\text{ST}})/2}$

[5] Derivation of the most diagnostic heuristic associations among classification and engagement decisions as well as evidence for the accuracy of these heuristics is provided in the Supplemental Materials.

[6] Alternative theoretical conceptualizations of how human capital resource should be represented at the organizational level have been suggested that do not rely on the mean of individual-level KSAOs (e.g., Ployhart, 2006; Aguinis & O'Boyle, 2014). The various operationalizations suggested by these models (e.g., variance, maximum, etc.) could easily be accommodated in the present simulation. However, the present model demonstrates that the consequential impact of ST effects are *cumulative*; that is, they compound and exacerbate over time. Such cumulative effects should be expected to permeate most all summary characteristics of a distribution, and thus the overall interpretations should be similar.

[7] The R code for running the computational model and simulations described in this paper are available in the Supplemental Materials accompanying this manuscript.

[8] The apparent "spike" observed at time *t* = 2 for the individual-level effect sizes shown in Figure 9 reflects the interaction between the decreasing power law used to operationalize performance improvement in the simulation and the impact of turnover in the organization. Differences in the rate parameter of a power law function (*b* in Equation 3) lead to the largest discrepancies when *t* is small. However, as *t* increases and the power law function approaches its lower asymptote, resultant differences in the rate parameter become less pronounced. Additionally, the simulated employees in every organization were "new" employees (tenure = 0) at the start of the simulation.

Consequently, for the first few time points, organizations were primarily populated by employees with low tenures (small *t*) and thus the discrepancy in observed performance improvement rate was larger during these time points. As organizations grew to include more employees with longer tenures, the drop-off in the observed average individual-level performance improvement at each time point decreases and leads to smaller effect sizes in the improvement of employee performance potential across conditions of ST.

[9] This assumption is neither unsupported nor unprecedented. For example, child development researchers have found that awareness of group stereotypes and the recognition that group stereotypes are used by others to make judgments develop at relatively early ages and continue to strengthen into adulthood (McKown & Weinstein, 2003; Pitner, Astor, Benbenishty, Haj-Yahia, & Zeira, 2003). Furthermore, this is the same rationale of existing research that has sought to identify ST effects by examining group differences in archival performance data (e.g., Cullen et al., 2004; Cullen et al., 2006; Stricker & Ward, 2004), though such studies failed to measure critical person and/or situational characteristics.

Table 1
Knowledge Concepts in TANDEM

Focus	Concept	Description	KS Label
Decision-making	1. Identify contact Type as Air	Classify contact as an Aircraft	idAir
	2. Identify contact Type as Surface	Classify contact as Surface	idSurf
	3. Identify contact Type as Sub	Classify contact as Submarine	idSub
	4. Identify contact Class as Civilian	Classify contact as Civilian	idCiv
	5. Identify contact Class as Military	Classify contact as Military	idMil
	6. Identify contact Intent as Peaceful	Classify contact as Peaceful	idPeac
	7. Identify contact Intent as Hostile	Classify contact as Hostile	idHost
	8. Make decision to Clear contact	Process contact as Clear	Clear
	9. Make decision to Warn contact	Process contact as Warn	Warn
	10. Make decision to Mark contact	Process contact as Warn	Mark
Task Operations	11. Gain/lose points	Indicator of task performance	Points
	12. Zoom out/zoom in	Change radar display resolution	Zoom
	13. Monitor inner perimeter	Track potential boundary intrusions across inner perimeter	MonInn
	14. Monitor outer perimeter	Track potential boundary intrusions across outer perimeter	MonOut
	15. Find/engage pop-up targets	Classify/engage new targets	PopUp
	16. Prioritize targets (engage targets likely to cross perimeter first)	Process targets likely to cross a defensive perimeter	Priority

Note. KS = knowledge structure. Terms listed in the Concept column present the concept labels used by participants to make similarity ratings. Terms listed in the KS Label column identify the label used for each concept in the knowledge structure figures (Figure 2).

Table 2
Descriptive Statistics and Correlations of Study Variables

Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Condition	.49	.50	—													
2. Perceived stereotype threat	3.09	.54	**.37**	(.67)												
3. Declarative knowledge (T1)	.63	.19	.02	.08	(.60)											
4. Declarative knowledge (T2)	.76	.19	-.16	.10	**.49**	(.65)										
5. Declarative knowledge (T3)	.74	.20	-.16	**.20**	**.52**	**.63**	(.69)									
6. Metacognitive activity (T1)	3.76	.54	-.09	-.03	**.35**	**.23**	**.24**	(.86)								
7. Metacognitive activity (T2)	3.82	.63	-.14	.05	**.30**	**.35**	**.35**	**.68**	(.91)							
8. Metacognitive activity (T3)	3.80	.78	**-.22**	.02	**.23**	**.22**	**.37**	**.59**	**.82**	(.95)						
9. Study time: Process contacts (L1)	78.04	17.25	.08	.00	**.17**	.08	.00	.06	.00	.01	—					
10. Study time: Process contacts (L2)	68.29	29.33	-.06	.04	-.15	.08	.04	-.02	.07	.12	.09	—				
11. Study time: Process contacts (L3)	50.53	39.02	**-.36**	-.10	-.12	-.04	.08	.01	.04	.12	.21	**.32**	—			
12. Study time: Task strategies (L1)	24.38	14.93	.04	.11	.05	.13	.10	**.26**	**.18**	.15	**-.69**	-.13	**-.29**	—		
13. Study time: Task strategies (L2)	27.68	21.48	-.09	.09	**.20**	**.23**	**.30**	.06	.09	.11	**.07**	**-.61**	-.09	.11	—	
14. Study time: Task strategies (L3)	14.12	20.59	**-.31**	.04	.13	**.25**	**.22**	-.07	-.06	.02	.00	**-.19**	-.06	-.04	**.50**	—
15. Number contacts processed (L1)	8.85	2.21	**.20**	.08	**.18**	.11	-.02	**.21**	.16	.12	**.21**	.02	-.04	**.17**	.02	-.16
16. Number contacts processed (L2)	10.93	2.21	.08	-.10	.14	.02	-.03	.13	.10	.06	**.21**	.07	.12	-.04	-.09	-.17
17. Number contacts processed (L3)	11.66	2.54	.11	.03	**.30**	**.25**	.16	**.23**	**.26**	.11	.04	.08	-.12	**.24**	-.02	-.09
18. Number contacts processed (P1)	18.35	4.11	.14	-.10	.11	-.07	-.17	.15	-.02	.01	**.21**	.09	.04	.02	-.14	**-.19**
19. Number contacts processed (P2)	18.37	4.19	.06	-.08	.02	.05	-.16	-.03	.02	-.05	.16	.11	.09	-.01	.00	-.10
20. Number contacts processed (P3)	19.61	4.77	.08	.05	.12	.11	.06	.05	.06	-.03	.01	.01	-.16	.18	-.05	-.09
21. Task performance (P1)	-2325.5	728.2	.02	.00	**.37**	**.33**	**.43**	**.21**	**.24**	.14	.13	**-.24**	-.15	.07	**.44**	.18
22. Task performance (P2)	-1586.9	1227.2	-.04	.08	**.51**	**.55**	**.50**	**.21**	**.33**	**.23**	-.06	-.08	**-.24**	**.29**	**.27**	.12
23. Task performance (P3)	-1015.8	1467.1	-.15	.06	**.47**	**.64**	**.60**	**.26**	**.36**	**.33**	.05	.07	-.12	**.24**	**.24**	**.20**

Note. Values in bold signify correlations where $p < .05$. Values reported on the diagonal are reliabilities (Cronbach's alpha). T1, T2, and T3 reflect measures taken at the end of Days 1, 2, and 3, respectively. L1, L2, and L3 reflect the value of variables averaged over the learning/practice trials on Days 1, 2, and 3, respectively. P1, P2, and P3 reflect measures taken during the performance trials on Days 1, 2, and 3, respectively.

	15	16	17	18	19	20	21	22	23
1. Condition									
2. Perceived stereotype threat									
3. Declarative knowledge (T1)									
4. Declarative knowledge (T2)									
5. Declarative knowledge (T3)									
6. Metacognitive activity (T1)									
7. Metacognitive activity (T2)									
8. Metacognitive activity (T3)									
9. Study time: Process contacts (L1)									
10. Study time: Process contacts (L2)									
11. Study time: Process contacts (L3)									
12. Study time: Task strategies (L1)									
13. Study time: Task strategies (L2)									
14. Study time: Task strategies (L3)									
15. Number contacts processed (L1)	—								
16. Number contacts processed (L2)	**.62**	—							
17. Number contacts processed (L3)	**.45**	**.60**	—						
18. Number contacts processed (P1)	**.66**	**.70**	**.33**	—					
19. Number contacts processed (P2)	**.47**	**.75**	**.52**	**.44**	—				
20. Number contacts processed (P3)	**.38**	**.50**	**.79**	**.25**	**.54**	—			
21. Task performance (P1)	.03	-.08	.13	**-.28**	.02	.01	—		
22. Task performance (P2)	.10	-.11	**.28**	**-.18**	-.04	.14	**.62**	—	
23. Task performance (P3)	.14	-.06	**.26**	-.13	-.03	.11	**.54**	**.82**	—

Table 3
Posterior Parameter Estimates from Hierarchical Bayesian Regression for Cognitive Learning Outcomes

Variables	*b*	95% Credibility Interval	% Posterior Above/Below Zero
DV: Declarative Knowledge[a]			
Intercept (β_{00})	.68	[.48, .89]	100% above
Condition (β_{01})	.01	[-.28, .31]	54% above
Day (β_{10})	.47	[.31, .63]	100% above
Condition*Day (β_{11})	-.25	[-.46, -.03]	99% below
DV: Metacognitive Activity			
Intercept (β_{00})	3.81	[3.69, 3.94]	100% above
Condition (β_{01})	-.09	[-.27, .09]	84% below
Day (β_{10})	.06	[-.02, .14]	94% above
Condition*Day (β_{11})	-.11	[-.23, .00]	97% below

[a]Coefficients reflect log odds

Note. All coefficients (*b*) are reported in original unstandardized units and represent the parameter estimate with the highest posterior probability. Interpretation of each coefficient is as follows:

β_{00} = intercept for Control condition at first time period

β_{01} = difference between ST and Control condition intercept at first time period

β_{10} = average change in DV over time for Control condition

β_{11} = difference between ST and Control condition in change in DV over time

Table 4
Descriptive Summary of Aggregate Knowledge Structures for Females in the Control and Stereotype Threat Conditions Over Time

Day	Condition	Max Degree	Central	Common Links	Similarity	Similarity minus Chance
1	Control	idHost	Points PopUp	6	.25	.18
	ST	idHost	Mark			
2	Control	Points	Points MonInn	9	.43	.36
	ST	Points	Points			
3	Control	Points	Points MonInn	7	.29	.22
	ST	Points Warn Mark Zoom	Points Mark			

Note. ST = Stereotype threat condition. Max Degree indicates the knowledge concept with the most associative links. Central indicates the knowledge concept with the minimum number of links between it and all other concepts. Similarity indicates the proportion of common to unique links between control and ST knowledge structures (Common Links / ($Links_{Control}$ + $Link_{ST}$ – Common Links)). Similarity minus Chance indicates the proportion of common to unique links between control and ST knowledge structures controlling for the number of common links expected by chance.

