

INVESTIGATING CHOKING UNDER PRESSURE IN NOVICE PARTICIPANTS
ACROSS DIGITAL AND LIVE PLATFORMS

By

Daisuke S. Katsumata

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology – Master of Arts

2023

ABSTRACT

The phenomenon of individuals underperforming relative to their typical skill level under stressful situations, or performance pressure, is colloquially known as “choking under pressure.” This project aims to investigate two major questions: (a) can a performance pressure paradigm that has proven successful in past laboratory investigations be replicated in an online, self-administered (OSA) environment? and (b) are there reliable individual differences in choking, and if so, do they correlate with psychological traits? These questions are investigated in two experiments in the cognitive and psychomotor domains by training novice participants in modular arithmetic (MA) and golf putting tasks, respectively. Performance pressure was manipulated in participants via a combination of monetary incentives and mock recording performances. In both experiments, the pressure manipulations failed to induce expected choking effects, which were based on the following: (a) changes in performance from baseline to pressure trials, (b) whether participants felt increased perceived pressure, and (c) how strongly participants believed the pressure manipulations. Participants from the in-person/golf putting study even saw improvements in performance from baseline to pressure trials. The reliability estimates for difference-based choking measures were low for both OSA/MA and in-person/golf putting, except for response time (RT) for MA. Correlations were found between the choking measure and some individual difference measures, but this should be interpreted with caution. There was also evidence of a speed-accuracy trade-off in MA, as accuracy and RT were negatively correlated.

TABLE OF CONTENTS

INTRODUCTION	1
STUDY 1	16
Method	16
Results	30
STUDY 2	44
Method	44
Results	53
GENERAL DISCUSSION	63
FOOTNOTES	76
REFERENCES	77
APPENDIX A: Modular Arithmetic Task Instructions (OSA)	81
APPENDIX B: Golf Putting Task Instructions (In-Person)	94

INTRODUCTION

We are sometimes unable to perform our best when it matters most. In these situations, factors such as incentives for success, a high level of expertise, and a desire to perform well seem to do little to save us as we succumb to the pressure. This phenomenon is colloquially known as “choking under pressure.” The common fear of public speaking is a prime example of this; many of us suddenly find it difficult to speak—a well-practiced skill—when confronted with an audience. In such situations, even being familiar with the topic of discussion may seem to do little to help us as we desperately try to find the words.

Definitions of Choking Under Pressure

Choking under pressure has been defined as “heightened levels of perceived pressure accompanied by a suboptimal performance level” (Beilock & Gray, 2012, p. 426). This notion of choking as *any* (statistically significant) inferior performance has been noted in a seminal review by Baumeister and Showers (1986) and experimental studies operating on this definition have contributed greatly to our understanding of choking as a phenomenon (Beilock, 2007). More recently, researchers have argued that a key characteristic of choking is a *significant* degree of performance decrements rather than *any* (statistically significant) performance decrement, separating it from less severe underperformances (Hill et al., 2010). Researchers have further proposed for expanding the definition of choking under pressure to include the phenomenological experience of the choker, stipulating that an athlete should label their own experience as a choke (Hill et al., 2009). Proponents assert that choking should be defined more concisely to capture how it is experienced in the real-world and failing to do so may lead to overlooking potential differences in cognitive processing among differing levels of underperformances (Mesagno & Hill, 2013; Hill et al., 2017; Mesagno et al., 2015). Precisely

defining what constitutes choking under pressure has not been without debate (cf. Mesagno & Hill, 2013; Jackson, 2013).

Pressure Manipulations in Choking Experiments

In a typical study, performance on a task in a baseline condition is contrasted with performance on the same task in a condition designed to create performance pressure. For example, in a study of modular arithmetic problems by Beilock et al. (2004), pressure was manipulated via a monetary incentive that was contingent on both members of randomly paired participants improving their performances. Each participant was informed that their partner had already met the improvement criterion, so it was now up to them to earn (or lose) the reward for the pair. Participants were additionally videotaped during the pressure manipulation while being told that their performance would later be analyzed by local math experts. Wang and Shah (2013) similarly manipulated pressure in a study involving mental arithmetic problems in Chinese third and fourth graders. During high-pressure trials, the first author stood behind the participants as they performed the task while also reminding them that the current trials were the “real” test and that previous trials were simply practice. Wang and Shah also videotaped participants under the pretense that their performance would later be evaluated by experts. As noted, many studies tend to employ a combination approach for pressure manipulations to mimic real-world conditions and to ensure the greatest chance of choking in their participants, and this tends to be true for choking studies that are focused on understanding the consequences of choking rather than the mechanisms of choking per se.

In contrast, other studies may seek to test the effectiveness of various pressure manipulations in conjunction with different tasks, skill levels, etc. to better understand the conditions under which choking occur and test various models that explain choking (DeCaro et

al., 2011; Mesagno et al., 2011). Notably, Mesagno et al. (2011) found that in an experimental study involving experienced field hockey players, self-presentation pressure manipulations (i.e., an audience of their peers and video recording to be analyzed later by their coaches) that threatened the players' identities as "athletes who do not choke under pressure" led to choking but motivational pressure manipulations (i.e., a money incentive) increased performance under pressure. A combined manipulation involving both self-presentation and motivational manipulations did not lead to choking. Mediation analysis revealed cognitive anxiety (i.e., how one may interpret the situation at hand via negative perceptions, expectations, worries, etc.) was a significant predictor of poor performance due to the self-presentation pressure manipulation, but somatic anxiety (i.e., one's awareness of physiological arousal related to the situation) was not. These findings suggest that researchers should be careful in selecting pressure manipulations according to research aims and be wary that with regards to pressure manipulations, more is not always better and may even lead to interference. This project utilized a combined approach to pressure manipulations, based on the successful track record in previous studies, particularly in those that used similar tasks (Beilock & Carr, 2001; Beilock et al., 2004; DeCaro et al., 2011).

Current Project

The operational definition of choking for this project was based on Beilock and Gray's (2012) definition. In this study, participants were classified as choking under pressure only if they perceived pressure prior to a performance decrement, but the degree of performance decrement required to classify the performance as a choke was not specified. Therefore, choking was not defined purely based on performance to avoid some of the potential pitfalls described above (see Definitions of Choking Under Pressure). Ideally, successful pressure manipulations should create psychologically realistic situations that compel participants to feel pressure, in turn

hindering their performance to a severe enough degree that they perform significantly worse than during baseline. However, it is possible for the pressure manipulations to fail despite researchers' efforts, meaning participants do not feel performance pressure and in turn their performance may stay the same or even improve. There is also an alternative possibility in the event of a lack of performance decrements; the pressure manipulations may succeed in making participants feel pressure, but participants may in turn succeed in resisting that pressure in their task performance (i.e., coming in "clutch").

This project involves two studies that strive to expand upon the work of previous research and develop an additional means for scientific inquiry into the phenomenon of choking under pressure. Foremost, the project aims to establish effective choking manipulations across an online, self-administered (OSA) research platform and an in-person setting among novice participants performing novel tasks. A secondary goal for this project is to examine choking under pressure as an individual-difference construct by assessing the reliability of difference-based choke scores as a psychometric measure, and whether differences in these choke scores are correlated with various psychological traits that have previously been linked to choking under pressure. These research goals are addressed by both the modular arithmetic task, adapted from Beilock et al. (2004), representing the cognitive domain and administered in an online format and in the golf putting task, adapted from Beilock and Carr (2001), representing the psychomotor domain and administered in person on a laboratory putting green. The research goals are intertwined, as part of establishing a new platform to test choking under pressure (i.e., Research Question 1) is to estimate the reliability of the choking measure derived from it (i.e., Research Question 2) to ensure that it is an effective new research platform. Research questions stemming from these goals are summarized in Table 1.

Table 1*Research Questions Addressed in Current Project*

Research Question 1: Research Platforms	Can a performance pressure paradigm that has proven successful in past laboratory investigations be replicated in an online, self-administered task environment?
Research Question 2: Individual Differences	Are there reliable individual differences in choking, and if so, do they correlate with psychological traits?

Note. An overview of the research questions this project addresses via two of studies consisting of an online, self-administered platform using a modular arithmetic task and an in-person platform using a golf putting task.

Research Question 1

To date, most choking experiments have been conducted in “live” settings in which participants interact with experimenters and pressure manipulations face-to-face (Mesagno et al., 2011; Beilock & Carr, 2001; Lewis & Linder, 1997). In contrast, no studies to my knowledge have used an OSA paradigm for choking experiments. There are advantages and disadvantages to each method.

In-person studies constitute the “default” model of choking experiments and have been utilized across a range of domains and tasks, including psychomotor tasks such as golf putting and cognitive tasks such as mental arithmetic (cf. Beilock & Carr, 2001; Beilock et al., 2004). Some tasks, particularly psychomotor tasks such as basketball free throws, field hockey penalty strokes, golf putting, etc., are most effective when performed in-person or cannot reasonably be self-administered (Wang et al., 2004; Mesagno et al., 2011; Beilock & Carr, 2001). In-person experiments may also allow for pressure manipulations, such as an audience, to be more salient. Being face-to-face with their participants may also allow researchers to record and quantify subtleties about their participants, such as the participants’ engagement level and even ensure that they are completing the task properly. Such nuances cannot be fully captured by self-report

measures employed by OSA studies nor can a researcher be on hand to ensure compliance and understanding of instructions. In some cases, participant engagement or lack thereof may be directly linked to exclusion criteria and even lead to new hypotheses in subsequent experiments. In-person studies are not without their drawbacks; they can be resource-intensive and time-consuming, ultimately resulting in opportunity sampling that is local to the researcher or consisting of smaller, less diverse samples.

OSA studies can be conducted similarly to in-person studies of cognitive tasks such as math problems (Beilock & Carr, 2001; DeCaro et al., 2011), as participants engage in the task via a computer. Thus, in principle, an OSA platform would not impact how the cognitive tasks per se are executed relative to previous literature. Presenting these tasks online also takes advantage of existing benefits of computerized studies, as data can be collected unobtrusively during the performance of the task itself (e.g., accuracy and response time data) and feedback can be immediate and consistent. In recent years, there has been an increase in the proliferation of high-speed internet in the United States due to the investments made in the wake of the lockdowns enacted in response to COVID-19 (Read & Wert, 2021). Additionally, software platforms designed to run experiments online have received increased attention, especially by researchers looking for new ways to conduct research after in-person research was halted in the wake of the COVID-19 pandemic. Due to these factors, data collection via experiments posted on a website has become an increasingly appealing way to recruit participants. The self-administered nature of an online study also allows for considerable advantages for data collection compared to in-person studies. Online studies potentially allow for data to be collected at a lower cost across larger samples, beyond university undergraduates. It is also easier to collect data from specialized or rare populations. Running a study online also leaves open the

possibility of long-term data collection with minimal monitoring by researchers. In the case of the Harvard Implicit Association Test, the experiment was made available online in the early 2000's and data collection is ongoing (Kraut et al., 2003). The digital nature of the OSA platform also makes it possible to rapidly test different tasks and manipulations. The automated process also ensures a consistent experience across all participants. The potential disadvantages faced by OSA studies are that it is still subject to sampling bias due to the digital divide and the sheer quantity of data that can be obtained may necessitate new or more elaborate techniques to clean and analyze the data. Researchers also have no explicit control over the digital environment in which participants complete the study, including software configurations, hardware compatibility, and network connection issues, which can potentially impact measurements such as reaction times. Fortunately, software developed for conducting experiments online have attempted to account for these potential differences, showing that for some software like Lab.js, latency times across different web browsers are comparable (Henninger et al., 2021). Additionally, issues may arise from participants completing the study while multitasking on the computer or being distracted in the physical environment that they are in, leading to more noise in the data. However, this problem is not insurmountable, as the substantial sample sizes made possible by OSA studies may ameliorate these concerns or even render them moot.

Regardless of the research platform, designing consistently effective pressure manipulations are a core aspect of any choking study (Beilock & Carr, 2001; DeCaro et al., 2011). A major goal of this project is to develop effective pressure manipulations that are comparable across OSA and in-person studies by reproducing choking effects regardless of the platform. Many in-person choking experiments have employed a combination approach to pressure manipulations, putting together elements of monetary incentives, peer pressure, and

social evaluation (e.g., Beilock & Carr, 2005). Many of these manipulations may seem straightforward to translate into an online environment. For example, instead of video recording participants with a physical camera during task performance, the screen can be recorded. A key characteristic of effectively *implemented* manipulations is that participants should become sufficiently engaged with the experiment so that they care that they are being video recorded for later analysis, and in turn perceive enough of a rise in pressure to then choke. Here, the distinction is that convincing participants to become invested with their own performance in the experiment is not just a property of the manipulations per se but is also a function of the platform it is conducted on. In this respect, the human interactions afforded by in-person studies allow experimenters to interact dynamically with participants. For example, experimenters can give instructions to participants in a conversational tone and use body language to signal the importance of trying their best, both of which help establish rapport and can ultimately help compel participants to be engaged and compliant with experimental procedures. In comparison, OSA studies must rely on relatively static means of conveying instructions such as text, pictures, and animations. While these methods can offer consistency unmatched by human experimenters, they may also come off as impersonal and dry, leaving it up to the participants to become invested in the experiment and accountable for their own performance. Of course, there are trade-offs. Even with standardized protocols, human experimenters can inadvertently deviate and introduce method variance, possibly even to the degree of undermining the original aims of the experiment. It remains to be seen what aspects of a human presence is required (e.g., if auditory instructions could be pre-recorded and replayed, or if a live interaction at some point in the experiment is necessary) for maximizing participant engagement from a research platform standpoint (or if it even is required). It is also unclear if these challenges are unique to choking

studies or if other fields also have concerns over their results in the face of such scrutiny. Nevertheless, it is an open question as to whether these modified manipulations work as effectively without the intangibles of a face-to-face interaction and successfully adapting the manipulations across platforms is a goal of this project.

Tasks based on modular arithmetic and golf putting have been successfully used to demonstrate and study choking effects in previous research in the cognitive and psychomotor domains, respectively (Beilock & Carr, 2005; Beilock & DeCaro, 2007; Beilock & Carr, 2001). The track record of these tasks in experimental choking studies makes them ideal for the purposes of the current project. These tasks were also chosen because they were feasible to test in their respective settings (OSA for modular arithmetic and available physical lab space for golf putting). Differences in the tasks mean that they each require task-specific skills and measures of individual differences. These differences can be beneficial in offering a robust test for estimating the reliability of choking measures across a range of domains, tasks, and platforms.