Table 5
Posterior Parameter Estimates from Hierarchical Bayesian Regression for Learning Engagement Behaviors

Variables	*b*	95% Credibility Interval	% Posterior Above/Below Zero
DV: Study time—Processing contacts[a]			
Intercept (β_{00})	.29	[.15, .42]	100% above
Condition (β_{01})	.31	[.11, .50]	100% above
Trial (β_{10})	-.03	[-.06, -.01]	100% below
Condition*Trial (β_{11})	-.07	[-.10, -.04]	100% below
DV: Study time—Task strategies[a]			
Intercept (β_{00})	-2.04	[-2.32, -1.76]	100% below
Condition (β_{01})	.08	[-.31, .48]	66% above
Trial (β_{10})	-.46	[-.56, -.36]	100% below
Condition*Trial (β_{11})	-.14	[-.28, .00]	98% below
Trial^2 (β_{20})	.04	[.03, .05]	100% above
Condition*Trial^2 (β_{21})	.03	[.01, .04]	100% above
DV: Number contacts processed during practice trials			
Intercept (β_{00})	7.77	[7.26, 8.26]	100% above
Condition (β_{01})	.83	[.11, 1.54]	99% above
Trial (β_{10})	.30	[.25, .34]	100% above
Condition*Trial (β_{11})	-.03	[-.10, .03]	83% below

[a]Coefficients reflect log odds

Note. See note in Table 3 for description of regression coefficients β_{00} through β_{11}.
β_{20} = average change in change over time for Control condition
β_{21} = difference between ST and Control condition in change in change over time

Table 6
Posterior Parameter Estimates from Hierarchical Bayesian Regression for Performance Outcomes

Variables	*b*	95% Credibility Interval	% Posterior Above/Below Zero
DV: Performance trial score			
Intercept (β_{00})	-2320	[-2510, -2140]	100% below
Condition (β_{01})	57.8	[-216, 326]	66% above
Day (β_{10})	764	[598, 930]	100% above
Condition*Day (β_{11})	-246	[-487, -14.9]	98% below
DV: Number contacts processed during performance trials			
Intercept (β_{00})	17.7	[16.8, 18.5]	100% above
Condition (β_{01})	1.04	[-.24, 2.28]	95% above
Day (β_{10})	.57	[-.09, 1.23]	95% above
Condition*Day (β_{11})	-.17	[-1.10, .77]	64% below

Note. See note in Table 3 for description of regression coefficients

Table 7
Pseudocode for Computational Model and Simulation of Stereotype Threat Effects During Learning/Training

Step	Description
1	Initialize time clock to $t = 0$
2	Create k organizations each containing n employees, with $k/2$ organizations containing ST
3	Create a voluntary turnover schedule for each organization
4	Increment time clock to $t = t + 1$
5	Determine change in employee performance potential as a result of learning
6	Compute cumulative employee and organizational performance potential
7	Invoke voluntary turnover and immediate replacement scheduled for time t
8	If $t < t_{stop}$, return to Step 4
9	Stop simulation

Note. ST = stereotype threat; t = time period. For all simulations reported in the manuscript, $k = 500$, $n = 100$, and $t_{stop} = 30$.

Figure 1. Regression estimated changes in proportion of declarative knowledge items answered correctly by female participants in the control and stereotype threat conditions.

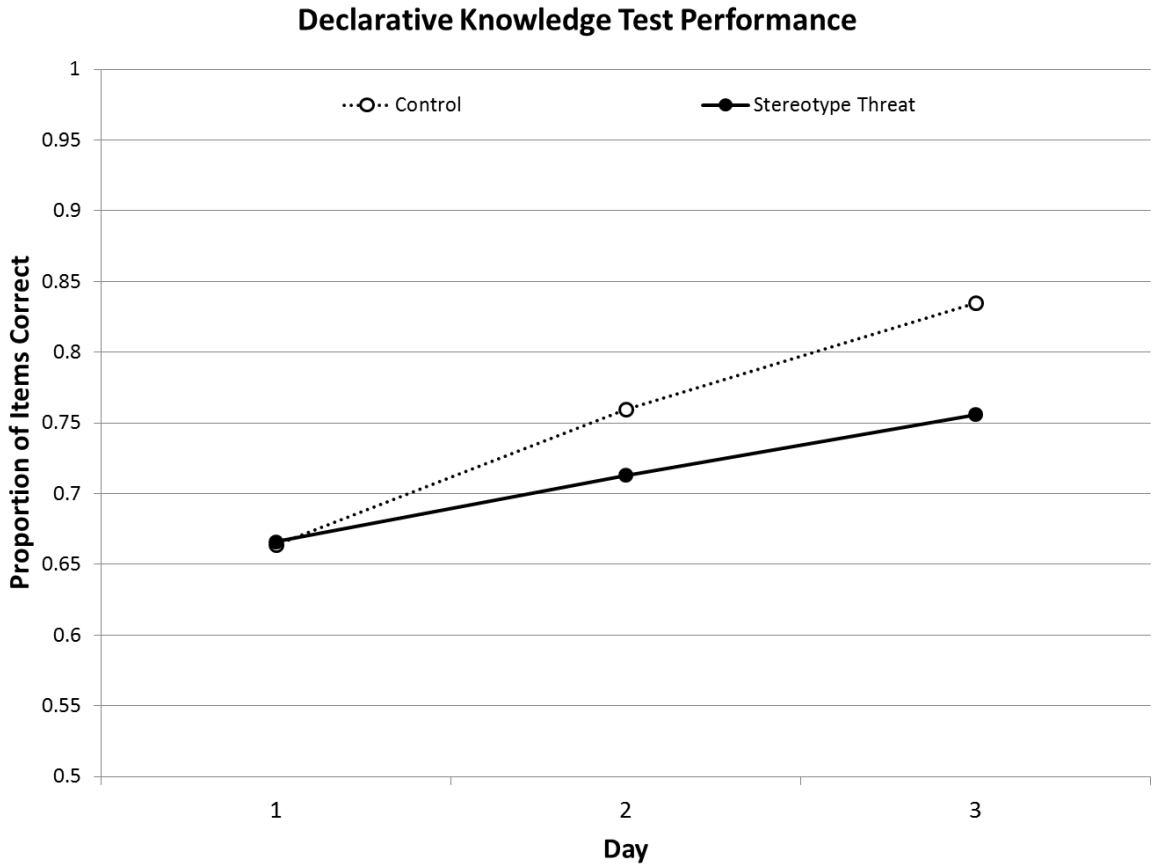
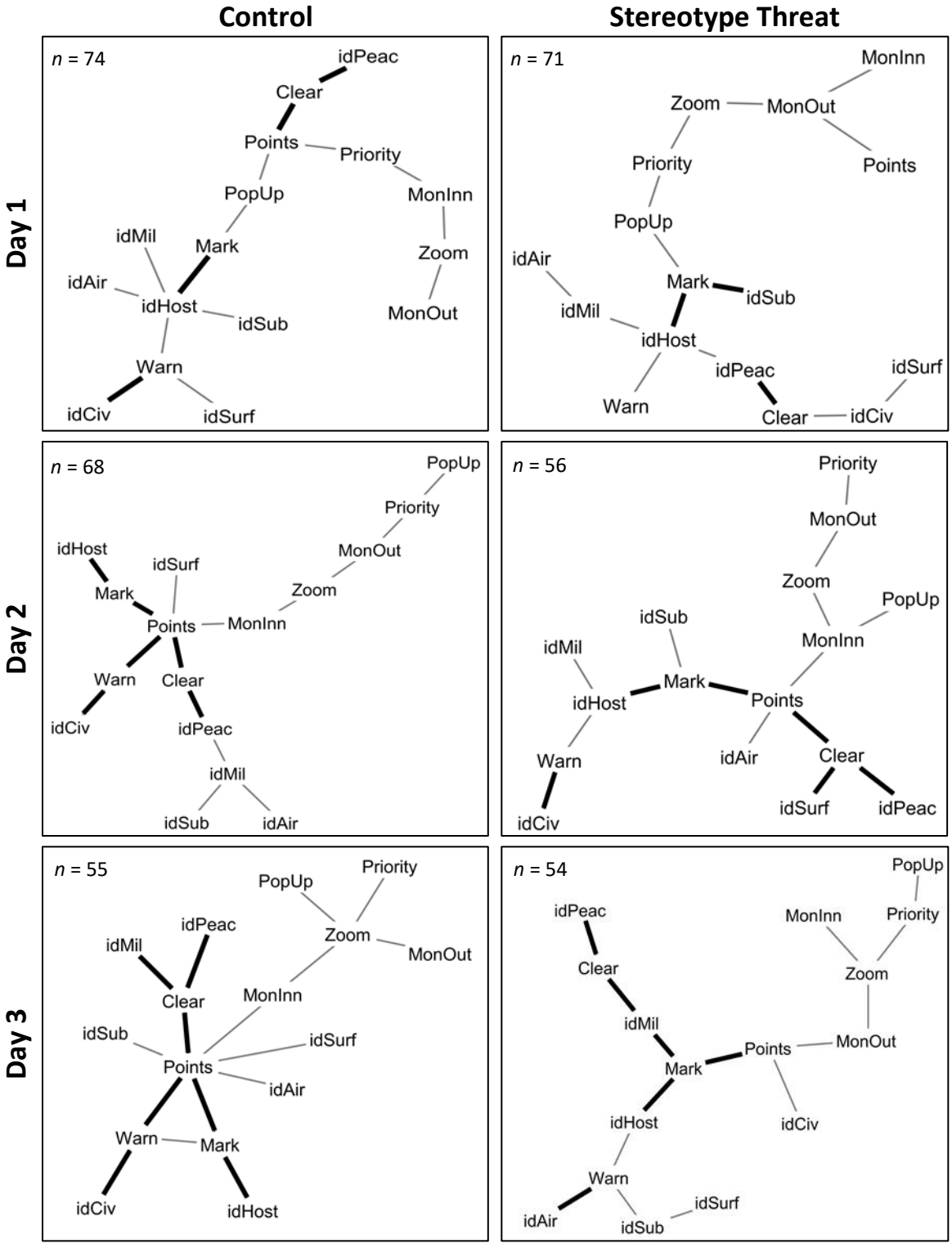


Figure 2. Aggregate knowledge structures for female participants in the control and stereotype threat conditions across day.



Note. Concept labels are defined in Table 1. *n* reflects the number of individual knowledge structures used to create the aggregate knowledge structure shown at each day. Bolded links highlight associations among concepts that share heuristic and performance-oriented relationships.

Figure 3. Regression estimated changes in metacognitive activity for female participants in the control and stereotype threat conditions.

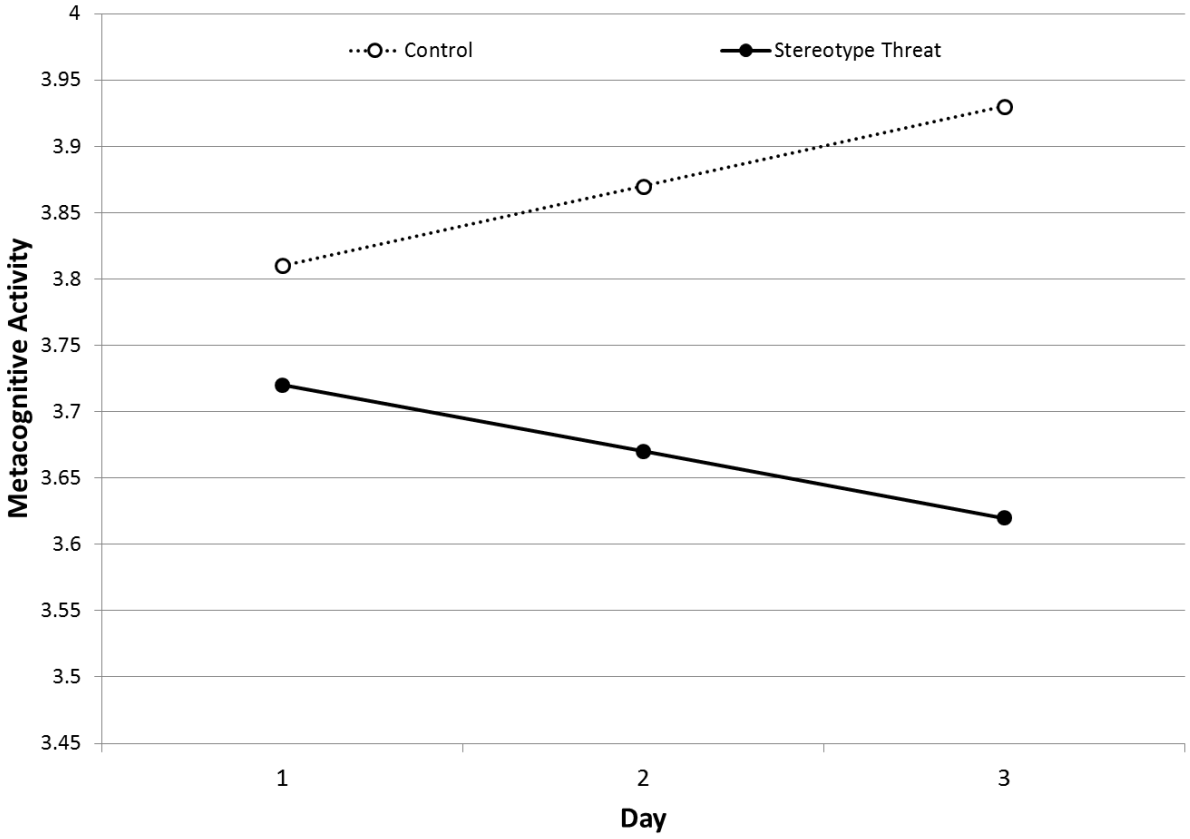
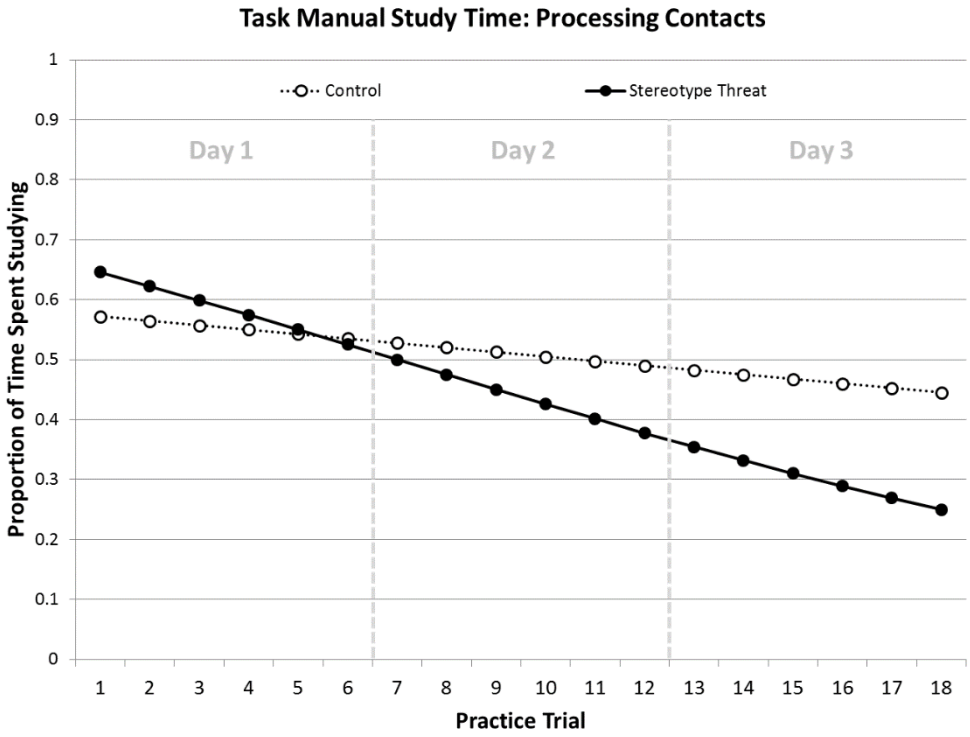


Figure 4. Regression estimated changes in (a) time spent studying how to process contacts and (b) time spent studying task strategies during practice trials for female participants in the control and stereotype threat conditions.