Research Question 2

Within the choking under pressure literature, there has been a relative lack of attention paid to examining choking under pressure as an individual-difference construct, including research to investigate the psychometric properties of the phenomenon (Beilock & Gray, 2012). The focus of the second research question is to develop a procedure to reliably measure individual differences in choking under pressure as broadly defined by any decrements in performance under pressure compared to baseline. The most basic requirement for the measurement of individual differences, reliability refers to the consistency of the scores an individual may attain on a given test across various occasions, testing conditions, and with equivalent sets of items (Anastasi & Urbina, 1997). To investigate sources of individual

differences in choking under pressure, the reliability of the effect must first be established, because the reliability of a measure limits the degree to which it can correlate with any other measure. Past studies have found variability in choking effects across individuals (Beilock & Gray, 2012). However, reliability has not been formally estimated and thus it is unclear as to what extent this variability is systematic across individuals or reflects random error.

To estimate reliability, two sets of analyses were performed. Firstly, the effects of the experimental manipulations were measured using a standard procedure in which participants perform the task under a baseline condition and again on designated trials under experimentally induced pressure. Performance was measured based on mean accuracy (correct or incorrect) and mean response times (RTs) on correct trials for modular arithmetic and mean accuracy (cm to target) for golf putting, and then compared between conditions to obtain a difference score. In each task, participants repeated this process multiple times over the course of the experiments to provide multiple measures of the choking effect. These measures were combined to provide a single, overall measure of a difference score between baseline performance and pressure performance. Estimates of reliability were computed for the choking measure based on these overall means in the baseline and pressure conditions for each individual participant. Secondly, the reliability of the choke effect was estimated at multiple points in the experiment by keeping the multiple measures of choking separate and computing reliability estimates for each of them. Given that this project recruited and trained novice participants in novel tasks, it is possible and even expected that participants would initially improve in performance regardless of pressure early in the study but choke later as their skill matures (see information on skill level below). Therefore, reliability was estimated on a per-block basis (i.e., after each set of baseline

and pressure trials) to account for potential differences in the choking effect over the course of the experiments (which included multiple sets of baseline and pressure trials).

Controlling the skill level of participants for choking studies is a critical matter. In studying expert performance under pressure, recruiting experts (Mesagno et al., 2011; Masaki et al., 2017) ensures that participants come ready at a level of skill far greater than could be reasonably attained in a short-term experiment (e.g., sports-related tasks). Studies that utilize novel or unusual tasks may recruit novices (Beilock & DeCaro, 2007), while other studies may recruit novices to control prior knowledge and experience with a task and the conditions under which they gain expertise (Beilock & Carr, 2001; Beilock et al., 2004; Oudejans & Pijpers, 2009).

The decision to recruit participants at a particular skill level for the task at hand is important, as there are critical expert-novice differences in choking studies. Skill execution, particularly with respect to attentional mechanisms, seems to change as novices become experts (Beilock, 2007). In general, novices tend to choke because of distractions as they need to process the task via working memory and are unable to effectively filter out further cognitive demands (Hill et al., 2010). Evidence suggests that performance pressure that is not distraction-based but encourages greater attention to the execution of the task may lead to better performance in relative novices or in unfamiliar tasks (Baumeister, 1984; Beilock & Carr, 2001). Experts, however, tend to choke due to self-focus when they process their well-practiced skill via working memory when they typically no longer do so; the same pressure manipulations that would cause novices to perform better can lead to a choke for experts. Consequently, skill level can act as a moderator for choking under pressure in some cases (Hill et al., 2010). Given these details, participants need to be trained to a sufficient level to respond to the chosen pressure

manipulation(s) or the pressure manipulation(s) themselves need to be properly matched with the participant pool.

However, care must be taken in the *training* of novice participants. Training while under performance pressure may subsequently lead to better performance under pressure (Beilock & Carr, 2001; Gröpel & Mesagno, 2017; Lewis & Linder, 1997; Oudejans & Pijpers, 2009; Oudejans & Pijpers, 2010). For example, a study investigating effective training practices against choking under pressure found that training participants under mild anxiety conditions helped them perform better under higher anxiety conditions in a novel dart throwing task (Oudejans & Pijpers, 2009). The challenge then, is that inexperienced participants must be carefully trained at the beginning of the study in a way that sufficiently allows them to acquire the skill level necessary to choke, but in the process, avoids inoculating them against the performance pressure manipulations that are to follow. Otherwise, the choking study may fail to find choking effects. In this project, ensuring that the learning process occurred over the course of the study provided an opportunity to observe the skill acquisition trajectory during and after training. This was due to the study design in which participants completed multiple blocks of practice and evaluation trials.

An additional goal of the research question is to investigate psychological correlates of individual differences in choking. There are at least two personality characteristics that have been shown to correlate with the propensity to choke under pressure (Beilock & Gray, 2012; see also Masters et al., 1993; Omoregie & Adegbesan, 2011). The first is *dispositional self-consciousness*: a person's awareness of their own internal psychological states and processes (Baumeister, 1984; Fenigstein et al., 1975). As previously mentioned, excessive self-focus on a task could be detrimental to its performance depending on skill level, so it may not come as a

surprise that an associated personality trait (i.e., self-consciousness) can impact choking under pressure in a similar way. According to *explicit monitoring theory* (EMT), performance pressure increases self-consciousness, which causes the performer to direct their attention towards the execution of the skill to compensate (Baumeister, 1984). In experts, this shift in attention can be detrimental for well-practiced, automated skills, reverting them to a novice-like state of deliberate step-by-step execution. In contrast, this may be helpful for novices, allowing them to focus their attention on learning the task (Beilock & Carr, 2001). Despite seemingly contradictory results, research on the link between self-consciousness and performance pressure in the psychomotor domain support this pattern (cf. Baumeister, 1984; Wang et al., 2004). Baumeister found that those high in self-consciousness performed better when performing an unfamiliar ball roll-up motor task, whereas Wang et al. found that those high in self-consciousness performed worse in a well-practiced basketball free-throw shooting task. Beilock and Gray (2012) have proposed that in these studies, self-consciousness may have helped the relative novices who were attempting an unfamiliar task, as attending to the execution of a novel task is common and helpful, whereas the same tendency to focus inwardly for well-practiced tasks may have been harmful for the experienced athletes.

The second personality characteristic is *trait anxiety*: the relatively stable disposition of an individual to interpret and report various situations as potentially threatening or negative (Mascarenhas and Smith, 2011). While state anxiety measures have been commonly used in many choking studies, state and trait anxiety measures are closely related (Mesagno et al., 2011). Furthermore, trait anxiety measures may provide a better basis for understanding individual differences. Research has shown high trait anxiety may be detrimental to performance in contexts ranging from academic test taking (Eysenck, 1992) to sports (Wang et al., 2004). Wang

et al. found that trait anxiety was a significant predictor of choking. Furthermore, they found that athletes who were high in both trait anxiety (assessed using the Sport Anxiety Scale-2; Smith et al., 2006) and self-consciousness performed even worse.

Based on past findings, both measures were expected to correlate significantly with choking under pressure. For example, Wang et al. (2004) analyzed self-consciousness and trait anxiety by comparing sub-scales within the overall measures. They found that the private self-consciousness sub-scale of the self-consciousness measure and somatic anxiety sub-scale of the trait anxiety measure were key predictors in choking under pressure.

Working memory capacity is another important individual difference variable to consider for the modular arithmetic task. Beilock and Carr (2005) paradoxically linked choking while solving hard modular arithmetic problems with high working memory but did not find the same association with low working memory. In the study, participants mentally solved modular arithmetic problems that ranged in working memory demands from low to high, depending on whether larger numbers and carry-over operations were involved. For accuracy, all participants maintained their level of accuracy regardless of pressure on the low-demand math problems. However, for high-demand problems, while high working memory participants were more accurate than low working memory participants while under low pressure, this advantage disappeared under high pressure. Analyzing response times for correct problems yielded three overall results: high working memory participants were faster than low working memory participants, response times were slower for high-demand problems than for low-demand problems, and all participants were slower under low pressure than high pressure. Working memory is correlated with superior performance on mentally demanding tasks (Cowan et al., 2005), perhaps because it allows for resource-intensive mechanisms or strategies. The authors

postulated that working memory is also important for dealing with pressure, as pressure can cause anxiety and worry, which in turn taxes working memory as participants attempt to stay on task. Therefore, high working memory participants may have tapped into their working memory to both deal with the pressure and to solve the high-demand math problems, but because these demands exceeded their capacity, ended up choking. These findings suggest that choking can only occur if both the participants are skilled enough and the problems are complex enough. It is also possible that experimentally producing a choking effect may be more nuanced than simply crafting all-purpose pressure manipulations that are effective regardless of individual differences.

STUDY 1

Study 1 consisted of the modular arithmetic (MA) task over an online, self-administered (OSA) platform. A major goal of the study was to establish working pressure manipulations over the OSA platform using a task (MA) that has had a successful track record in prior choking studies. Alongside this, calculating a reliability estimate of the choking measure and analyzing whether that choking measure would correlate with individual difference variables with successful track records was also important. Establishing the reliability of the choking measure itself is critical for it to correlate with other measures and doing so by utilizing a previously successful choking task and correlated measures was seen as a natural extension of the literature.

Method

Participants

Institutional review board approval was obtained before experimentation began. The study was completed by 180 undergraduate students over the age of 18 who were recruited through an online research recruitment service and were enrolled in undergraduate psychology courses in a large Midwestern university during the spring of 2021. After applying exclusion criteria described in the Results section below, data from a total of 144 participants were included in the analyses.

Demographics. After providing their consent, participants answered demographics-related questions. Participants included 47 men (32.6%), 92 women (63.9%), and five (3.5%) who declined to answer. Participants were asked about their prior math experience, as it may impact their performance on the modular arithmetic task. Most participants (132, or 91.7%) responded they were “never a math major,” several (6, or 4.9%) responded that they “were a math major at one point but not anymore,” and a similar number (5, or 3.5%) declined to

respond. When asked about their prior modular arithmetic(MA) experience and provided with an example, 109 participants (75.7%) responded that they were not familiar with MA, a portion of participants (31, or 21.7%) responded that they had some prior experience with MA but did not remember how to solve the example problem, and three (2.1%) responded that they had prior experience with MA and did know how to solve an example problem that was provided, and one participant (0.7%) declined to answer. Participants were not excluded on account of prior experience with MA problems. Participants were also asked about their vision. Responses indicated that 136 participants (94.4%) were not color-blind, six (4.2%) were color blind, and two participants (1.4%) declined to answer. 122 participants (84.7%) reported corrected-to-normal vision, 21 (14.6%) reported that they did not have corrected-to-normal vision, and one participant (0.7%) declined to answer.

Materials

Measures. Multiple measures were used to assess the psychological correlates of choking under pressure.

Self-consciousness was assessed using the Self-Consciousness Scale (SCS) (Fenigstein et al., 1975), which includes 23 items (e.g., *I'm concerned about what other people think about me.*). The scale is composed of a 10-item private self-consciousness sub-scale, a 7-item public self-consciousness sub-scale, and a 6-item social anxiety sub-scale (see Table 2). SCS scores are calculated by summing the individual items.

Table 2

Self-Consciousness Scale (Fenigstein et al., 1975)

	Extremely uncharacteristic				Extremely characteristic
Private self-consciousness					

Table 2 (cont'd)

I'm always trying to figure myself out.	0	1	2	3	4
Generally, I'm not very aware of myself.	0	1	2	3	4
I reflect about myself a lot.	0	1	2	3	4
I'm often the subject of my own fantasies.	0	1	2	3	4
I never scrutinize myself.	0	1	2	3	4
I'm generally attentive to my inner feelings.	0	1	2	3	4
I'm constantly examining my motives.	0	1	2	3	4
I sometimes have the feeling that I'm off somewhere watching myself.	0	1	2	3	4
I'm alert to changes in my mood.	0	1	2	3	4
I'm aware of the way my mind works when I work through a problem.	0	1	2	3	4
Public self-consciousness					
I'm concerned about my style of doing things.	0	1	2	3	4
I'm concerned about the way I present myself.	0	1	2	3	4
I'm self-conscious about the way I look.	0	1	2	3	4
I usually worry about making a good impression.	0	1	2	3	4
One of the last things I do before I leave my house is look in the mirror.	0	1	2	3	4
I'm concerned about what other people think about me.	0	1	2	3	4
I'm usually aware of my appearance.	0	1	2	3	4
Social anxiety					

Table 2 (cont'd)

It takes me time to overcome my shyness in new situations.	0	1	2	3	4
I have trouble working when someone is watching me.	0	1	2	3	4
I get embarrassed very easily.	0	1	2	3	4
I don't find it hard to talk to strangers.	0	1	2	3	4
I feel anxious when I speak in front of a group.	0	1	2	3	4
Large groups make me nervous.	0	1	2	3	4

Note. Participants were asked about their dispositional self-consciousness to understand its connection with individual propensity for performance decrements under pressure.

Trait anxiety was assessed using the Trait Anxiety subscale of State-Trait Anxiety Inventory for Adults (STAI-AD) Form Y (Spielberger, 1983), which includes 20 items (e.g., *I am a steady person.*) on a 4-point scale (see Table 3). STAI scores are calculated by summing the individual items. Scores may range from a minimum of 20 to a maximum of 80. The commonly accepted classifications of STAI scores are “no or low anxiety” for 20-37, “moderate anxiety” for 38-44, and “high anxiety” for 45-80 (Kayikcioglu et al., 2017).

Table 3

State-Trait Anxiety Inventory for Adults Form Y-2 (Spielberger, 1983)

	ALMOST NEVER	SOMETIMES	OFTEN	ALMOST ALWAYS
I feel pleasant	1	2	3	4
I feel nervous and restless	1	2	3	4

Table 3 (cont'd)

I am satisfied with myself	1	2	3	4
I wish I could be as happy as others seem to be	1	2	3	4
I feel like a failure	1	2	3	4
I feel rested	1	2	3	4
I am "calm, cool, and collected"	1	2	3	4
I feel that difficulties are piling up so that I cannot overcome them	1	2	3	4
I worry too much over something that really doesn't matter	1	2	3	4
I am happy	1	2	3	4
I have disturbing thoughts	1	2	3	4
I lack self-confidence	1	2	3	4
I feel secure	1	2	3	4
I make decisions easily	1	2	3	4
I feel inadequate	1	2	3	4
I am content	1	2	3	4
Some unimportant thought runs through my mind and bothers me	1	2	3	4
I take disappointments so keenly that I can't put them out of my mind	1	2	3	4
I am a steady person	1	2	3	4
I get in a state of tension or turmoil as I think over my recent concerns and interests	1	2	3	4

Note. Participants were asked about their trait anxiety to understand its connection with individual propensity for performance decrements under pressure.