a



b

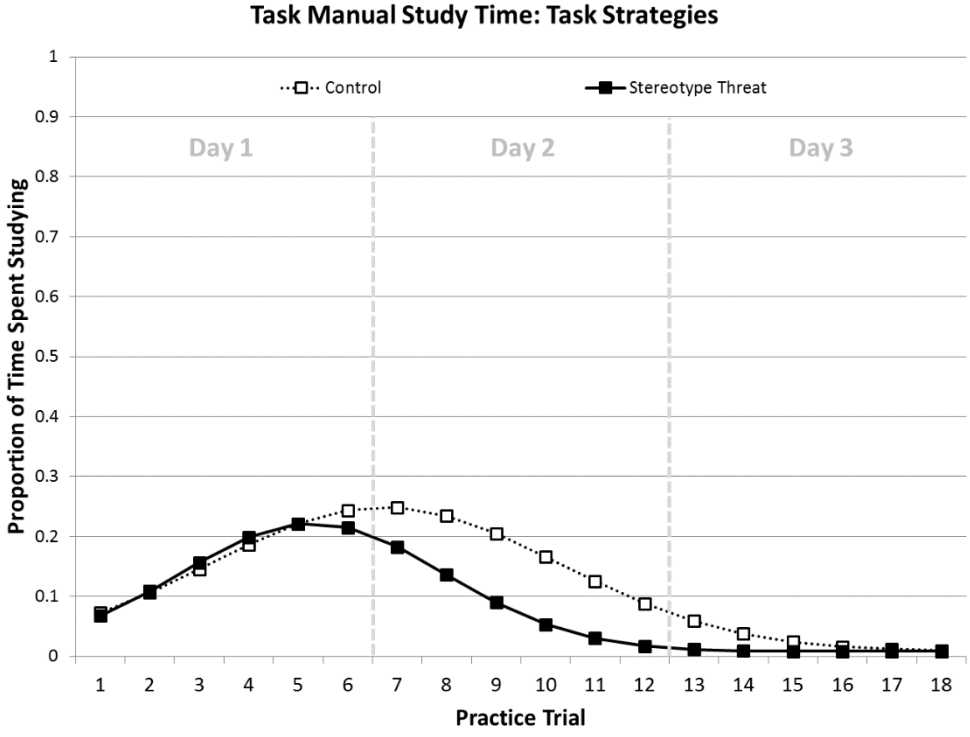
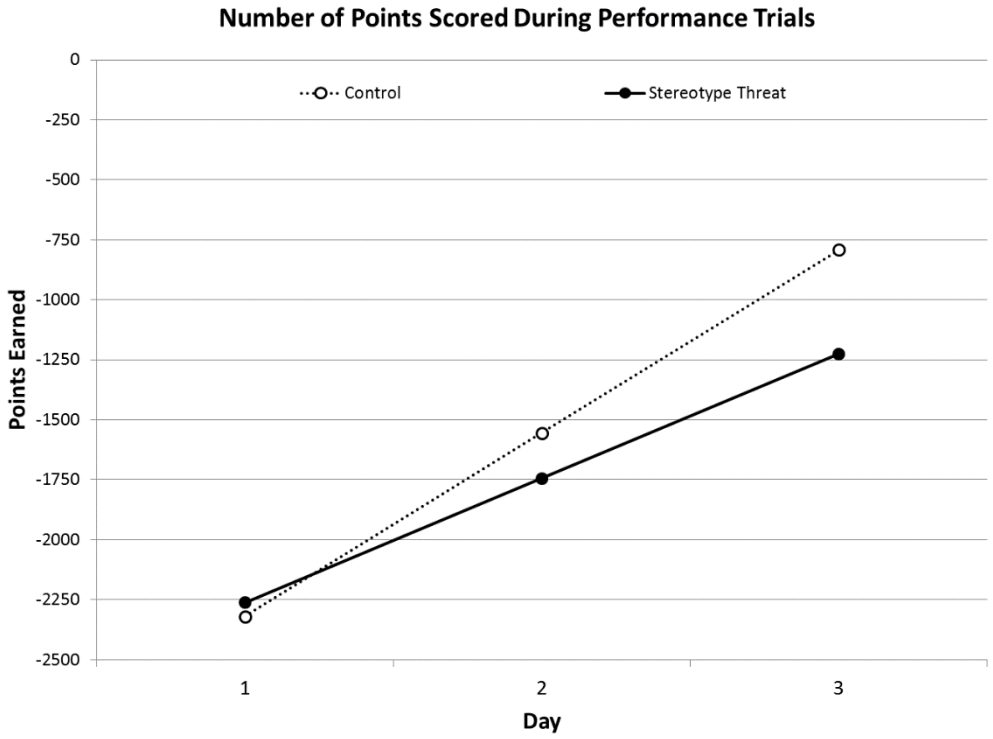


Figure 5. Regression estimated changes in (a) number of points earned and (b) number of contacts processed during the performance trials for female learners in the control and stereotype threat conditions.

a



b

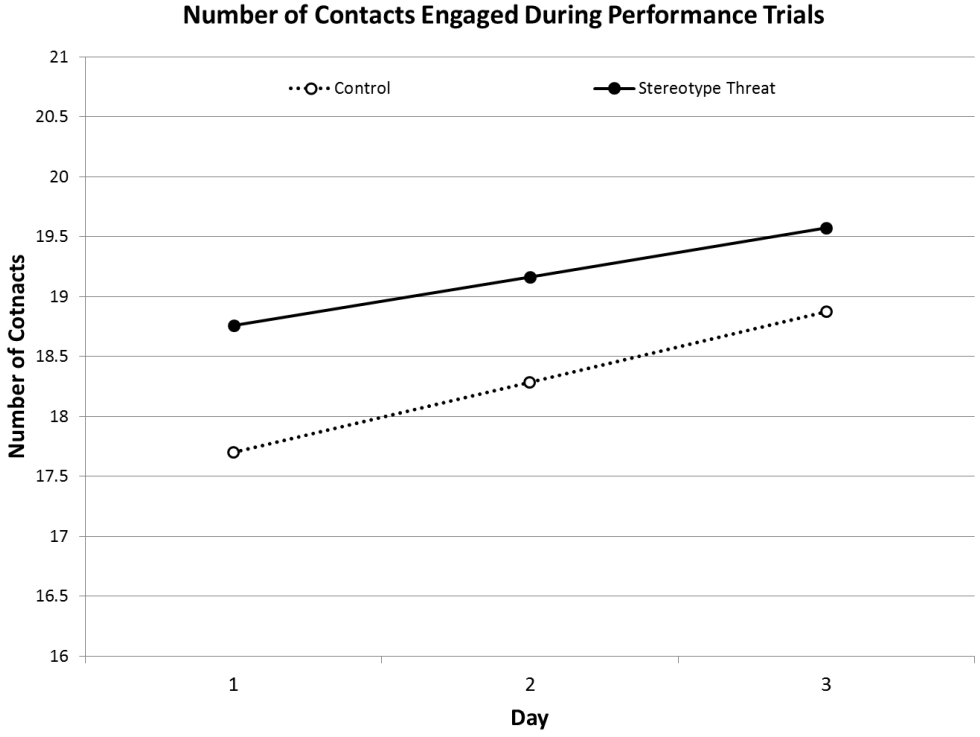
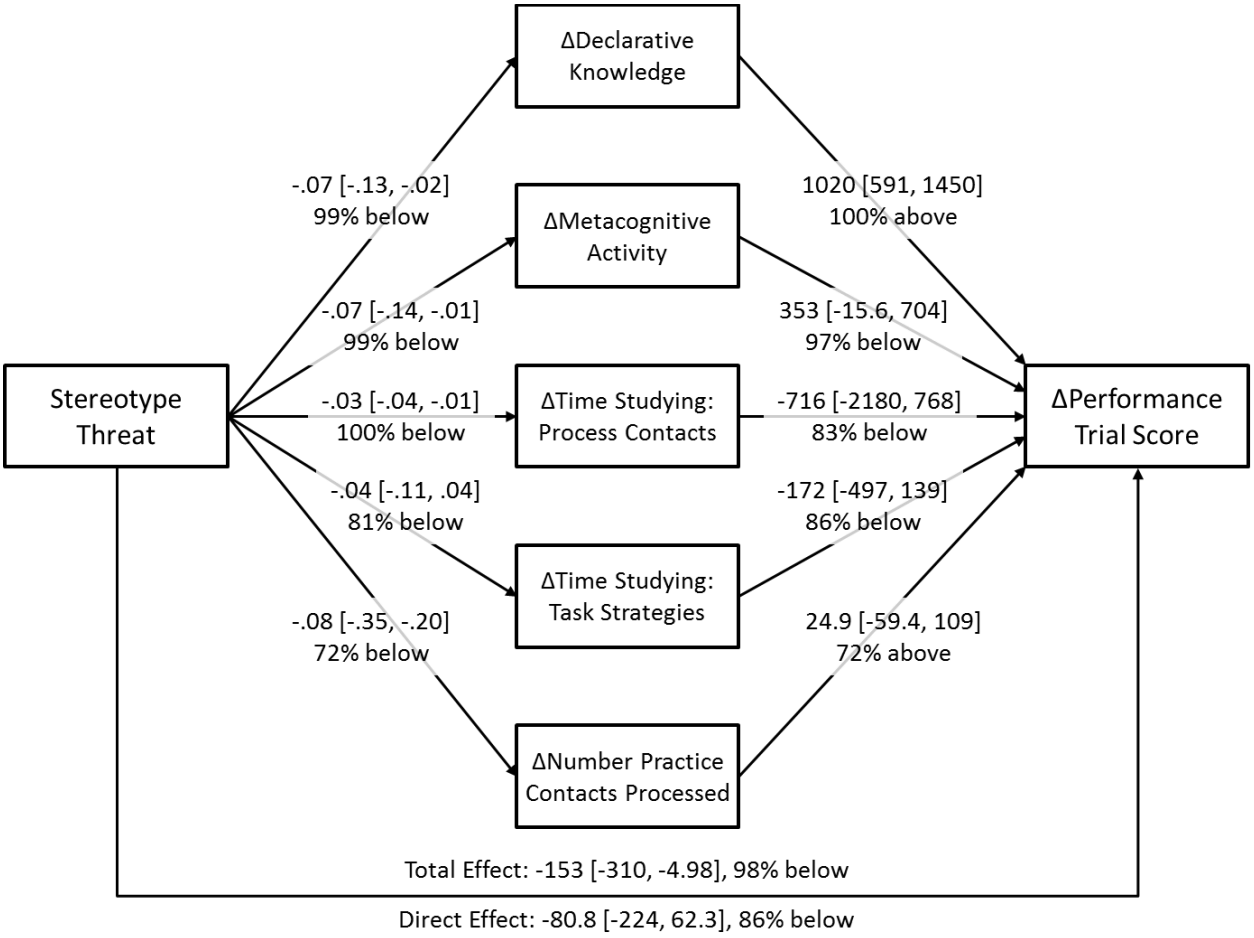
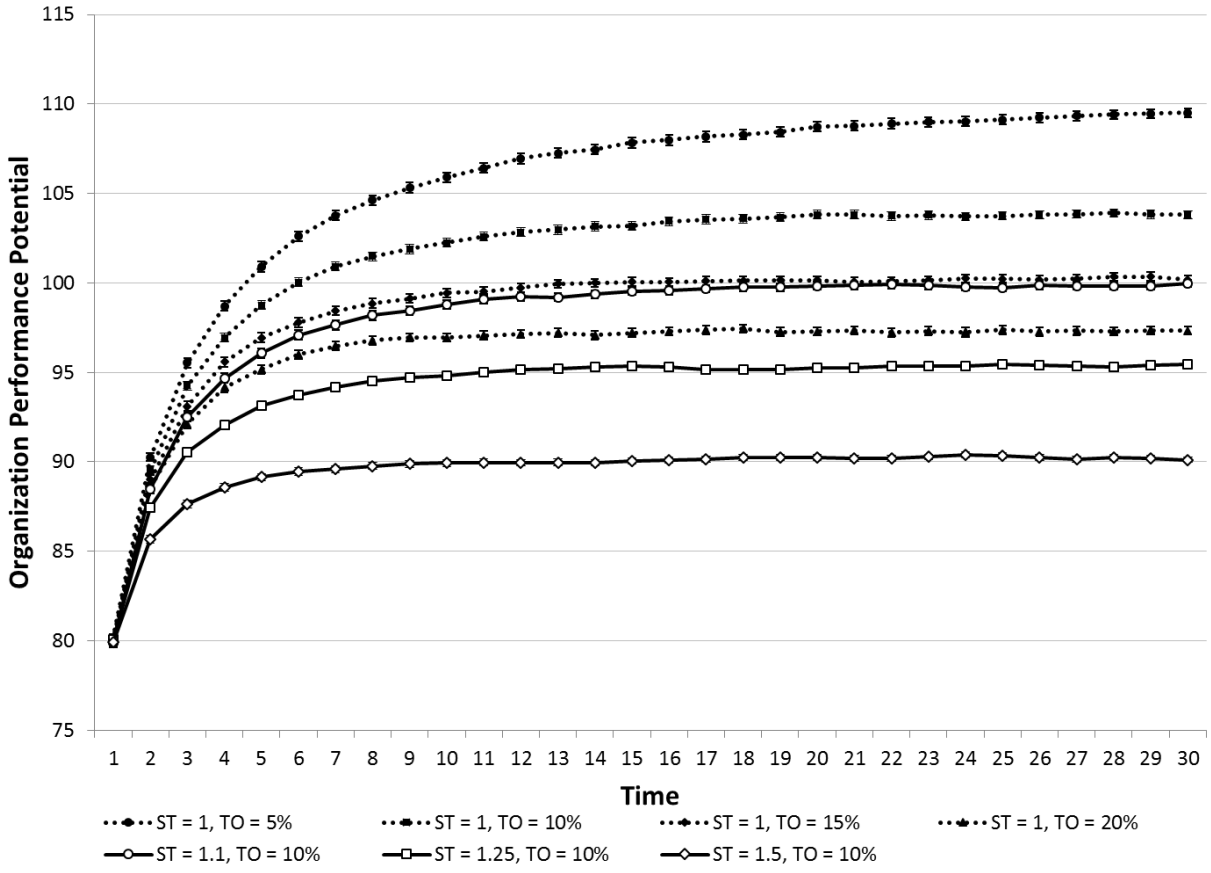


Figure 6. Summary of results for mediation analyses testing influence of stereotype threat on performance improvement though rates of change in knowledge acquisition and learning engagement.



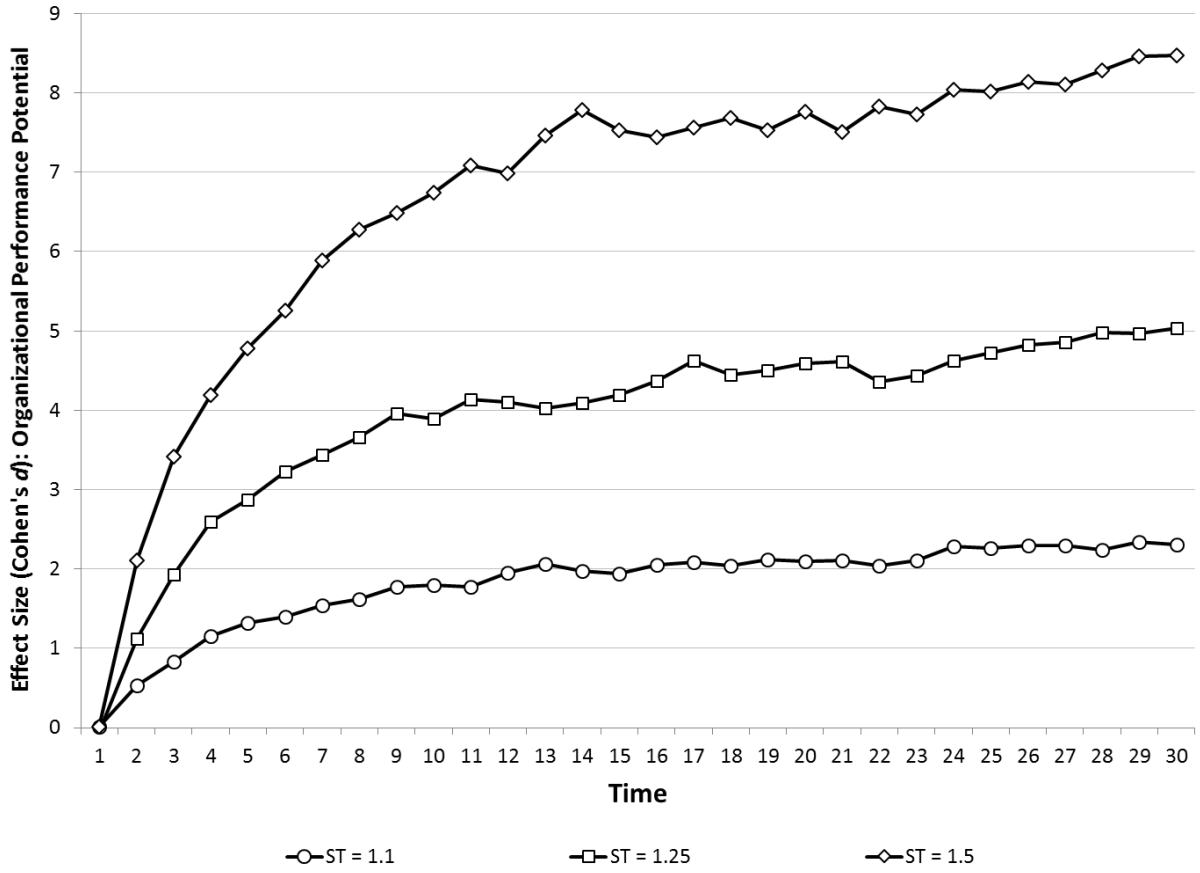
Note. Values in brackets represent 95% credibility intervals of the estimated regression parameters. % above/below represents the proportion of the posterior distribution of the estimated regression parameter that is above/below zero. Stereotype Threat was a dummy coded variable (0 = control, 1 = stereotype threat). The mediator and outcome variables used in the mediation analyses were regression coefficients representing the estimated change in each variable over time (see Equation 2). The ΔDeclarative Knowledge, ΔTime Studying: Process Contacts, and ΔTime Studying: Task Strategies variables are in log-odds units, while the ΔMetacognitive Activity, ΔNumber Practice Contacts Processed, and ΔPerformance Trial Score variables are in their unstandardized units.

Figure 7. Average organizational performance potential for simulated organizations with different levels of stereotype threat in learning/training practices and voluntary turnover rates.



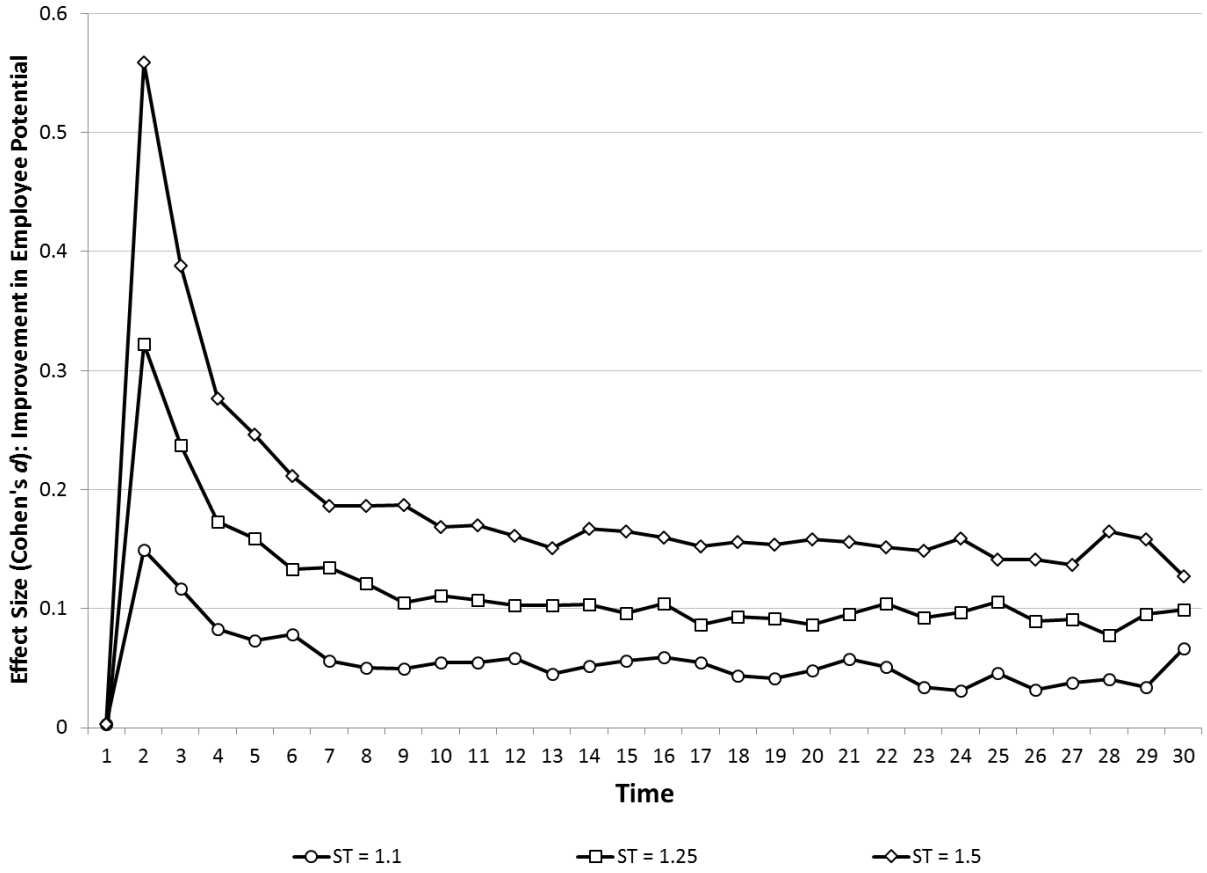
Note. ST = relative difference in rate of change in employee performance potential of agents experiencing versus not experiencing stereotype threat (θ in Equation 4). TO = average voluntary turnover rate across organizations. Error bars reflect the 95% confidence interval for the observed value. Each line summarizes results from a total of 250 simulated organizations comprised of 100 employees.