Working memory capacity was assessed using a version of the backwards digit span (BDS) (Case & Globerson, 1974), the reverse digit span (RDS). The basic concept of RDS is the same as BDS, as participants must remember an ordered sequence of digits, mentally reverse it, and enter the reversed sequence by typing or clicking numbers on the keyboard or mouse. The version used by Vock and Holling (2008) was implemented. Specifically, RDS consisted of showing participants spans of digits ranging from 2 to 9 for 1.5 seconds each. There were two unique sets of digits shown for each span length, for a total of 16 sets. The sets were always presented in ascending order, meaning participants saw sets of progressively longer spans (i.e., 2, 2, 3, 3, 4, 4...) but the digits within the spans were randomized for each participant. RDS performance ranged from 0-16 and was based on the summed total of spans that were correctly answered out of 16. For example, a participant could score a 12 on the RDS if they got 12 spans correct, regardless of which length spans were answered correctly.

Two difference-based choke scores were calculated on a participant-by-participant basis for task performance over the course of the entire experiment: one based on accuracy and one based on RT during correctly solved problems. The accuracy choke score was calculated by comparing percentage accuracy during baseline trials with pressure trials. The same steps were repeated on a per-block basis to calculate four additional difference-based choke scores. These choke scores compared baseline and pressure performance within each block (each block consisted of baseline trials that were immediately followed by pressure trials). Negative scores indicated that the participant potentially choked (i.e., did better during baseline trials than during pressure trials). Positive scores indicated either a “clutch” if participants perceived pressure, thus presumably overcoming that pressure to overperform relative to baseline, or a failure of the

pressure manipulations to induce performance decrements if participants did not perceive pressure.

Task. Modular arithmetic (MA) problems have been successfully used in prior choking studies (Beilock & Carr, 2005) and was used as the task for the current study. MA problems take the form of, “ $47 \equiv 24 \pmod{3}$ ” and are solved by judging whether the statement is true.

Participants were instructed to solve it by first subtracting the middle number from the first number (i.e., $47 - 24$), followed by dividing by the last number (i.e., $23 \div 3$). If the resulting number is a whole number, then the problem is considered true; in this case, the number is not a whole number and is therefore considered false. The MA task is a preferred laboratory task because solving it this way only requires basic arithmetic and is therefore highly accessible to most participants, yet the operations involved are an uncommon combination so that even those well-versed in math may not have experience on the task.

As noted, Beilock and Carr (2005) found that, with respect to accuracy, only MA problems that had high demand on working memory led to choking, and only among those high in working memory. MA problems with low demand on working memory did not lead to choking among all participants and there were no differences between pressure conditions with respect to reactions times. High-demand problems were defined by whether the first step had large numbers (>20) or required a borrow operation. These attributes necessitate a longer sequence of mental calculations and more intermediate products to hold in working memory, thereby placing greater demand upon it (Ashcraft, 1992; Ashcraft & Kirk, 2001). For example, “ $7 \equiv 4 \pmod{2}$ ” is a low-demand problem, as the first step only involves small numbers (i.e., $7 - 4$) and there is no borrow operation. In contrast, “ $47 \equiv 19 \pmod{4}$ ” is a high-demand problem, as the first step involves large numbers (i.e., $47 - 19$) and involves a borrow operation.

To increase the likelihood of choking under pressure, the current study exclusively used high-demand MA problems. MA problems were formulated based on prior studies (Beilock & Carr, 2005; Beilock & DeCaro, 2007). Specifically, high-demand problems had the following attributes: the first value was between 20 and 99, the last value was between 2 and 9, and the first step required a double-digit carry-over operation. Each problem had two versions of itself, a true version and a false correlate that only differed as a function of the value in the mod statement. For example, “ $77 \equiv 59 \pmod{3}$ ” has large number in the first step (i.e., $77 - 59$), the last value is between 2 and 9 (i.e., mod 3) and has a carry-over operation (i.e., $77 - 59$ requires more intermediate steps to solve than $77 - 51$, which can be solved without a carry-over operation). As “ $77 \equiv 59 \pmod{3}$ ” can be solved as true (i.e., $77 - 59 = 18$, $18 \div 3 = 6$), the false correlate would be “ $77 \equiv 59 \pmod{4}$ ” (i.e., $77 - 59 = 18$, $18 \div 4$ is not a whole number). Problems were formulated using the random number generator (RNG) at random.org. The RNG was set to randomize a value between 20 and 99, and 2 and 9 to determine the first and last values of a true problem, respectively. Next, a number between 2 and 9 was randomized to act as a multiplier for the last value and the resulting product was subtracted from the first value to complete the true problem. For example, 32 and 9 are randomly generated as the first and last values. 2 is randomly generated as the multiplier and multiplied with 9, then subtracted from 32 to obtain 14 to complete the expression (i.e., $9 \times 2 = 18$, $32 - 18 = 14$). The resulting true problem is expressed as, “ $32 \equiv 14 \pmod{9}$.” Practically, the possible range for the multiplier is restricted to what can be divisible and still result in a positive whole number, such as the values 2 and 3 being the only possible values for this example ($9 \times 2 = 18$ and $9 \times 3 = 27$). Regardless, the exact number used for the stimuli was always randomized. False correlates were manually generated by adjusting the last value up or down by 1 while staying within the predefined 2 to 9 range.

Exceptions were made in certain cases when the value obtained from the first step (e.g., $32 - 14$) could easily exclude certain answers. For example, with mod 3, if the answer obtained from the first step was odd, then a false correlate of 2 would be very easy to answer. In contrast, for mod 2, a false correlate of 3 would not be as easy to exclude as the two can potentially share a common factor (i.e., 18).

The study was coded in lab.js, a free, open-sourced, online study builder. This experiment builder platform has excellent documented timing performance across the most popular web browsers today (i.e., Chrome, Firefox, Safari, Edge) and across different operating systems (i.e., Linux, Mac OS, Windows) (Henninger et al., 2021). Web browsers and operating systems were not restricted in this study.

Procedure

Modular Arithmetic. Participants started the study by navigating to the website hosting the online experiment where they were informed of the purpose of the study and gave informed consent to continue (see Appendix A). Throughout the experiment, instructions and stimuli were given via on-screen text and/or visuals (see Appendix A). Participants indicated their understanding of these instructions with a mouse click or by pressing a specified key on the keyboard before proceeding. Participants were instructed that they should complete the study in one sitting and that if they were unable to do so (e.g., a network connection failure, navigating away from the experiment web page, etc.), they were required to restart the experiment from the very beginning. Participants then answered demographic questions, personality questionnaires (SCS and STAI), and completed a working memory assessment (RDS).

Participants were then given instructions on how to solve the MA problems in a practice block and reminded that the task should be completed independently and without any outside

help, including the use of a phone, calculator, scrap paper, etc. (see Appendix A5). Participants were then shown two examples of MA problems that were solved on-screen and explained in detail, before completing a practice block. Participants indicated if a problem was true by pressing “z” and if a problem was false by pressing “m.” These instructions were visible for all trials throughout the experiment. During the practice block, participants were required to correctly solve a set of four randomly ordered MA problems in a row, which consisted of two problems and their true and false correlates. If a participant failed to correctly answer all four problems in a set, they were given additional trials until they either correctly solved all the problems in a set or reached 99 trials. Participants who reached 99 trials and were still unable to correctly solve the MA problems were excluded from analysis. As an aside, it should be acknowledged that having four problems in a set would mean that if participants could not complete the practice block after the 96th trial (i.e., 24 sets), then they would not be able to complete the practice block by the 99th trial (since there are only 3 problems left and not enough to complete another set).

Once the participants successfully completed the practice block, they proceeded to the experiment, which was presented as another practice block. Participants were instructed to solve the additional practice MA problems as quickly and accurately as possible and to learn the task to the best of their ability. Participants were specifically instructed so that all problems in the experiment should be completed on their own without any outside help, including the use of their phone, calculators, scrap papers, etc. Participants were also informed that they would later be tested on their progress twice with the possibility of earning rewards based on their performance (i.e., the pressure manipulations) but that the exact nature of the rewards would be elaborated upon later. Participants then completed their first baseline trials, consisting of 24 MA problems,

followed by pressure trials that also consisted of 24 MA problems. Participants were informed about the details of the pressure manipulations before the start of the first pressure trials. Together, the baseline-pressure trials constituted the first block of the experiment. Participants completed the experiment by doing a second block of baseline-pressure trials for a total of 96 MA problems, in which the order of the problems was randomized for each participant. Participants were briefly reminded of the pressure manipulations at the start of the second baseline trials and reminded in more detail at the start of the second pressure trials. Notably, five MA problems were inadvertently duplicated due to experimenter error (see Table 4), but this did not affect any results.¹

Table 4

An Overview of the Research Design for the Modular Arithmetic Task

Task	Modular Arithmetic				
Block	Practice	1		2	
# of Trials	4-99	24*	24*	24*	24*
Pressure Condition	-	Baseline	Pressure	Baseline	Pressure

Note. Due to experimenter error, each condition within a block did not in fact, contain 24 MA problems but slightly less for a total of 91 MA items across the two baseline-pressure blocks.

After the MA problems, participants answered post-study questionnaires that consisted of manipulation checks of the pressure manipulations and participant engagement. Participants were debriefed on the true nature of the study, including the deception involved in the pressure manipulations. The pressure manipulations included a team- and performance-based monetary incentive that was compensated to all participants regardless of performance.

Pressure Manipulations. Performance pressure was manipulated by informing participants that their performance would be tested twice over the course of the study and that

rewards could be earned based on performance during those evaluations. Participants were made aware of this briefly before the start of the first and second baseline trials and in detail before the start of the first and second pressure trials.

The pressure manipulations consisted of informing participants that a performance score would be calculated based on their speed and accuracy of solving the MA problems; participants were explicitly instructed to solve the problems as quickly as possible without sacrificing accuracy. Participants were notified that they could earn a \$5 reward based on their final performance score across the two tests, but that they had been randomly paired with another participant and the reward was contingent on the pair's average final performance score being within the top 20% of all pairs in the study. Each participant was also informed that their partner had already completed the study, that their partner's performance was currently within the top 18%, and that now it was up to them to secure the reward or squander the opportunity. The partner's performance of 18% was chosen based on the assumption that a non-rounded number would increase verisimilitude and the number was also shown in a way as if it looked calculated and was not scripted (see Appendix A8). Participants were also instructed that their performance during the tests would be recorded via a screen capturing software as part of a university initiative on assessing undergraduate math curriculum and that their performance would be evaluated later by experts. Participants were also shown that a red dot on the top right corner of the screen would indicate that screen capturing was in progress. In reality, there was no other participant, screen recording or evaluation, and all of the pressure manipulations were ruses to induce performance pressure. Participants were notified of this in the debriefing at the end and received \$5 regardless of their performance.

Manipulation Checks. The manipulation checks were presented at the end of the session after all baseline and pressure trials had been completed and consisted of three parts.

Following procedures from Balk et al. (2013), a 5-item pressure/tension subscale from the Intrinsic Motivation Inventory (Ryan, 1982) and a 4-question competitiveness questionnaire on a 7-point Likert scale on how important, engaging, difficult, and exciting participants perceived the tasks to be (Veldhuijzen et al., 2002) were presented (see Tables 5 and 6).

Table 5

Five-item Pressure/Tension Subscale from the Intrinsic Motivation Inventory (Ryan, 1982)

	Not at all true			Somewhat true			Very true
I did not feel nervous at all while doing this.	1	2	3	4	5	6	7
I felt very tense awhile doing this activity.	1	2	3	4	5	6	7
I was very relaxed in doing these.	1	2	3	4	5	6	7
I was anxious while working on this task.	1	2	3	4	5	6	7
I felt pressured while doing these.	1	2	3	4	5	6	7

Note. Participants were asked about how they felt while performing the tasks and if the performance pressure manipulations worked as intended.

Table 6

Four-Question Competitive Questionnaire on a 7-point Likert Scale (Veldhuijzen et al., 2002)

	Not at all						Extremely
I found the task to be competitive	0	1	2	3	4	5	6
I found the task to be engaging	0	1	2	3	4	5	6
I found the task to be difficult	0	1	2	3	4	5	6

Table 6 (cont'd)

I found the task to be exciting	0	1	2	3	4	5	6
---------------------------------	---	---	---	---	---	---	---

Note. Participants were asked about how engaged they were with the task and if the performance pressure manipulations worked as intended.

A series of questions related to participant experiences was asked (see Table 7) and participants were instructed that their answers would not impact their potential to earn rewards (this deception was not revealed until after the manipulation checks). Participants answered how they completed the task and depending on the answer, participants were excluded from analysis in instances when they did not appropriately solve the MA problems (e.g., use a calculator, scrap paper, count on fingers, etc.). Participants also indicated their feelings of perceived pressure during the pressure trials and their perceived importance of the task based on a 7-point Likert scale (0 = *Not at all*, 3 = *Somewhat*, 6 = *Absolutely*). These measures were collected as situational importance of the task is regarded as an important component of perceived pressure (Hill et al., 2010). Participants were also asked about whether they believed the pressure manipulations or not based on a 6-point Likert scale (0 = *Not at all*, 5 = *Absolutely*). The Likert scale for assessing belief in the manipulations was an even scale to allow for coding as a dichotomous variable and to classify participants as “believers” or “non-believers” of the manipulations. Finally, participants indicated if they had any issues completing the study in one sitting. Participants were also explicitly asked what strategies they used to maximize their performance during the experiment and that their answers would not disqualify them from earning potential rewards. This was important to get an honest response from participants so that only those who completed the experiment as intended (on their own, without any outside help).