Figure 8. Average effect size (Cohen's *d*) of differences in organizational performance potential between simulated organizations with versus without stereotype threat effects during learning/training over time.



Note. ST = relative difference in rate of change in employee performance potential of agents experiencing versus not experiencing stereotype threat (θ in Equation 4). Higher values on the y-axis indicates that organizations with no ST achieved higher levels of organizational performance potential relative to organizations with ST. Turnover was equal to 10% in all conditions shown. Each line summarizes results comparing 250 simulated organizations with stereotype threat effects and 250 simulated organizations without stereotype threat effects.

Figure 9. Average effect size (Cohen's *d*) of differences in employee performance potential improvements between agents experiencing versus not experiencing stereotype threat effects during learning/training over time.



Note. ST = relative difference in rate of change in employee performance potential of agents experiencing versus not experiencing stereotype threat (θ in Equation 4). Higher values on the y-axis indicates that agents not experiencing ST improved their performance potential more than agents experiencing ST. Turnover was equal to 10% in all conditions. Each line summarizes results from 25,000 simulated employees experiencing stereotype threat effects against 25,000 simulated employees not experiencing stereotype threat effects.

APPENDIX A
DESCRIPTION OF BAYESIAN ANALYSES

The motivations for using Bayesian inferential statistics methods in the present study were threefold (cf., Kruschke, 2015; Kruschke et al., 2012; Zyphur & Oswald, 2015). First, the results from Bayesian methods allowed direct interpretation of each prediction based on the observed data. Conceptually, Bayesian methods compute the "believability" of values for a parameter (e.g., a mean, regression coefficient, variance, etc.) by updating a researcher's explicitly stated beliefs/predictions about the most likely values for those parameters (stated in the form of a *prior* distribution) on the basis of observed values for those parameters (the *likelihood* distribution). The results of this process are the most credible estimates for the modeled parameter given the data that was collected (stated in the form of a *posterior* distribution). Note that Bayesian parameter estimation is not dependent on sample size/stopping rules in the same way as null hypothesis testing or the interpretation of confidence intervals (Kruschke, 2015); empirical observations simply contribute to the believability of the posterior parameter estimates.

Second, Bayesian methods rely on interpretation of credibility intervals rather than confidence intervals for evaluating parameter estimates. In the context of null hypothesis testing, a 95% confidence interval represents the range of parameter values which would not be rejected by a (two-tailed) test for statistical significance that allows for a 5% false alarm rate. An alternative but analytically identical definition is that if the same experiment (with exact same sample size and stopping rule) was replicated repeatedly, the parameter values reported in any particular confidence interval would contain the true population value for that parameter 95% of the time. Note that for either definition, the confidence interval for a single sample reveals no information about the most believable parameter estimate based on the data; furthermore, the interpretation of a confidence interval depends on the sample size and stopping rule intended by the researcher. In contrast, a 95% credibility interval provides the range of parameter values which constitute 95% of a posterior distribution's density (i.e., the range of parameter estimates whose "believability" sums to 95%) and therefore reflects the most believable parameter values based on the observed data. The most probable point estimate from this interval is simply the value with the highest believability in the posterior distribution (typically given by the mean). Thus unlike a confidence interval, the credibility interval offers a direct evaluation of the believable value(s) for a parameter given a researcher's prediction and observations.

Lastly, the results obtained through Bayesian methods offer a natural route to accumulating evidence through empirical replication. As described earlier, the posterior distribution for a parameter offers a complete description of its most believable estimates for a given dataset. Consequently, this distribution can easily be used to inform prior distributions in subsequent replications. Thus future efforts to replicate a particular finding/parameter value can incorporate the findings of previous research and contribute to generative evidence through the use of empirically informed prior distributions.

Hierarchical Bayesian regression models

In order to conduct Bayesian parameter estimation, investigators must specify 1) prior distributions that reflect their explicit beliefs/predictions about the parameter values to be estimated, and 2) a likelihood distribution that adequately reflects the distribution of observed data. As noted in the manuscript text, noncommittal prior distributions were provided for all Level-2 regression coefficients. More specifically, the priors for the Level-2 parameters were specified as uniform distributions whose range encompassed the entirety of believable posterior estimates for the parameters. Such a noncommittal prior allowed the observed data (i.e., the likelihood distribution) to be the primary contributor to estimation of the posterior distribution (cf., Gelman & Hill, 2004). Every analyses was also run using a model in which the priors for the Level-2 regression coefficients were specified as uninformed normal distributions centered at zero [~$N(\mu = 0, SD = 31.6)$] to assess whether the choice of prior distributions may have unduly influenced the results. As expected, given that both sets of prior distribution specifications are largely "weak" and noncommittal, the distribution of posterior parameter estimates under both prior specifications were virtually identical.

Figures A1 through A4 provide graphical representations of the final hierarchical Bayesian regression models evaluated in the present study. The graphical depiction corresponds with those described by Kruschke (2015) and shows each estimated parameters along with its accompanying distribution. The underlying regression models are identical (general linear model; see Equation 1 in manuscript text); the only differences are in the selection of the likelihood distributions used to represent the observed data. A normal likelihood distribution was used to model the data for metacognitive activity, number of contacts processed during practice/performance trials, and points scored during performance trials. Declarative knowledge was operationalized as the proportion of items correct out of 11, and thus was most appropriately modeled using a binomial distribution. Data for the proportion of time spent studying

how to process contacts was bimodal such that, across all trials, there were many observations in which participants spent all 120 seconds reading this section each trial while others spent very little time on this manual section. Consequently, a beta distribution—which can be used to represent bimodal data with modes near the extremes—was used to represent the likelihood distribution for this variable. Parameterization of the beta regression model followed the recommendations of Figueroa-Zúñiga, Arellano-Valle, and Ferrari (2012). Lastly, the observed study times for the task strategies portion of the manual was extremely positively skewed such that the majority of participants spent very little time reading this section each trial with increasingly fewer participants spending more time on this section. Such data is best fit using an exponential likelihood distribution. In the case of regression using an exponential likelihood distribution, model coefficients reflect the log-odd estimates for a linear model predicting the rate parameter (λ) of the exponential probability density function ($\lambda e^{-\lambda x}$). However, it is typically more informative to interpret regression coefficients in terms of their mean rather than their rate. The mean of the exponential probability density function distribution is 1/λ. The following formula can thus be used to interpret the estimated regression coefficients in terms of the mean:

$$1 \Big/ \left(\frac{e^b}{1+e^b}\right), \tag{A1}$$

where *b* is the estimated value for λ (i.e,. the regression coefficient) from the exponential regression model. For example, β_{00} in Table 5 was estimated as -2.04. Substituting -2.04 for *b* in Equation A1 reveals that the average number of seconds spent studying the task strategies portion of the operations manual at Day 1 for female control condition participants was 8.69 seconds.

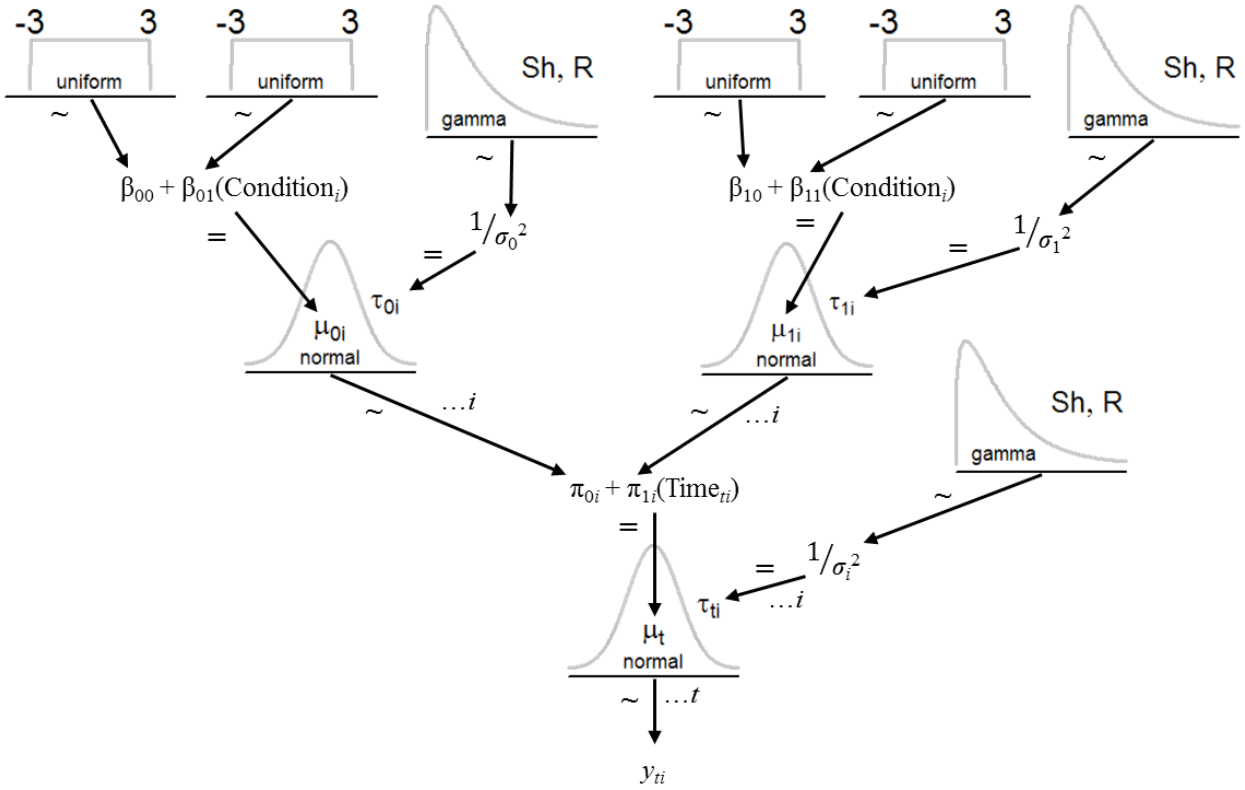


Figure A1. Multilevel hierarchical Bayesian regression model with normal likelihood distribution. This model was used to examine metacognitive activity, number of contacts processed during practice/performance trials, and points scored during performance trials (Tables 3, 5, and 6 in manuscript). The tilde (~) symbol indicates that parameter values were drawn from the corresponding distribution, while the ellipses (…) indicate that multiple parameters were computed across the corresponding units (*t* = time point, *i* = subject). All normal distributions were parameterized using the precision (τ) rather than variance, which is equivalent to 1/variance. The shape (Sh) and rate (R) parameters of the gamma distributions were parameterized according to their mode rather than mean; for clarity of presentation, the distributions over these parameters are not shown. However, the R code available in the Supplemental Materials provides the full specification of these parameterizations.