Table 7*Modular Arithmetic-Specific Manipulation Checks Administered at the End of the Study*

Study-Specific Manipulation Checks
Describe any strategies you used to maximize performance during the experiment. (Open-ended response)
Did you feel more pressure during the test blocks than the practice blocks during the experiment? (0-6; Not at all, Somewhat, Absolutely for endpoints and midpoint)
Explain why you felt pressure (or did not feel pressure) to perform well during the experiment. (0-6; Not at all, Somewhat, Absolutely for endpoints and midpoint)
Did you feel it was important to perform at a high level during the test blocks? (0-6; Not at all, Somewhat, Absolutely for endpoints and midpoint)
Did you believe that you would not get the \$5 reward if you and your partner's performance score did not finish in the top 20%? (0-5; Not at all to Absolutely)
Did you believe that you had a partner that was counting on you to finish in the top 20% for each of you to get the \$5 reward? (0-5; Not at all to Absolutely)
Did you believe that you were being recorded when the red dot was present on the screen? (0-5; Not at all to Absolutely)
Were you able to complete the study in one sitting without any interruptions (e.g., internet connection issues, accidentally closed the webpage, etc.)? (Yes/No answer)

Note. Participants were asked about how they completed the task, their perceived level of pressure, perceived importance of the task, whether they believed the pressure manipulations or not, and if they had any issues completing the study in one sitting.

Results***Exclusion Criteria***

To ensure the highest quality data possible, exclusions were handled in two stages: participant-based exclusions and trial-based exclusions. Accounting for both participant-based and trial-based exclusion criteria resulted in 144 participants and 12,949 trials being analyzed, which were narrowed down from an initial pool of 180 participants who completed the study. Of the 180 participants, 36 participants (20%) were excluded for various reasons. 17 of the 36 participants (9.4%) were excluded for using improper strategies for solving the MA problems:

five participants reported using a calculator, three participants reported using a scrap paper, six verbalized the problems aloud, and three used fingers to count. An additional 17 of the 36 participants (9.4%) were excluded from analysis for failing to properly progress through the practice block in under 99 trials. Two participants (1.1%) were excluded from analysis for having an excessive number of trials (for both accuracy and response times) being deemed ineligible (see trial-based exclusions below) and having near chance-level accuracies (50%). Participants were also checked for whether their average RT across all answered MA problems (up to 91, based on the five excluded duplicate problems) was greater than 3.5 *SD* over the grand mean (i.e., average RT of all participants), but no participants were excluded based on this criterion. Out of a potential 13,824 trials (i.e., 144 participants completing 96 trials), trial-based exclusions consisted of removing both accuracy and response times of 875 problematic trials (6.3%). Of the 875 problematic trials removed, 28 trials (3.2%) were excluded for having a negative or 0 ms response time, indicating that participants did not wait for the stimuli to appear on the screen before responding. 720 trials (82.3%) were excluded for being repeat occurrences of duplicate problems (five MA problems were duplicated due to experimenter error). 127 trials (14.5%) were excluded because the response time was 3.5 *SD* above the mean for a given MA problem; means and *SDs* were calculated for each MA problem to account for any outliers (e.g., participants taking an abnormally short or long time to answer).

Pressure Manipulations

The pressure manipulations were assessed based on the definition of choking as “heightened levels of perceived pressure accompanied by a suboptimal performance level” (Beilock & Gray, 2012, p. 426). This was accomplished by computing three scores: a perceived pressure score, a belief score, and a choke score. For a performance to be classified as a choke

and not just as a performance decrement (see Definitions of Choking Under Pressure in Introduction), participants should perceive heightened levels of pressure in response to the pressure manipulation, believe that there is something real at stake due to the pressure manipulation, and consequently experience a performance decrement under pressure relative to baseline performance.

The perceived pressure score was based on participants' responses to the manipulation checks asking: a) if participants felt more pressure during the pressure blocks than during the baseline blocks and b) if participants felt it was important to perform at a high level during pressure blocks (see Table 7). These manipulation checks were based on a 7-point Likert scale (0 – 6) and combined for this measure. A score of 6 or higher on this measure was the cut-off for participants feeling an elevated level of perceived pressure, as this would indicate that participants scored an average of 3 (at midpoint) on each manipulation check.

The belief score was calculated based on participants' responses to the manipulation checks asking if they believed elements of the pressure manipulations (e.g., did you believe that you truly had a teammate counting on you?). As the belief score was composed of three different 6-point Likert scales (0 – 5) that were summed, a score of 9 or higher (average score in the upper half of the scale) was taken to be the cut-off for belief in the manipulations.

The choke score was derived from taking a difference score for accuracy and response times between baseline trials and pressure trials, with negative scores (i.e., a performance decrement) indicating a potential choke. Positive scores indicated either a “clutch” if participants perceived pressure, thus presumably overcoming that pressure to overperform relative to baseline, or a failure of the pressure manipulations to induce performance decrements if participants did not perceive pressure.

Out of 140 participants who completed the manipulation checks, 111 participants (79.3%) had heightened levels of perceived pressure. This was also reflected in the mean, which at 7.04, was above the threshold of feeling elevated perceived pressure (6 or higher). Out of 140 participants who completed the manipulation checks, only 60 participants (42.9%) believed the manipulations were real. This was also reflected in the mean, which at 7.44, was below the threshold of a believer (9 or higher). Out of 144 participants who completed the task, only 61 participants (42.4%) had a negative accuracy-based choke score, meaning that the pressure manipulations were not effective in producing performance decrements for most participants. Choke scores were calculated so that negative scores reflected performance decrements from baseline to evaluation (i.e., potentially a choke) and positive scores reflected performance enhancement from baseline to evaluation (i.e., potentially a clutch). The mean accuracy-based choke score among those that did choke was -7.2%, meaning chokers' accuracy during pressure trials were on average 7.2% worse than during baseline trials. Out of the same 144 participants, only 63 participants (43.8%) had a negative RT-based choke score. The mean RT-based choke score among those that did choke was -617.84 ms, meaning chokers' RT for correctly solved problems during pressure trials were on average 617.84 ms slower than during baseline trials. Among the four measures of the pressure manipulations, only the perceived pressure score and belief score were significantly correlated with each other $r(137) = .424, p < .001$, as well as the accuracy-based choke score and the RT-based choke score $r(142) = -.524, p < .001$. This negative correlation between the choke scores suggested a speed-accuracy trade-off. Given that the choke scores are calculated so that negative scores indicate a performance decrement (lower accuracy, slower RT) from baseline to pressure and positive scores indicate a performance

enhancement (higher accuracy, faster RT), there are four possible states of the world (SoWs), two of which are relevant for speed-accuracy trade-offs (see Table 8).

Table 8

Four Possible States of the World (SoWs) for Interpreting Correlations Between Choke Scores

State of the World (SoW)	Direction of Choke Score	Interpretation	Direction of Correlation
I	Positive accuracy-based choke score, negative RT-based choke score	Higher accuracy, slower RT	Negative
II	Positive accuracy-based choke score, positive RT-based choke score	Higher accuracy, faster RT	Positive
III	Negative accuracy-based choke score, negative RT-based choke score	Lower accuracy, slower RT	Positive
IV	Negative accuracy-based choke score, positive RT-based choke score	Lower accuracy, faster RT	Negative

Note. Interpretation of choke scores indicate that performance during pressure trials had higher/lower accuracy and slower/faster RT relative to baseline trials.

It is possible that SoW I or IV is possible in interpreting the negative correlation between accuracy-based choke score and RT-based choke score. Further analysis revealed that slower RT was associated with higher accuracy (SoW I) among: a) participants who did not experience performance decrements accuracy-wise (see Table 10) and b) participants who experienced performance decrements RT-wise (see Table 11). Faster RT was associated with lower accuracy (SoW IV) only among participants who experienced performance decrements accuracy-wise (see Table 9). There were no significant correlations among participants who did not experience performance decrements RT-wise (see Table 12) and participants who experienced performance decrements both accuracy- and RT-wise (see Table 13). Participants who did not experience both accuracy- and RT-based performance decrements (SoW II) had an association between higher accuracy and faster RT (see Table 14). Taken together, it appears that there is evidence for a

speed-accuracy trade-off among participants who resisted accuracy-based performance decrements and those who had performance decrements in RT, of which both groups slowed down performance during pressure trials to prioritize accuracy (SoW I), whereas only participants who had negative accuracy-based choke scores seemed to have prioritized speed over accuracy (SoW IV). Curiously, there was no correlation between the two choke scores among those who had negative choke scores on both.

Table 9

Participants with Negative Choke Scores (Accuracy %)

	<i>M</i>	<i>SD</i>	<i>N</i>	<i>r</i>	SoW	Interpretation
Accuracy CS	-0.0724	0.0618	61	-0.286*	IV	Lower accuracy
RT CS	455.6908	766.4508	61			Faster RT

Note. * $p < .05$

Table 10

Participants with Positive Choke Scores (Accuracy %)

	<i>M</i>	<i>SD</i>	<i>N</i>	<i>r</i>	SoW	Interpretation
Accuracy CS	0.0755	0.0648	83	-0.560**	I	Higher accuracy
RT CS	-170.034	1066.894	83			Slower RT

Note. ** $p < .01$

Table 11

Participants with Negative Choke Scores (RT)

	<i>M</i>	<i>SD</i>	<i>N</i>	<i>r</i>	SoW	Interpretation
Accuracy CS	0.0584	0.09453	63	-0.538**	I	Higher accuracy
RT CS	-617.839	911.1975	63			Slower RT

Note. ** $p < .01$

Table 12*Participants with Positive Choke Scores (RT)*

	<i>M</i>	<i>SD</i>	<i>N</i>	<i>r</i>	SoW	Interpretation
Accuracy CS	-0.0226	0.08337	81	-0.168	IV	Lower accuracy
RT CS	649.483	651.5061	81			Faster RT

Table 13*Participants with Negative Choke Scores (Accuracy % & RT)*

	<i>M</i>	<i>SD</i>	<i>N</i>	<i>r</i>	SoW	Interpretation
Accuracy CS	-0.0583	0.050807	16	-0.402	III	Lower accuracy
RT CS	-197.759	163.9665	16			Slower RT

Table 14*Participants with Positive Choke Scores (Accuracy % & RT)*

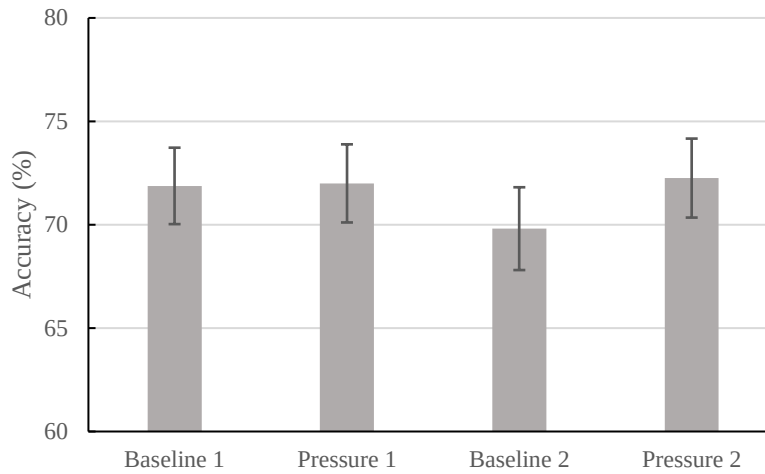
	<i>M</i>	<i>SD</i>	<i>N</i>	<i>r</i>	SoW	Interpretation
Accuracy CS	0.0217	0.09772	128	-0.580**	II	Higher accuracy
RT CS	131.6283	1051.449	128			Faster RT

*Note. ** $p < .01$*

The pressure manipulations were also evaluated by performing a 2 x 2 RM-ANOVA for accuracy and response time in baseline and pressure conditions across two baseline-pressure blocks. There were no significant effects of the pressure manipulations on accuracy, indicating that participants performed at a consistent level throughout the experiment (see Figure 1).

Figure 1

Effects of the Pressure Manipulations on Accuracy (Modular Arithmetic)

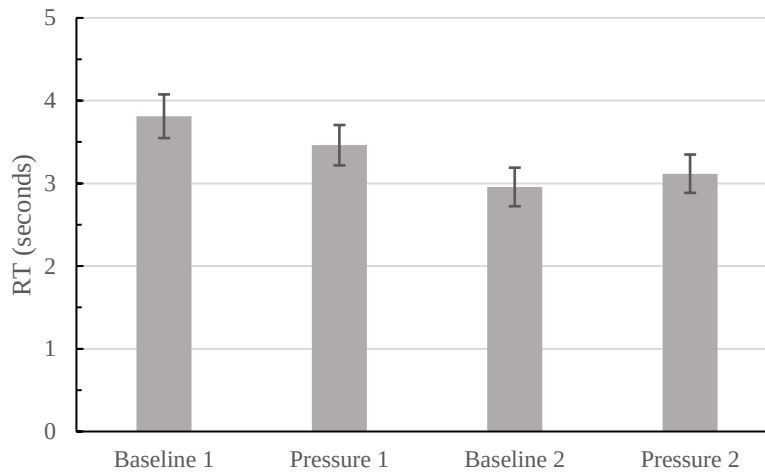


Note. This figure shows the 2 x 2 RM-ANOVA (pressure x block) for accuracy (higher is better accuracy) in MA problems. Participants' accuracy did not significantly differ throughout the experiment. Error bars represent one standard error.

For response time, there was a significant effect of block ($F(1,143) = 24.755, p < .001$) and an interaction ($F(1,1) = 9.816, p = .002$), as participants were faster in Block 2 than in Block 1, and their response times decreased more from baseline to pressure conditions in Block 1 than in Block 2 (see Figure 2). While participants were expected to improve over time, it was not expected that they consistently performed faster during the pressure condition than during the baseline condition.

Figure 2

Effects of the Pressure Manipulations on RT (Modular Arithmetic)



Note. This figure shows the 2 x 2 RM-ANOVA (pressure x block) for RT (lower is faster) in MA problems. There was a block effect and an interaction, but in the opposite direction of a choking effect. Error bars represent one standard error.

The same analysis was repeated with participants' perceived pressure from the pressure manipulations as a covariate to account for any confounds related to the effectiveness of the pressure manipulations to produce performance decrements (i.e., decreased accuracy and slower RTs from baseline to pressure trials) being dependent on whether participants felt pressure from the manipulations. There were no significant effects of the pressure manipulations on accuracy, regardless of whether participants felt pressure from the manipulations (see Table 15). For response time, there were also no significant effects (see Table 16). These results are in line with previous results, indicating that participants' task performance did not differ from baseline to pressure trials (i.e., no performance decrements) regardless of their level of perceived pressure.