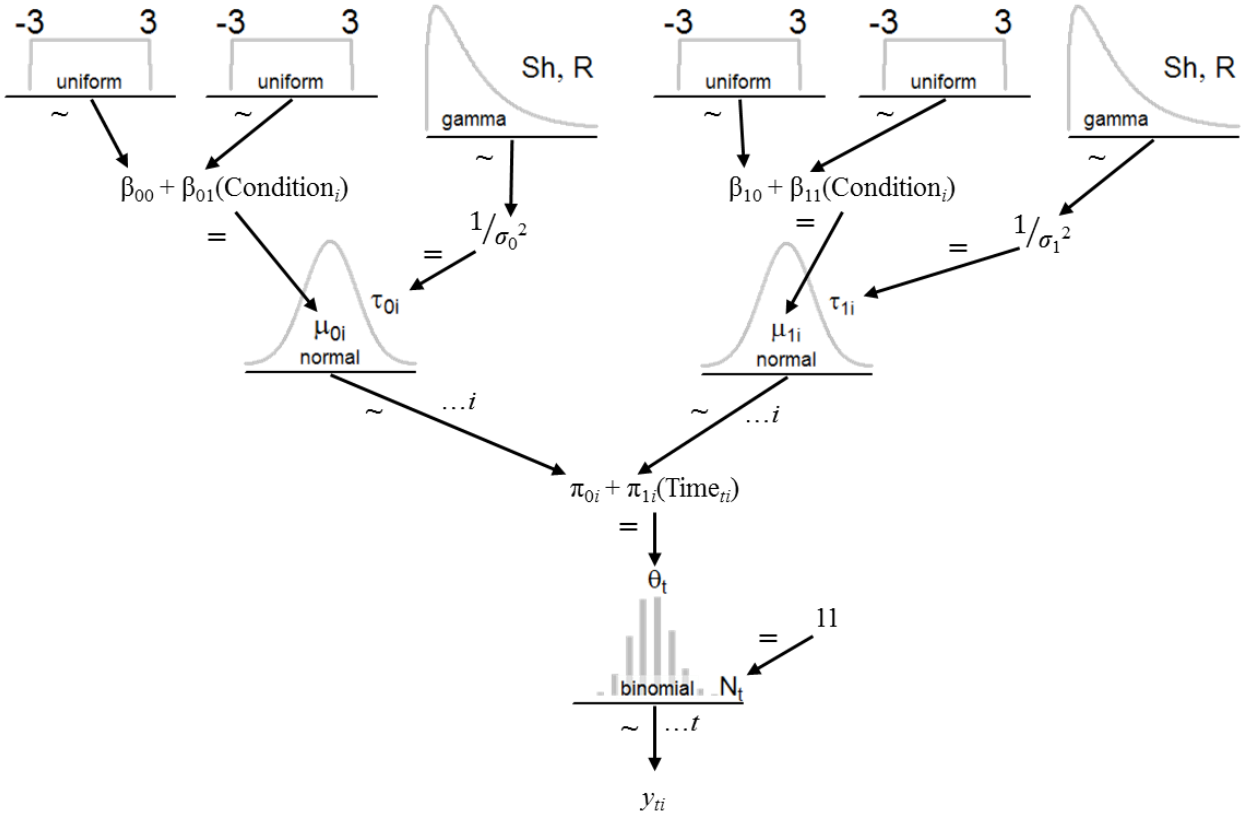


Figure A2. Multilevel hierarchical Bayesian regression model with binomial likelihood distribution. This model was used to examine declarative knowledge acquisition (Table 3 in manuscript). Note that all normal distributions were parameterized using the precision (τ) rather than variance, which is equivalent to 1/variance. The shape (Sh) and rate (R) parameters of the gamma distributions were parameterized according to their mode rather than mean; for clarity of presentation, the distributions over these parameters are not shown. However, the R code available in the Supplemental Materials provides the full specification of these parameterizations.

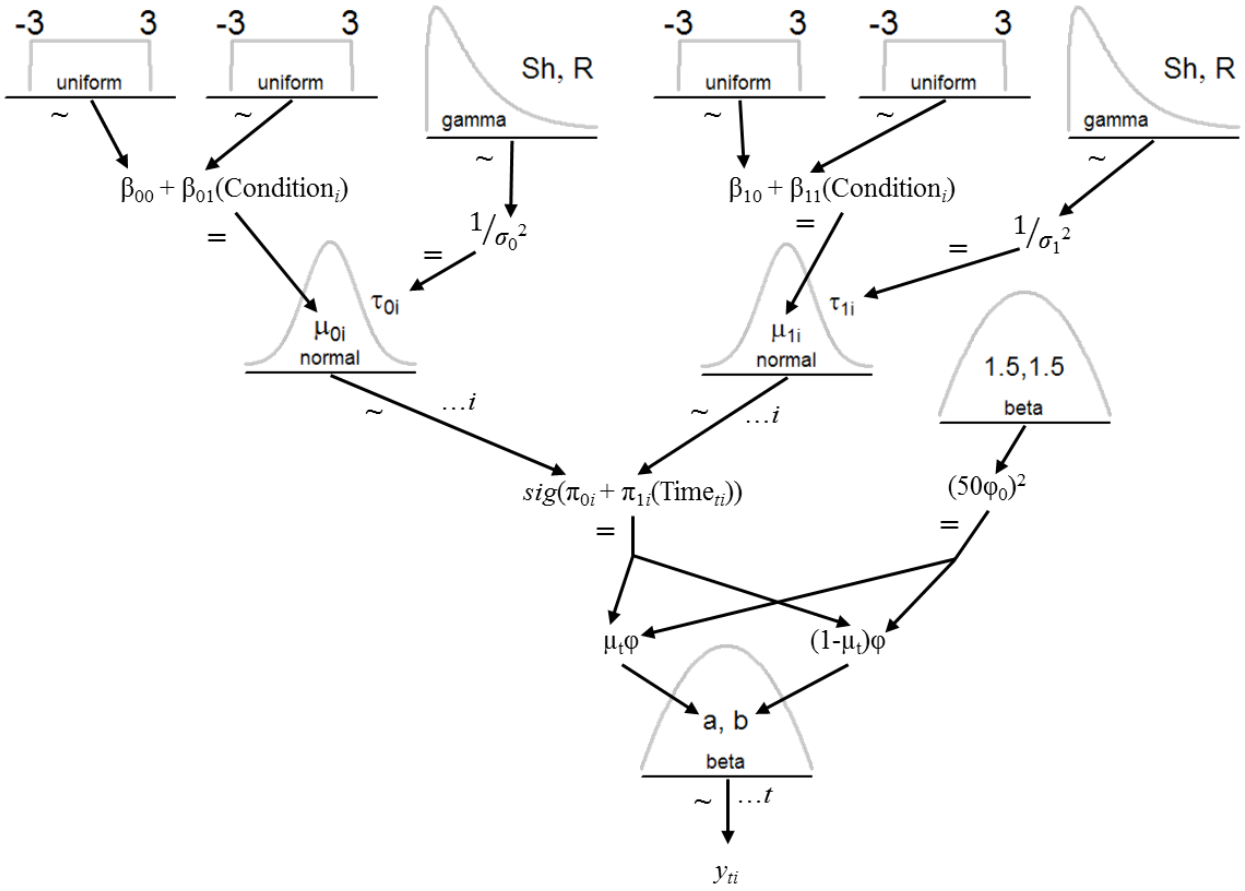


Figure A3. Multilevel hierarchical Bayesian regression model with beta likelihood distribution. This model was used to examine time spent studying how to process contacts in the operations manual (Table 5 in manuscript). Note that all normal distributions were parameterized using the precision (τ) rather than variance, which is equivalent to 1/variance. The *sig* function on the Level-1 regression denotes use of the inverse-logit linking function. The shape (Sh) and rate (R) parameters of the gamma distributions were parameterized according to their mode rather than mean; for clarity of presentation, the distributions over these parameters are not shown. However, the R code available in the Supplemental Materials provides the full specification of these parameterizations.