Table 15

Effects of the Pressure Manipulations on Accuracy with Perceived Pressure from the Manipulations as a Covariate (Modular Arithmetic)

Effect	<i>F</i>	<i>df</i>	<i>p</i>
Pressure	.232	138	.631
Pressure * Perceived Pressure	1.194	138	.276
Block	.141	138	.708
Block * Perceived Pressure	.003	138	.954
Pressure * Block	.060	138	.806
Pressure * Block * Perceived Pressure	.604	138	.439

Note. * $p < .01$

Table 16

Effects of the Pressure Manipulations on RT with Perceived Pressure from the Manipulations as a Covariate (Modular Arithmetic)

Effect	<i>F</i>	<i>df</i>	<i>p</i>
Pressure	.066	138	.797
Pressure * Perceived Pressure	.027	138	.870
Block	.006	138	.938
Block * Perceived Pressure	3.425	138	.066
Pressure * Block	.556	138	.457
Pressure * Block * Perceived Pressure	.213	138	.645

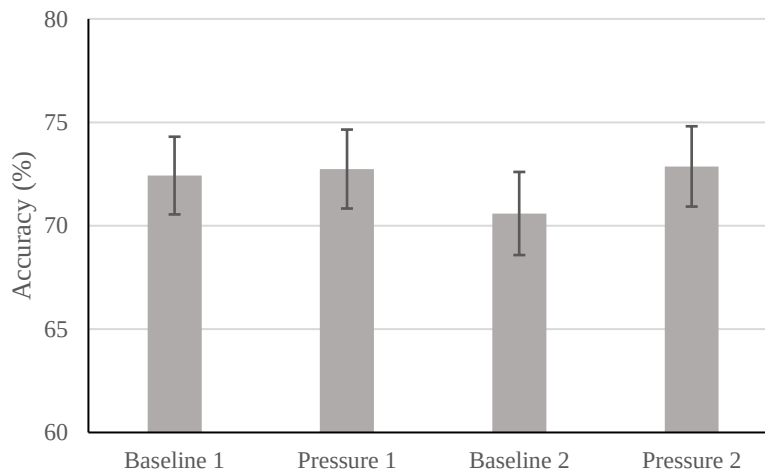
Note. * $p < .01$

The same analysis was repeated with participants' belief in the pressure manipulations as a covariate to account for any confounds related to the effectiveness of the pressure manipulations to produce performance decrements (i.e., decreased accuracy and slower RTs from baseline to pressure trials) being dependent on whether the participants believed the manipulations. There were no significant effects of the pressure manipulations on accuracy, regardless of their level of belief in the manipulations (see Figure 3). For response time, there was a significant interaction without the covariate ($F(1,1) = 9.012, p = .003$). This was consistent with the pattern of results observed earlier, indicating that participants' task performance did not

differ from baseline to pressure trials (i.e., no performance decrements) regardless of their level of belief that the pressure manipulations were real (see Table 17).

Figure 3

Effects of the Pressure Manipulations on Accuracy with Belief in the Manipulations as a Covariate (Modular Arithmetic)



Note. This figure shows the 2 x 2 Repeated measures ANCOVA (pressure x block) for accuracy (higher is better accuracy) in MA problems based on participants' belief in the manipulations. Participants' accuracy did not significantly differ throughout the experiment regardless of their belief. Error bars represent one standard error.

Table 17

Effects of the Pressure Manipulations on Accuracy with Belief in the Manipulations as a Covariate (Modular Arithmetic)

Effect	<i>F</i>	<i>df</i>	<i>p</i>
Pressure	.073	138	.788
Pressure * Belief	.914	138	.341
Block	1.392	138	.240
Block * Belief	2.477	138	.118
Pressure * Block	9.012	138	.003*
Pressure * Block * Belief	2.563	138	.112

Note. * $p < .01$

Reliability of Choking Measure

The reliability of the choke score was estimated by computing a split-half correlation for accuracy (based on percentage correct to account for uneven n because of duplicate problems) and RT (based on correctly solved problems only). Two choke scores were calculated on each of the measures for each participant ($N = 144$) based on subtracting the odd-numbered baseline trials from the odd-numbered pressure trials and doing the same with even-numbered trials. These two difference scores were correlated with each other. The reliability estimate for the MA task based on accuracy was 0.22, indicating that this measure had poor reliability. The reliability estimate for the choke score based on RT was 0.63, indicating that this measure had reasonably high reliability.

Additionally, choke score reliability was estimated using the same method as above on a per-block basis to account for potential differences in changes in skill level over the course of the experiment. The correlation for accuracy for Block 1 was 0.24 and Block 2 was 0.13. The correlation for RT for Block 1 was 0.71 and Block 2 was 0.55, which was in line with the overall reliability estimate.

Correlations with Individual Differences

Correlations were computed between the accuracy and RT-based choke scores and individual difference variables (working memory, dispositional self-consciousness, trait anxiety, and its subscales) to assess the association between choking under pressure and certain psychological traits. There were no significant correlations between the choke scores with any of the variables (see Tables 18 and 19).

Table 18*Correlations Between Choke Score (Accuracy %) and Individual Difference Variables*

Measure	Working Memory	Dispositional Self-Consciousness	Private Self-Consciousness	Public Self-Consciousness	Social Anxiety	Trait Anxiety
Choke Score (Accuracy %)	.036	-.003	.049	-.057	-.004	.114

Note. * $p < .05$ **Table 19***Correlations Between Choke Score (RT) and Individual Difference Variables*

Measure	Working Memory	Dispositional Self-Consciousness	Private Self-Consciousness	Public Self-Consciousness	Social Anxiety	Trait Anxiety
Choke Score (RT)	.013	.014	.007	-.019	-.069	-.106

Note. * $p < .01$

Reliabilities for the individual difference measures (except for working memory) were also estimated from the data based on Cronbach's alpha and maximum correlations were calculated using the disattenuation formula using overall reliability estimates for accuracy and RT (Schmidt & Hunter, 1999; see Hambrick et al., 2014 for an example). Observed reliabilities for SCS (Dispositional S-C) and its subscales (Private S-C, Public S-C, Social Anxiety) were unexpectedly low (see Table 20), thus limiting the maximum possible correlations (see Tables 21 and 22). In fact, all observed correlations (see Tables 18 and 19) were well below their respective maximum correlations (see Tables 21 and 22). Note that the extent to which measures may correlate with each other is limited by their reliabilities; measures with low reliabilities will have smaller correlations.

Table 20

Observed Reliabilities for Choke Scores and Individual Difference Variables Based on the Current Study Sample

Measure	Choke Score (Accuracy %)	Choke Score (RT)	Dispositional S-C	Private S-C	Public S-C	Social Anxiety	Trait Anxiety
Observed Reliabilities	.222	.629	.498	.367	.134	.057	.810

Table 21

Maximum Possible Correlations Between Choke Score (Accuracy %) and Individual Difference Variables Based on Observed Reliabilities

Measure	Dispositional S-C	Private S-C	Public S-C	Social Anxiety	Trait Anxiety
Maximum Correlations for Choke Score (Accuracy %)	.332	.285	.172	.112	.424

Note. See Table 20 for observed reliabilities of each individual difference variable and see Table 18 to compare with observed correlations.

Table 22

Maximum Possible Correlations Between Choke Score (RT) and Individual Difference Variables Based on Observed Reliabilities

Measure	Dispositional S-C	Private S-C	Public S-C	Social Anxiety	Trait Anxiety
Maximum Correlations for Choke Score (RT)	.560	.480	.290	.189	.714

Note. See Table 20 for observed reliabilities of each individual difference variable and see Table 19 to compare with observed correlations.

STUDY 2

The major question for Study 2 was if successful pressure manipulations could be created using a golf putting task via in an-person setup and to test for the reliability of the choking effect. This study also analyzed whether the calculated choke score would correlate with individual difference variables.

Method

Participants

Institutional review board approval was obtained before experimentation began. The study was completed by 111 undergraduate students over the age of 18 who were recruited through an online research recruitment service and were enrolled in undergraduate psychology courses in a large Midwestern university from the fall of 2021 through the spring of 2022. After applying exclusion criteria described in the Results section below, data from a total of 109 participants were included in the analyses.

Demographics. After providing consent, participants answered demographics-related questions. One participant did not respond to any of the demographics-related questions. Of the 108 participants that did respond, 35 were men (32.1%), 70 were women (64.2%), and three (2.8%) indicated “other” as their gender. While there was no restriction on prior golf experience, participants were asked about their familiarity with golf (see Table 23). 62 participants (56.9%) responded that they had no experience, 33 (30.3%) responded that they were beginners, four (3.7%) responded that they were novices, and nine (8.3%) responded that they were experienced. Participants were also asked about their own golf skill (see Table X). 84 participants (77.1%) reported that they were a complete beginner, 14 (12.8%) reported that they were a beginner, five (4.6%) reported that they were a novice, and five (4.6%) reported that they were experienced.

Table 23

Golf Experience/Skill Level Questionnaire

How would you rate your experience level in golf? Check the option that fits you best. (Golf does not include mini-golf, but includes play on the putting green, 9-hole play, or 18-hole play.)
☐ No Experience: I have never played a round of golf or mini golf in my life.
☐ Beginner: I have played fewer than 20 rounds of golf or mini golf in my life.
☐ Novice: I have played 20-50 rounds of golf or mini golf in my life.
☐ Experienced: I have played 50+ rounds of golf or mini golf in my life.
How would you rate your skill level in golf? Check the option that fits you best.
☐ Complete Beginner: I am not skilled enough to record an actual score for 18 holes.
☐ Beginner: My average for 18 holes is greater than 110.
☐ Novice: My average for 18 holes is 100-110.
☐ Experienced: My average for 18 holes is below 100.

Note. Participants were asked questions on their experience and skill level in golf prior to the study.

Materials

Measure. Multiple measures were used to assess the psychological correlates of choking under pressure.

Self-consciousness was assessed using the Self-Consciousness Scale (SCS) (Fenigstein et al., 1975), which includes 23 items (e.g., *I'm concerned about what other people think about me.*). The scale is composed of a 9-item private self-consciousness sub-scale, a 7-item public self-consciousness sub-scale, and a 6-item social anxiety sub-scale (see Table 2). SCS scores are calculated by summing the individual items.

Trait anxiety was assessed using the Trait Anxiety subscale of State-Trait Anxiety Inventory for Adults (STAI-AD) Form Y (Spielberger, 1983), which includes 20 items (e.g., *I am a steady person.*) on a 4-point scale (see Table 3). STAI scores are calculated by summing the individual items. Scores may range from a minimum of 20 to a maximum of 80. The

commonly accepted classifications of STAI scores are “no or low anxiety” for 20-37, “moderate anxiety” for 38-44, and “high anxiety” for 45-80 (Kayikcioglu et al., 2017).

For the golf putting task, sports-specific anxiety was assessed using the Sports Anxiety Scale-2 (SAS-2) (Smith et al., 2006), which includes 15 items (e.g., *My body feels tense.*) on a 4-point scale (see Table 24). SAS-2 scores are calculated by summing the individual items and is composed of 5-point somatic subscale, 5-point worry subscale, and 5-point concentration disruption subscale.

Table 24

Sports Anxiety Scale-2 (Smith et al., 2006)

	Not At All	A Little Bit	Pretty Much	Very Much
It is hard to concentrate on the game.	1	2	3	4
My body feels tense.	1	2	3	4
I worry that I will not play well.	1	2	3	4
It is hard for me to focus on what I am supposed to do.	1	2	3	4
I worry that I will let others down.	1	2	3	4
I feel tense in my stomach.	1	2	3	4
I lose focus on the game.	1	2	3	4
I worry that I will not play my best.	1	2	3	4
I worry that I will play badly.	1	2	3	4
My muscles feel shaky.	1	2	3	4
I worry that I will mess up during the game.	1	2	3	4
My stomach feels upset.	1	2	3	4

Table 24 (cont'd)

I cannot think clearly during the game.	1	2	3	4
My muscles feel tight because I am nervous.	1	2	3	4
I have a hard time focusing on what my coach tells me to do.	1	2	3	4

Note. Participants were asked about their sports-specific trait anxiety to understand its connection with individual propensity for performance decrements under pressure.

A difference-based choke score was calculated on a participant-by-participant basis for performance over the course of the entire experiment based on accuracy (i.e., cm to target). This choke score was calculated by comparing accuracy during baseline trials with performance during pressure trials. The same steps were repeated on a per-block basis to calculate three additional difference-based choke scores. This choke score compared baseline and pressure performance within each block (i.e., each block consisted of baseline trials that were immediately followed by pressure trials). Negative scores indicate that the participant potentially choked (i.e., did better during baseline than pressure trials). Positive scores indicated either a “clutch” if participants perceived pressure, thus presumably overcoming that pressure to overperform relative to baseline, or a failure of the pressure manipulations to induce performance decrements if participants did not perceive pressure.

Task. The golf putting task was administered on an 18-foot x 10-foot synthetic green with a Stimpmeter reading (a standardized measure of the speed of golf greens) of approximately 10, a typical speed for a recreational (non-championship) golf course. Participants were instructed to putt standard golf balls (Bridgestone e6) that were white with the university logo during baseline trials and yellow during pressure trials using a standard golf putter (Odyssey White Hot Pro 2.0 Rossie, with both right- and left-handed models) from a starting point

(marked by a 2-cm x 2-cm piece of masking tape) at a target 8 feet away (marked by a 2-cm x 2-cm piece of masking tape) with the perfect putt being one that stops rolling on the center of the marker. For each putt, Experimenter B measured the distance of the ball (in cm) from the center of the ball to the center of the target and Experimenter A recorded the results as they were called out by Experimenter B. Putts that lie directly on the center of the target were assigned a value of 0 cm. Putts that hit the wall behind the target (located 120 cm away) were marked as out of bounds (OB) and treated as missing values. Putts that did not hit the wall behind the target but rolled to the edge of the putting green up the “lip” (the edges of the putting green did not lie perfectly flat were curled slightly upwards in a “lip”) were marked as 118 cm. Participants set up each of the balls on their own and were instructed to properly center the ball on at the starting point by either of the experimenters when warranted.

Procedure

Study 2 consisted of two sessions. Participants were instructed to complete Session 1 (web-based questionnaire) after signing up for the study but prior to coming in for Session 2 (in-person golf putting); Session 2 was held during weekdays and typically occurred within 24 hours of Session 1. However, there were some exceptions that could result in the interval between Session 1 and Session 2 being much greater than 24 hours or much shorter than 24 hours. Nevertheless, given that the questionnaires in Session 2 asked about trait-based participant characteristics, the differences in delay were not expected to matter. In cases when the interval was greater than 24 hours, this was because participants signed up for the study up to a week in advance, meaning they signed up and immediately completed the web-based questionnaire before coming in for Session 2 which occurred up to five days later. In cases when the interval was shorter, participants completed Session 1 immediately before Session 2 or during Session 2

(participants were reminded and asked if they completed Session 1 and completed Session 1 before beginning in-person trials). In rare circumstances ($n = 2$), participants did not complete Session 1 or there were issues with data recordkeeping and did not complete Session 1 at all or completed it after Session 2. Data from Session 2 were used regardless.

Session 1. After acknowledging their consent, participants completed a web-based questionnaire about their dispositional self-consciousness, trait anxiety, sports-specific anxiety, demographics, and prior golf putting experience. The golf experience/skill level questionnaire asked participants to rate their own level of experience and skill level (see Table 23).

Session 2. Experimenter A was primarily tasked with directing instructions to the participants and recording measurement data. Experimenter B was primarily tasked with setting up the pressure manipulations (e.g., set up the camera, swap out the golf balls, place the poster with information on the pressure manipulations on the wall, etc.) and measuring the distance of each putt and calling out the measurements for Experimenter A to manually transcribe. Manually entered measurements were then later transcribed digitally and checked for inaccuracies. Experimenter B also collected and passed the golf balls back to the participant in sets of 10.

Reading to participants from a script, Experimenter A informed participants that the goal of the study was to investigate how people develop skill in golf putting (see Appendix B). Participants were then shown a brief instructional video on how to putt titled “Putting Basics” (<https://www.youtube.com/watch?v=X4ZT9HHvX88>), given the appropriate equipment (right-handed or left-handed golf putter), and instructed on the task. Participants performed several (<5) practice golf putts based on these directions before beginning the tasks to ensure basic competence. During this practice block, participants were required to make two consecutive putts that were not considered out of bounds (OB) before they could proceed, regardless of whether

they indicated that they were ready. A putt is considered OB if the ball bounces off the back wall (located 120 cm behind the target). Participants were not made explicitly aware of this requirement unless their attempts were repeatedly OB. In such cases, participants were gently told to aim at the target without hitting OB to minimize any pressure they may feel.

Once participants started the task, performance pressure was manipulated within-subjects (baseline trials vs. pressure trials); participants completed a block of 40 baseline putts, followed by 20 pressure putts and repeated this three times for a total of 180 trials (see Table 25). After each baseline-pressure block (60 trials), participants were given a 3-minute break. This design was chosen to account for several considerations: (a) many participants may be beginners and may require training to reach a sufficient skill level to show performance decrements under pressure, and (b) ensure sufficient pressure trials for robust choke score reliability.

Table 25

An Overview of the Research Design for the Golf Putting Task

Task	Golf Putting						
Block	Practice	1		2		3	
# of Trials	≈5	40	20	40	20	40	20
Pressure Condition	-	Baseline	Pressure	Baseline	Pressure	Baseline	Pressure

Note. Participants were given a 3-minute break after the first two baseline trials (indicated by double border lines). Manipulation checks and debriefing were provided after the final baseline-pressure block.

Pressure manipulations. Performance pressure was manipulated by swapping the white golf balls used in the baseline condition with yellow balls and telling participants that they were eligible to earn rewards by performing at a high level, and that it was a team effort. Participants were also told that there would be a video camera recording their performance to track their

progress and for later evaluation by experts (see below for details), but only during the pressure trials. In reality, nothing was recorded, and all participants received \$5 at the end of their participation regardless of their performance.

A Sony (HDR-CX440) camcorder was placed on a tripod behind the participant in a pre-marked location, within their peripheral view while fixating on the ball as they take their backswing. During the baseline trials, the camera was turned off and a black cinch bag was placed over it to completely cover the camera. During the pressure trials, Experimenter B turned the camera on and with the viewfinder facing away from the participant (to prevent the participant from knowing that it was not recording) and adjusted the camera to ensure that the participant was clearly in the frame before pretending to set it to record. In reality, the camera made a noise when it is turned on but did not make a noise when it began recording, allowing for the experimenter to give the illusion that the camera had begun recording when it has not. At the start of the first pressure trials, Experimenter A told the participants that there were three ways to earn rewards during the pressure trials. The first is that participants would be video recorded as they performed the task and that the recording would be analyzed by a golf coach and golf pro at an upcoming sports psychology meeting to be sponsored by the Kinesiology Department. The golf coach and golf pro would analyze each participant's performance and would select up to five participants whom they deem have made the most improvements (based on non-performance factors such as form and level of concentration and accounting for novice skill levels) and those participants would earn a \$5 "most improved players" award. Participants were also told that the second way to earn rewards is that they had been paired with a random participant to form a team and that if both of their averaged performance (accuracy to target) during their pressure putts is in the top 20% of all participants tested, that they would earn \$5

each. The participants were then told that the other participant had already successfully completed the study and is currently in the top 20%, leaving it up to the participant to earn or lose the “team reward” for both. Finally, participants were told that their individual performance made them eligible for an additional \$10 for each of the team members if they hit the closest putt to the target during the pressure trials. These pressure manipulations were designed to keep participant engagement high even if participants perceived their overall performance to be insufficient for some of the rewards.

Manipulation Check. The manipulation checks were presented at the end of the session after the task was completed. They were the same as in the MA task in that they consisted of three parts but differed in the study-specific manipulation checks to account for the differences in pressure manipulations (see Table 26).

Table 26

Golf Putting-Specific Manipulation Checks Administered at the End of the Study

Study-Specific Manipulation Checks
Describe any strategies you used to maximize performance during the experiment. (Open-ended response)
Did you feel more pressure during the evaluation blocks than the practice blocks during the experiment? (0-6; Not at all, Somewhat, Absolutely for endpoints and midpoint)
Explain why you felt pressure (or did not feel pressure) to perform well during the experiment. (0-6; Not at all, Somewhat, Absolutely for endpoints and midpoint)
Did you feel it was important to perform at a high level during the test blocks? (0-6; Not at all, Somewhat, Absolutely for endpoints and midpoint)
Did you believe that you would earn the monetary incentive if the averaged performance between you and your partner was in the top 20% of all participants? (0-5; Not at all to Absolutely)
Did you believe that your performance was being video recorded? (0-5; Not at all to Absolutely)

Table 26 (cont'd)

Did you believe that you would earn the monetary incentive if the golf coach and golf pro evaluated your performance as the “most improved” during the experiment? (This evaluation is based not just on your overall performance but on your form, concentration, etc.) (0-5; Not at all to Absolutely)
Did you believe that you would earn the monetary incentive if you managed to get the closest putt to the target? (0-5; Not at all to Absolutely)

Note. Participants were asked about how they completed the task, their perceived level of pressure, perceived importance of the task, whether they believed the pressure manipulations or not, and if they had any issues completing the study in one sitting.

Results

Exclusion Criteria

Accounting for both participant-based and trial-based exclusion criteria resulted in 109 participants and 19,614 trials being analyzed, which were narrowed down from an initial pool of 111 participants who completed the study. Of the 111 participants, two participants (1.8%) withdrew consent after the being debriefed on the deception used in the pressure manipulations. Out of 19620 trials, six golf putts (0.03%) were excluded due to difficulties in transcribing measurements.

Putts that missed the target long and hit the wall behind the target, located 120 cm away, were counted as missing values. Putts that were hit long but were not going fast enough to hit the wall (to count as a missing value) but instead rolled down from the slight “lip” formed by the edge of the putting green were assigned a value of 118 cm (the distance from the lip to the target). This lip was formed due to the imperfect nature of the putting green, as it did not lie completely flat from end to end, meaning that the putting green next to the wall formed a slight “lip” that curled up near the back wall. These putts represented one of the clearest indications of poor performance, as the vast majority of these out of bounce (OB) putts—coded as missing

values—and “roll down” putts—assigned a value of 118 cm—occurred during the baseline trials in the first block before rapidly decreasing in frequency as participants completed more trials (see Table 27). This frontloaded distribution of these poor performing putts limited its utility as a potential indicator of choking. OB putts were excluded from analysis due to the difficulty of assigning them an accurate value, but roll down putts were included (i.e., given a value of 118 cm). Balls that missed short were not restricted in this way and could potentially have values exceeding 120 cm from the target.

Table 27

Frequencies of OB and Roll Down Putts in the Golf Putting Task

Task	Golf Putting					
Block	1		2		3	
OB Putts (% of block total)	79 (1.8%)	14 (6.4%)	22 (0.5%)	1 (0.0%)	7 (0.2%)	4 (0.2%)
Roll Down Putts (% of block total)	36 (0.8%)	6 (0.1%)	15 (0.3%)	4 (0.2%)	8 (0.2%)	2 (0.1%)
Pressure Condition	Baseline	Pressure	Baseline	Pressure	Baseline	Pressure

Note. This table shows the frequencies of OB and roll down putts in the golf putting task with percentages indicating what proportion they made up of the total putts in the given block in parentheses. OB putts were defined as missing the target long and hitting the wall behind the target located 120 cm away and were assigned a missing value, whereas roll down putts were defined as missing long and nearly hitting the wall but “rolling down” from the lip of the putting green adjacent to the wall and were assigned a value of 118 cm.

Pressure Manipulations

The effect of the pressure manipulations was evaluated in the exact same way as in Study 1, except for the specific manipulation checks involved in assessing participants’ belief in the manipulation (see Table 26). Since there were four manipulation checks related to belief, the cut-off for belief in the manipulations was a 12 or higher (average of 3 or higher on each manipulation check). Performance decrements were based on taking a difference score for

accuracy (cm to target) between baseline trials and pressure trials, with negative scores indicating a potential choke. Positive scores indicated either a “clutch” if participants perceived pressure, thus presumably overcoming that pressure to overperform relative to baseline, or a failure of the pressure manipulations to induce performance decrements if participants did not perceive pressure.

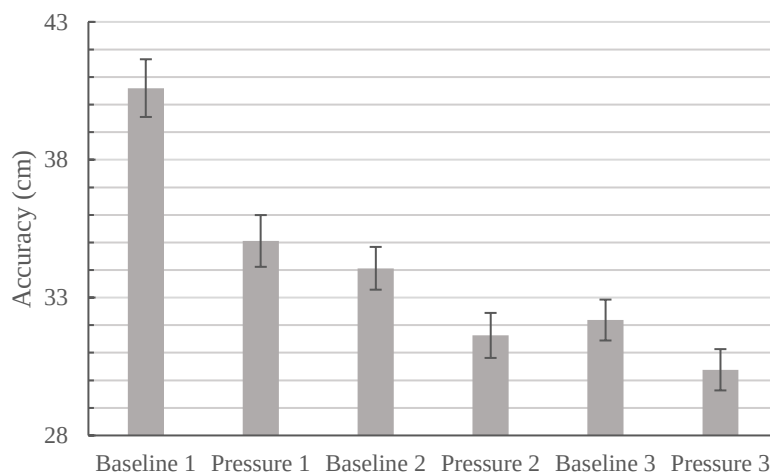
Out of 108 participants who completed the manipulation checks, 92 participants (85.2%) had heightened levels of perceived pressure. This was also reflected in the mean, which at 8.44, was above the threshold of feeling elevated perceived pressure (6 or higher). Out of 109 participants who completed the manipulation checks, 77 participants (70.6%) believed the manipulations were real. This was also reflected in the mean, which at 13.63, was above the threshold of a believer (12 or higher). Out of the 109 participants who completed the task, only 27 participants (24.8%) had a negative choke score, meaning that the pressure manipulations were not effective in producing performance decrements for most participants. The mean choke score (cm to target) was -1.52 cm, meaning among those that did choke, their accuracy during pressure trials ($M = 33.23$ cm) were on average 1.52 cm or 4.8% worse than during baseline trials ($M = 31.71$ cm). Among the three measures of the pressure manipulations, only the perceived pressure score and belief score were significantly correlated with each other $r(107) = .331, p < .001$.

The pressure manipulations were also assessed by computing a 2 x 3 RM-ANOVA for accuracy in baseline and pressure conditions across 3 baseline-pressure blocks. There was a significant effect of pressure ($F(1,108) = 82.829, p < .001$), block ($F(2,216) = 76.107, p < .001$), and an interaction ($F(2,216) = 8.668, p < .001$). Participants’ putts were significantly closer to the target during pressure trials than baseline trials, their performance significantly improved

across blocks, and the interaction was mainly driven by the much greater improvement from baseline to pressure conditions in Block 1 than in all subsequent blocks (see Figure 4). While participants were expected to improve over time, it was not expected that they consistently performed better during the pressure condition than during the baseline condition.

Figure 4

Effects of the Pressure Manipulations on Accuracy (Golf Putting)



Note. This figure shows the 2 x 3 RM-ANOVA (pressure x block) for accuracy (lower is higher accuracy) in golf putting. There was a significant effect of pressure, block, and a pressure x block interaction, but no indication of choking. Error bars represent one standard error.

The same analysis was repeated with participants' perceived pressure from the pressure manipulations as a covariate to account for any confounds related to the effectiveness of the pressure manipulations to produce performance decrements (i.e., decreased accuracy from baseline to pressure trials) being dependent on whether participants felt pressure from the manipulations. There was an effect of pressure ($F(1, 106) = 12.497, p < .001$) and block ($F(2, 212) = 5.450, p = .006$). As sphericity was violated ($\epsilon = .914$), Huynh-Feldt corrected results are reported (see Table 28). This did not fundamentally contradict the pattern of results observed

earlier, indicating that participants' task performance did not differ from baseline to pressure trials (i.e., no performance decrements) regardless of their level of perceived pressure. As previously shown, most participants were classified as having elevated levels of perceived pressure from the pressure manipulations. Therefore, participants could be characterized as overall resisting the pressure manipulations despite the pressure they felt from them and performing better under pressure than under baseline during each block of the study.