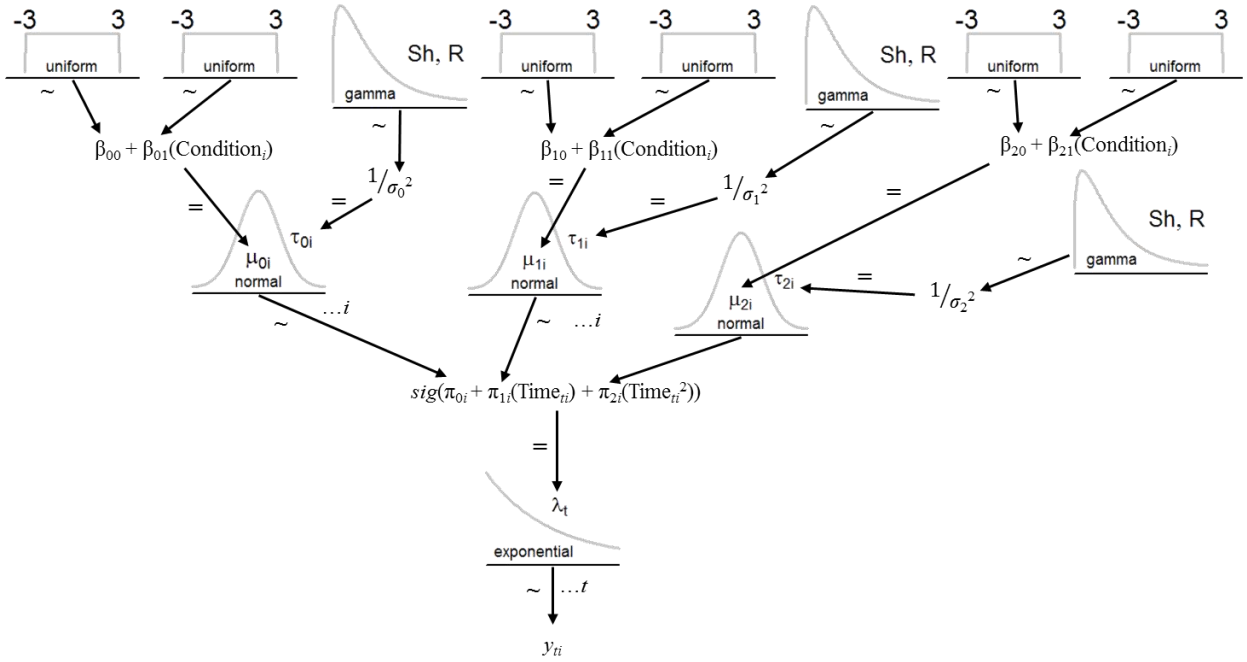


Figure A4. Multilevel hierarchical Bayesian regression model with exponential likelihood distribution. This model was used to examine time spent studying task strategies in the operations manual (Table 5 in manuscript). Note that all normal distributions were parameterized using the precision (τ) rather than variance, which is equivalent to 1/variance. The *sig* function on the Level-1 regression denotes use of the inverse-logit linking function. The shape (Sh) and rate (R) parameters of the gamma distributions were parameterized according to their mode rather than mean; for clarity of presentation, the distributions over these parameters are not shown. However, the R code available in the Supplemental Materials provides the full specification of these parameterizations.

APPENDIX B
SPECIFICATIONS AND DESCRIPTION OF COMPUTATIONAL MODEL

Table 7 in the text summarizes the steps of the computational model; here, a more detailed description of the model specifications and operationalizations are provided. To begin, a set of employees for every organization was created (Step 2). Each employee was given an initial performance potential by sampling a random number from a normal distribution (mean = 50, SD = 10). This value represented the human resource capital possessed by an individual when they were first selected for entry prior to any learning/training within the organization. Next, each employee was assigned a unique learning curve that determined how much their performance potential changed as a result of learning/training. Well over a century of research on learning has shown that the effects of learning/training on performance tends to follow a pattern of diminishing returns over time adhering to a power law function (e.g., Anderson, 2002):

$$y_{it} = a_i t^{b_i}, \quad \text{(A2)}$$

in which y_{it} represents the change in performance potential for individual *i* at time point *t*, a_i reflects the initial change in performance potential for individual *i* following the first time period (i.e., intercept at time *t* = 1), and b_i equals the rate of change in performance potential as a result of learning/training for individual *i*. The *b* coefficients for all agents were always negative, reflecting that improvements in performance potential as a result of training tended to decay over time. Consequently, this parameterization reflects that agent's performance potential did not grow linearly ad infinitum, but eventually reached an asymptote corresponding to its maximum performance potential.

Each agent received a unique *a* parameter by multiplying its initial performance potential by a randomly sampled percentage. This made performance potential improvement positively correlated with an individual's overall capabilities (i.e., more capable employees tend to learn more than less capable employees, e.g., Blume et al., 2010). The sampled value for this percentage was determined by selecting a random value from a Beta(12,8) distribution for each agent. Beta distributions are defined on the interval [0,1] and are commonly used for representing distributions of probability densities (Kruschke, 2015). The Beta(12,8) distribution used in this simulation was selected so as to approximate a normal distribution with mean = .6 and SD = .11 and reflect that simulated employees increase their performance potential by an amount close to 60% of their initial performance potential at the first time period as a result of job-relevant learning/training. The *b* parameter in Equation A2 controlled the rate at which an agent's performance potential changed over time as a result of learning/training; more positive values for *b* correspond with smaller rates of decay in performance potential over time, while more negative values reflect greater rates of decay in performance potential over time. The results from the empirical study revealed that the rate at which both task

performance and knowledge acquisition changed over time was generally slower for individuals experiencing ST relative to those not experiencing ST (e.g., Figures 5 and 6). As such, a similar pattern should be reflected in the parameterization of the simulation as well. To do so, the rate parameter for agents not experiencing ST was randomly sampled from a normal distribution with mean = -1.5 and SD = .1 (all $b < 0$). This distribution reflected that the rate at which the performance potential for a simulated employee not experiencing ST should change as a result of learning/training was expected to decrease by approximately 1.5 each time period. The *b* parameter for agents experiencing ST was also randomly sampled from a normal distribution with SD = .1, though the mean of this distribution was weighted proportionately to the mean of the *b* distribution used for agents in organizations without ST. More specifically, the mean for the *b* parameter in the learning curves of agents with ST was calculated as:

$$b_{ST} = \theta b_{NoST}, \tag{A3}$$

where b_{ST} is the mean of the sampling distribution for the *b* parameter of agents experiencing ST, b_{NoST} is the mean of the sampling distribution for the *b* parameter of agents not experiencing ST, and θ is a free parameter in the interval [0, ∞] reflecting the relative difference in rates of change in performance potential between agents not experiencing compared to those experiencing ST. A value of θ = 1 indicates no difference in rates of performance potential change between agents experiencing versus not experiencing ST. Alternatively and because the values for *b* were always negative, θ > 1 indicates a faster rate of decay for agents experiencing ST (i.e., larger negative *b* value) and thus a poorer performance potential trajectory.

The actual values selected for the *a* and *b* parameters in the simulation will certainly influence the rates at which agents' and organization's performance potential change; it would thus be most desirable to base these parameter values on existing empirical observations. However, when such data is not available or not well understood—as in the case of the present model—values should be assigned that are both plausible and permit evaluation of the phenomena (Harrison et al., 2007). Given that the purpose of the present model was to demonstrate the potential impact of small ST effects during learning/training in an organizational context, it was most important to ensure that the *relative* magnitude and parameterization of the *a* and *b* values were consistent with the overall pattern of results observed in the empirical data (i.e., agents experiencing ST should have a slower rate of performance improvement due to knowledge acquisition that agents not experiencing ST). Thus, although different values for the coefficients change how quickly employees and organizations reach their performance potential in the simulations, the actual values selected will not greatly change the overall pattern of results or substantive

conclusions. Interested readers are encouraged to test this conclusion by evaluating the model using different *a* and *b* parameters using the model code available in the Supplemental Materials.

Once the workforces were constructed, voluntary turnover schedules were created for each organization in Step 3. The creation of voluntary turnover schedules within each simulated company followed the same procedure outlined by Scullen et al. (2005, p. 9) and conceptualized voluntary turnover as an arrival/departure process that follows a periodic rate. Such phenomena are characteristic of a Poisson process, and thus the Poisson distribution can be used to estimate the number of events expected to occur within a discrete interval given a particular rate of occurrence. Every organization received a unique turnover schedule indicating the number of employees leaving the company at each time period by randomly sampling from a Poisson distribution with a rate parameter equal to a specified voluntary turnover rate. Note that although each company followed its own turnover schedule, the long-run average number of employees leaving any given organization equals the voluntary turnover rate.

Steps 4 through 7 of the pseudocode in Table 7 constitute the fundamental growth process for employee and organizational performance potential modeled in the simulation. In Step 5, the amount by which each simulated employee's performance potential increased as a result of learning/training at a given time period was computed according to Equation A2. Note that the operationalization of this model is such that *all* agents in an organization with ST experienced a ST-induced decrement to their performance potential increase. This assumption is equivalent to an empirical study that *only* examines differences in performance potential among employees who are expected to be influenced by ST (i.e., females/groups facing a negative stereotype). Modeling individuals not influenced by ST (i.e., males/groups not facing a negative stereotype) was unnecessary in the present simulation as the performance potential of these simulated employees would not be expected to differ across conditions of ST.

Step 6 recorded the cumulative performance potential for simulated employees and organizations. At the employee-level, this computation was determined as:

$$PP_{it} = PP_{i(t-1)} + y_{it}, \tag{A4}$$

where PP_{it} is the cumulative performance potential of employee *i* at time *t*, and y_{it} is given by Equation A2. Following the computation of Equation A4 for all agents, each organization's performance potential was recorded by computing the average of its employees' cumulative performance potentials.

The final stage of note in the simulation occurs in Step 7 when employees departed from an organization through turnover. As in Scullen et al. (2005), turnover was enacted by randomly selecting employees from each organization in accordance with its voluntary turnover schedule and removing those members from the company. All

departed employees were immediately replaced by a new hire following the same procedure and parameterization as described for Step 2. If the goal of the simulation was to develop a more fully comprehensive model of human resource management, it would be desirable to model alternative turnover (i.e., turnover correlated with performance potential, tenure, etc.) and hiring (i.e., hire employees with higher potential to replace more experienced departing employees, introduce ST into selection) processes. However, for purposes of the present simulation, such mechanisms simply add "moving parts" to the model that do not contribute to its intended objective—demonstrating the impact that small-to-moderate ST effects experienced during training/learning can exert on an organization's human capital resources. Steps 4 through 7 of the simulation were repeated until the time limit (t_{stop}) was reached. In all reported simulations, t_{stop} = 30 to allow adequate opportunity for the modeled processes to unfold and for any potential asymptotic effects to be observed.