Table 28

Effects of the Pressure Manipulations on Accuracy with Perceived Pressure from the Manipulations as a Covariate (Modular Arithmetic)

Effect	<i>F</i>	<i>df</i>	<i>p</i>
Pressure	12.497	138	<.001**
Pressure * Perceived Pressure	.616	138	.434
Block	5.450	138	.006*
Block * Perceived Pressure	.955	138	.380
Pressure * Block	2.231	138	.110
Pressure * Block * Perceived Pressure	.425	138	.653

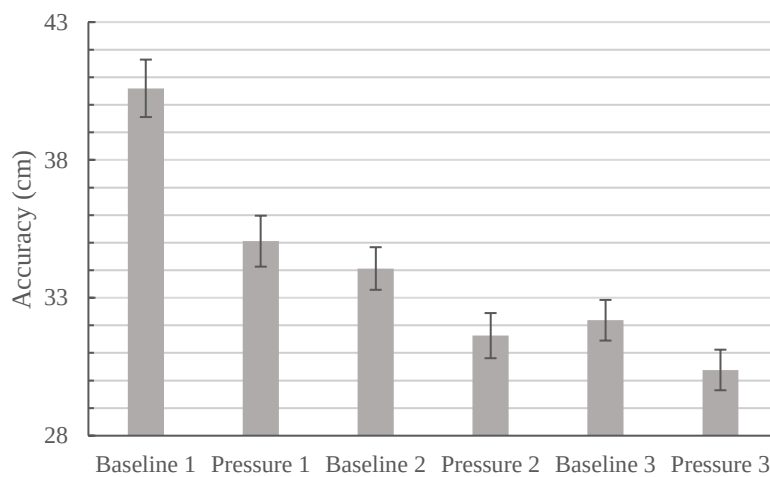
Note. ** $p < .001$, * $p < .05$

The same analysis was repeated with participants' belief in the pressure manipulations as a covariate to account for any confounds related to the effectiveness of the pressure manipulations to produce performance decrements (i.e., decreased accuracy from baseline to pressure trials) being dependent on whether the participants believed the manipulations. There was an effect of pressure ($F(1,107) = 6.130$, $p = .015$) and block ($F(2,214) = 12.516$, $p < .001$). There was no interaction and there were no effects of the covariate (see Figure 5). This did not fundamentally contradict the pattern of results observed earlier, indicating that participants' task performance did not differ from baseline to pressure trials (i.e., no performance decrements) regardless of their level of belief in the pressure manipulations. As previously shown, most participants were classified as having believed the pressure manipulations were real. Therefore,

participants could be characterized as overall resisting the pressure manipulations despite believing that they were real and performing better under pressure than under baseline during each block of the study.

Figure 5

Effects of the Pressure Manipulations on Accuracy with Belief in the Manipulations as a Covariate (Golf Putting)



Note. This figure shows the 2 x 3 Repeated measures ANCOVA (pressure x block) for accuracy (lower is better accuracy) in golf putting based on their belief in the manipulations. There was a significant effect of pressure and block, but no interaction. Participants' belief in the manipulations were not related to performance decrements from the pressure manipulations. Error bars represent one standard error.

Reliability of Choking Measure

The reliability of the choke score was estimated by taking a split-half approach similar to Study 1. Two choke scores were calculated for each participant ($N = 109$) based on subtracting the odd-numbered baseline trials from the odd-numbered pressure trials and doing the same with the even-numbered trials, then correlating the two scores. The correlation was $r = -0.12$, indicating that the choke score had very low reliability.

Additionally, choke score reliability was estimated using the same method as above on a per-block basis to account for potential differences in changes in skill level over the course of the experiment. The correlation for Block 1 was 0.06, Block 2 was -0.09, and Block 3 was -0.02, which was in line with the overall reliability estimate.

Correlations with Individual Differences

Correlations were computed between choke scores and individual difference variables (dispositional self-consciousness, trait anxiety, and sports-specific anxiety) to assess the association between choking under pressure and certain psychological traits. The total scale of the SAS-2 was correlated with choke score ($r(106) = -.205, p = .034$) and the worry subscale of the SAS-2 was correlated with choke score ($r(106) = -.279, p = .003$). The negative correlations shown in Table 29 and plotted in Figures 6 and 7 indicate that the greater the degree of performance decrements (i.e., more negative choke scores indicate greater performance decrements) experienced by participants, the higher they were in certain psychological traits (i.e., reported greater levels of dispositional self-consciousness, trait anxiety, sports anxiety, etc.). While this relationship was one that was expected given the previous literature, the exact interpretation of these results is unclear due to the low reliability of the choke score and the pressure manipulations facilitating performance enhancement rather than performance decrements. However, the finding of null effects at the group level is not inscrutable, as Miller and Schwarz (2018) offer a tantalizing approach to account for the possibility. The researchers argue that while it is possible that an on-average null effect may mean that the manipulation had no effect, it is also possible that the manipulation may have had an effect on many or all the individuals, but in opposing directions, thus resulting in the overall null effect. Miller and Schwarz propose that whenever an on-average null effect is found, an analysis should be

conducted to account for both possibilities. Nevertheless, this analysis lies beyond the scope of this paper.

Table 29

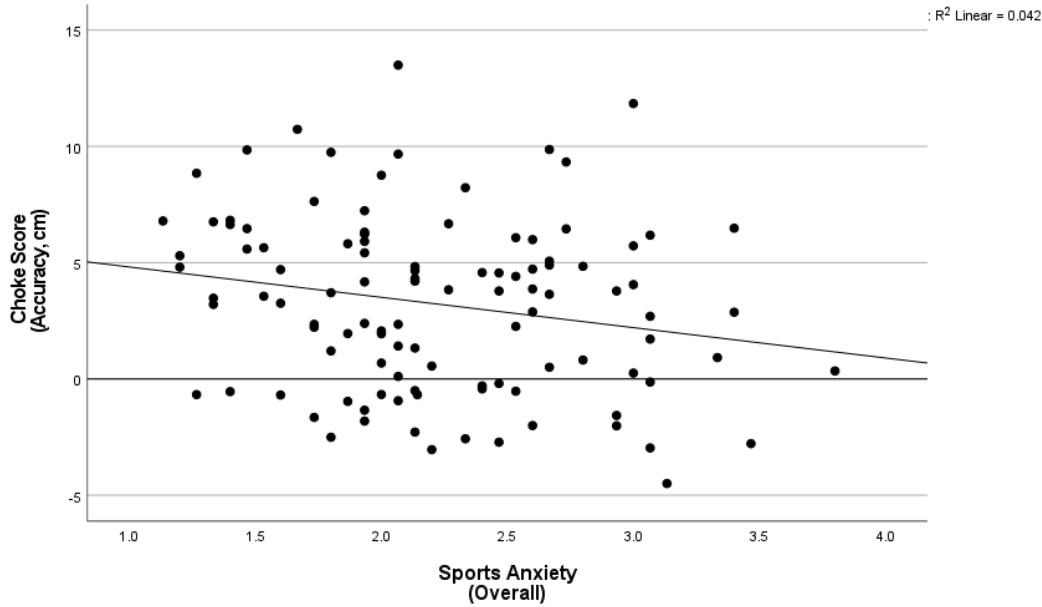
Correlations Between Choke Score (Accuracy, cm) and Individual Difference Variables.

Measure	Trait Anxiety	Dispos. S-C	Private S-C	Public S-C	Social Anxiety	Sports Anxiety (Overall)	Sports Anxiety (Somatic)	Sports Anxiety (Worry)	Sports Anxiety (Concentration Disruption)
Choke Score (Accuracy, cm)	.007	-.056	.061	-.140	-.066	-.205*	-.096	-.279**	-.113

Note. * $p < .01$, ** $p < .001$

Figure 6

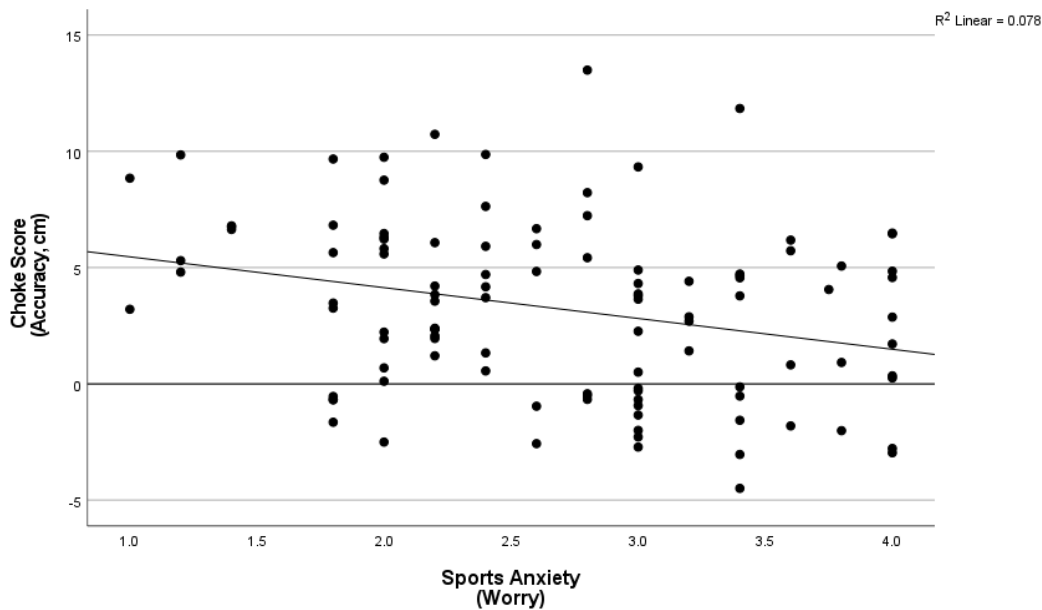
Correlation Between Choke Score and Sports Anxiety (Overall) in Golf Putting



Note. This figure shows the scatter plot of correlation between choke score (accuracy, cm) and sports anxiety (overall). Choke scores below 0 indicate performance decrements from baseline to pressure trials.

Figure 7

Correlation Between Choke Score and Sports Anxiety (Worry) in Golf Putting



Note. This figure shows the scatter plot of correlation between choke score (accuracy, cm) and sports anxiety (worry). Choke scores below 0 indicate a greater performance decrements from baseline to pressure trials.

Reliabilities for the individual difference measures (except for working memory) were also estimated from the data based on Cronbach's alpha and maximum correlations were calculated using the disattenuation formula using the overall reliability estimate for accuracy (Schmidt & Hunter, 1999; see Hambrick et al., 2014 for an example). The absolute value was used for the overall reliability because the correlation was negative. Observed reliabilities for all individual difference measures were all reasonably high (see Table 30), but all observed correlations (see Table 29) were well below their respective maximum correlations due to the low reliability of the choke score (see Tables 18 and 19). Note that the extent to which measures may correlate with each other is limited by their reliabilities; measures with lower reliabilities will have smaller correlations.

Table 30

Observed Reliabilities for Choke Score and Individual Difference Variables Based on the Current Study Sample

Measure	Choke Score (Accuracy, cm)	Dispos. S-C	Private S-C	Public S-C	Social Anxiety	Trait Anxiety	Sports Anxiety (Overall)	Sports Anxiety (Somatic)	Sports Anxiety (Worry)	Sports Anxiety (CD)
Observed Reliabilities	-.119	.846	.702	.751	.834	.911	.911	.826	.906	.839

Table 31

Maximum Possible Correlations Between Choke Score (Accuracy, cm) and Individual Difference Variables Based on Observed Reliabilities.

Measure	Dispos. S-C	Private S-C	Public S-C	Social Anxiety	Trait Anxiety	Sports Anxiety (Overall)	Sports Anxiety (Somatic)	Sports Anxiety (Worry)	Sports Anxiety (Concentration Disruption)
Maximum Correlations for Choke Score (Accuracy, cm)	.317	.289	.299	.315	.329	.329	.314	.328	.316

Note. See Table 30 for observed reliabilities of each individual difference variable and see Table 29 to compare with observed correlations.

GENERAL DISCUSSION

This project focused on establishing an experimental paradigm for studying choking under pressure on the OSA platform while also replicating past results in traditional in-person settings using tasks and individual difference measures with successful histories. Unfortunately, the project failed to see expected results on multiple fronts. First, the pressure manipulations for both OSA and in-person platforms failed to induce the expected choking effect; for the golf putting task, pressure even resulted in an overall performance enhancement. Second, the difference score-based choking measure was revealed to be low in reliability, except for RT for MA problems. Third, the reliability estimates for the individual difference variables for the MA study were substantially lower than for the golf putting study, despite the personality questionnaires being presented on an OSA platform for both studies (the tasks themselves were presented in OSA and in-person, respectively). Fourth, a few correlations were found between the choke score and individual difference measures, but this relationship should be interpreted with caution given the lack of clear results elsewhere. Given the overall null effect of performance decrements from the pressure manipulations, there was evidence for a somewhat predictable speed-accuracy trade-off.

Regarding the failure of the pressure manipulations to induce performance decrements, there are several things to note. It appears that the perceived pressure of the situation, scored based on combining participants' answers about how much pressure they perceived during pressure trials and how important they felt it was to do their best, is correlated with belief that the pressure manipulations are real but only moderately so and not correlated with performance decrements whatsoever. This suggests that while believing the pressure manipulations and feeling increased perceived pressure appear associated, it is possible to experience one without

the other. This seems to contradict most real-world situations, in which the two are often intertwined. This also suggests that the feeling that there is something real at stake and the feeling of being under pressure may play an important role in producing chokes (or lack thereof). Regarding belief in the pressure manipulations, while the golf putting study had all the advantages of a human interactions to sell the cover story, the simplistic design of the MA study may have been counterproductive. The instructions provided to the participants in the OSA studies were thorough but also basic and required participant engagement for comprehension and compliance, which may have been inconsistent or lacking in an undergraduate participant pool. Participants may need to be committed to the study above a certain threshold, which may be necessary for greater belief in the manipulations. This may be accomplished by creating more elaborately crafted instructions that include attention checks and incorporating more visually appealing user experience (UX) designs. Designing such a study would require a degree of creativity and UX expertise but is testable to determine if such changes would better entice undergraduate participants. Recruiting from the general population may also gather participants who are more enthusiastic about participation even in a more basic presentation.

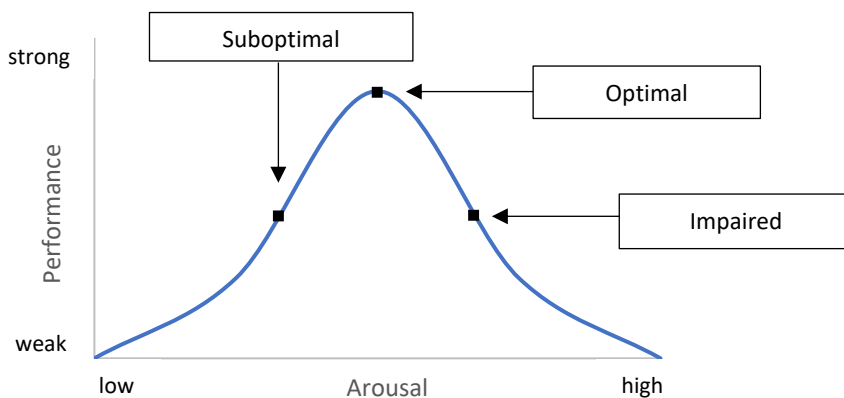
However, even the combination of increased perceived pressure and belief that there is something at stake also seems insufficient. On average, participants in the golf putting were high in both, yet RM-ANOVAs showed that they still performed better under pressure than during baseline conditions. RM-ANCOVAs also showed that the level of perceived pressure and belief did not covary with task performance, suggesting that there may be a missing piece to the choking puzzle. It is possible that performance may not be affected by any arbitrary increase in pressure and may only be impacted once a threshold has been met, which may individually vary.

Such results also underscore the rare of an incident choking is, and the challenges of capturing it under laboratory conditions.

Typically, pressure manipulations increase perceived pressure by offering participants incentives to increase their motivation by raising the importance of giving full effort towards performing the task as well as they can. This relationship can be plotted on a Yerkes-Dodson curve (see Figure 8), with stronger pressure manipulations affecting participants to increase their attention and interest in doing well on the task. Pressure manipulations that are strong, but insufficiently so may paradoxically lead participants to the height of the Yerkes-Dodson curve in which they are at optimal performance. Ideally, pressure manipulations will be strong enough to increase attention and interest beyond what is optimal and impair performance, labeling it as choking under pressure.

Figure 8

Hypothetical Yerkes-Dodson Curve of Performance Incentives Acting as a Pressure Manipulation.



It is also possible that the failure to induce a choking effect may not entirely be the fault of an insufficient pressure manipulation, but due to both the amount and design of the training, particularly for novice participants. In this respect, Beilock and Carr (2001) had participants perform 270 repetitions of a golf putting task during the training phase of the experiment after determining in pilot testing that task skill only seemed to level off after this amount of practice. The results showed that participants went from an average miss of about 28 cm in the first 18 practice trials to about 18 cm in the last 18 practice trials, which was an improvement of about 36%. This is consistent with the golf putting study here (see Figure 4), which showed improvement from an average miss of about 40.5 cm during Block 1 practice trials to an average miss of about 32 cm after a total of 180 trials, for an improvement of about 21%. Choking effects in Beilock and Carr's study did not emerge until a post-test after the 270 practice putts, and only appeared for participants who had practiced under video recording conditions designed to evoke explicit monitoring of their skill—precisely the type of pressure manipulation that was a part of the golf putting task in this project. The post-test manipulated pressure by invoking a team- and performance-based monetary incentive nearly identical to the one used in this study. Unlike this study, Beilock and Carr's participants continued to practice uninterrupted by an evaluation phase until after the training had concluded. The current study only utilized 180 putts total, had participants perform under pressure (including the use of a video recording manipulation and team- and performance-based monetary incentives) multiple times throughout the experiment, and averaged the pressure performance from these interspersed pressure trials. It should be noted that direct comparisons between this study and Beilock and Carr's should be made with caution, as the golf putting tasks slightly differed (Beilock and Carr had participants putt towards various targets at distances ranging from 4 to 5 feet, whereas the current study had participants putt

towards a single target 8 feet away). Nevertheless, it is plausible that the number of practice trials in the current study were insufficient, that the interspersed pressure manipulations may have inadvertently inoculated the participants from choking, and that any choking effect that could have still emerged may have been diluted by averaging scores from throughout the experiment. This also suggests that recruiting experts for experimental studies would typically yield more reliable and robust results than training novices, particularly university undergraduates. As pressure manipulations are typically adapted to maximally motivate the participants involved, testing experts, such as high-level athletes, while recording their performances and telling them that the footage would later be evaluated by their coach (with whom they have an established relationship with and would likely be personally motivated to impress) may incidentally lead to a far more robust pressure manipulation (Mesagno et al., 2011). Reliably motivating university undergraduates in a similar manner may necessitate creative ways to manipulate pressure, such as leveraging in-group/out-group dynamics to form teams competing against each other or offering incentives that may be more enticing such as increasing the monetary or course credit compensation.

Variability in task difficulty is not a pressure manipulation per se yet may also contribute to the overall lack of a pressure effect. In many real-life tasks, there are natural fluctuations in the degree of difficulty. For example, when taking a math test, students may solve a series of relatively easy problems before being faced with a much more difficult challenge. The variability of the difficulty, but more specifically the difficulty in anticipating these influxes may contribute to pressure in an unforeseen way. Research by Lyons and Beilock (2012) has shown that higher levels of math anxiety can activate pain-related activity in the brain in anticipation of difficult math problems but not during the solving of math problems per se. A baseline of difficulty-

matched word-task did not show the same results. It is unclear from this research if it was the mix of difficulties, which could lead to ambiguity on what to expect for each upcoming math problem, that contributed to the anticipation of pain or if the same results could be obtained from using problems of the same difficulty level (i.e., hard problems only). For choking studies using difference scores, easier problems did not result in choking, whereas harder problems did (Beilock & Carr, 2005, Beilock & DeCaro, 2007); if averaged together the overall choking effect may be diminished or even negated. If harder math problems that are included in mixed difficulty problem sets have any negative carryover effects of lowering overall math performance—meaning merely the anticipation of harder math problems can worsen performance during easier problems, then it is possible that a mixed difficulty task design would be superior to a uniform difficulty design. This notion presumes that the pain-response in *anticipation* of difficult math problems but not the *performance* of solving of the math problems themselves can impact the performance itself. Despite these potential complications, variations in task difficulty may be worthy of future investigation.

It is also possible that the overall null effect of pressure may be due to individual differences in how participants responded to the pressure manipulations. Miller and Schwarz (2018) point out that while a null effect may be just that, it is also possible that a manipulation could affect individuals in opposing directions, thus canceling the effect at the group level. In a choking study such as this one with many moving parts, such as attempting to train participants to a sufficient skill level and participants engaging in speed-accuracy trade-offs, it is possible that differences in how participants reacted to the pressure manipulations may be hiding an overall effect of the manipulations. Accounting for this possibility will be an important consideration moving forward.

Difference scores are notorious for having low reliability, and this reputation appears well-earned. For modular arithmetic, the difference-based choke score for accuracy had low split-half reliability, but a similarly calculated choke score for RT had high reliability. A notable difference between the two measures is in their precision and consequently the range in which participants could differ in their performance was far greater for RT (measured in time, down to milliseconds) than accuracy (measured in percentages for number of correct out of up to 24 problems). For example, as accuracy was calculated as a percentage, that percentage was based on up to 24 problems at a time (e.g., 24 odd baseline MA problems), meaning that the percentage could only vary as a dividend of 24, increasing variability. Therefore, a participant who got 21 out of 24 problems correct would have an accuracy of 87.5%, whereas a participant who got 22 out of 24 problems would have an accuracy of 91.7%. For RT, participants can vary much more precisely, such as 500 ms and 501 ms. However, it should be noted that while the precision of the measurements may be hampered by range restriction, this does not necessarily mean that the distribution of those measurements cannot vary sufficiently. This means that even if participants can only be distinguished in accuracy on a measure that is precise up to 24 units (e.g., 87.5% vs. 91.7% for accuracy), they may still meaningfully vary in accuracy among those 24 units and if so, then there is no issue with range restriction with respect to reliability. This is an important distinction, as only the latter is concerned with reliability; range restriction there can result in an inaccurate estimate of reliability (Sackett et al., 2002). Still, if most participants performed similarly in accuracy due to the restricted range, then the reliability would be compromised due to low variability. RT, however, may vary far more, as it is far less likely that multiple participants would get the same exact RTs; even if intra-individual variability was relatively stable (as in one participant consistently performed exactly 1 second slower on pressure trials),

inter-individual variability may be more likely to vary (that same exact magnitude of performance decrement would be unlikely to be repeated by other participants). For golf putting, the low reliability suggests that there was little that was systematic in the variability of the magnitude of people's response to the pressure manipulations (i.e., how badly they choked in response to the pressure manipulation). Here, the importance of sufficient practice cannot be understated, as improving baseline performance to asymptote removes variability of choke scores related to skill acquisition and can provide a purer measure of the influence of the pressure manipulations on performance. Deviations from baseline can be more readily interpreted as due to choking, provided that any performance decrements are accompanied by an increase in perceived pressure and belief in the pressure manipulations.

Although this project focused on internal consistency reliability for mean performance differences between baseline and pressure conditions, it is possible that pressure could also impact performance variability. In this sense, reduced precision (i.e., increased performance variability and hence decreased internal consistency of the choke score) could be the result of difficulty in regulating typically refined control structures under pressure. An example of this situation would be if variability during a series of baseline trials were low (suggesting competence in a task) and variability in a series of pressure trials immediately following them were high (suggesting a choke). Looking at the variabilities within the series of baseline and pressure trials that make up a choke score is also important, as novices in the process of skill acquisition could also have large fluctuations in performance as they refine their control structures, mimicking a choke. In such cases, novices may see that the variability in their performance decreases from baseline to pressure trials, as they continue to improve. Testing variability within difference-based choke scores as an indicator of choking is an interesting

direction but beyond the scope of this project, given how, overall, participants appeared to continue to improve over the course of the experiment and continued to decrease in variability.

Although much of the discussion has highlighted the importance of training participants to become “good enough” to choke, defining the threshold for such a skill level based on performance alone can lead to the circular logic of a participant choking because they are good enough to choke, and if they choke, then it was because they were good enough to choke. Fortunately, many researchers have identified expert-novice differences that indicate that experts and novices process tasks differently (Beilock & Carr, 2001; Hill et al., 2010). While the current project did not assess participants’ access to declarative knowledge during the golf putting task, doing so may be useful to cross compare with performance outcomes to see if novices have been trained to a sufficient level to choke, similar to how this study assessed the effectiveness of the pressure manipulations via a combination of choke scores, perceived pressure scores, and belief scores.

Given the inconsistent reliability of the difference-based choke scores and the pressure manipulations not having their expected effects, having only a few significant correlations was surprising but consistent with existing literature and expectations prior to data collection. It is important to consider that reliability estimates are just that—estimates, and therefore may be prone to error, just as any other statistic is. Presumably, greater correlations could be obtained with more robust instances of choking under pressure, which can be accomplished via various study designs (e.g., studying experts, stronger pressure manipulations, etc.). For example, Wang et al. (2004) found correlations between choking under pressure and private self-consciousness (-.49) and somatic trait anxiety (-.30) when they recruited experienced basketball players in a free throw shooting task and had a live audience as part of the pressure manipulation. It should

be noted that despite complications, these correlations are fairly close in magnitude to those observed in the golf putting study, which found correlations between choke scores and overall sports anxiety (-.205) and sports anxiety worry (-.279). This indicates that while the overall impact of the pressure manipulations was to unexpectedly enhance performance for the golf putting study, participants who reported greater sports-related anxiety did not experience as much of that improvement as those with lower anxiety.

Nevertheless, it is concerning that individual difference variables had inconsistent reliabilities despite being presented in a nearly identical format. Such a result underscores the importance of a “sanity check” for an individual differences study to ensure that correlations are not undermined by unexpected factors. Given how the personality questionnaires were given at the start of the study and in OSA format for both studies, it remains unclear if the uneven responses are due to how the task themselves are presented (i.e., OSA vs. in-person). Somewhat relatedly, assessing additional relevant psychological correlates such as math-based anxiety for math-based tasks may uncover individual differences with great real-world relevance.

Sample size may be an important consideration in the service of stabilizing correlations, especially in individual differences studies. Schönbrodt and Perugini (2013) showed via Monte Carlo simulations that achieving stable estimates for correlations depends on the effect size, the tolerable “corridor” of deviation around the true correlation, and the confidence level for that corridor. The authors note that for an assumed effect size of .21, a confidence level of 80% (typical for a power analysis), and acceptable fluctuations in the observed effect size at $<.10$, the required sample size would be 238. For larger assumed effect sizes such as .4, which can serve as an approximation to Wang et al. (2004) findings, the required sample size becomes 181. Requesting a higher confidence level of 90% would increase the required sample size to 260.

These sample sizes become prohibitively difficult to obtain for in-person studies, but such studies may also potentially benefit from larger effect sizes due to greater salience of the pressure manipulations afforded by human interaction. In contrast, OSA studies may produce smaller effect sizes but offer much greater access to the required sample sizes. Ultimately, careful consideration of sample size offers an additional way to tighten the design of experimental choking studies.

Speed-accuracy trade-offs are a common concern among psychology studies (DeCaro et al., 2010; Beilock et al., 2004). In the OSA study using the MA task, participants were instructed to prioritize both to maximize performance (see Appendix A8). Still, participants may opt to prioritize one over the other, including for strategic considerations (Förster et al., 2003). While the exact reasons that participants engaged in speed-accuracy trade-offs is unclear, it appears that among participants who had negative choke scores (i.e., had performance decrements from baseline to pressure trials), the tradeoffs occurred both ways. Participants who had negative accuracy-based choke scores had correlations suggesting they prioritized speed over accuracy and vice versa for participants who had negative RT-based choke scores. This makes intuitive sense, but evidence for speed-accuracy trade-off was only partially supported among participants who had positive choke scores (i.e., had performance improve from baseline to pressure trials); namely participants who had positive accuracy-based choke scores but not among those who had positive RT-based choke scores. Most participants (128 out of 144 participants) did not have negative choke scores in both accuracy and RT, yet both choke scores were still negatively correlated. Speed-accuracy trade-offs can lead to greater variability in the measure and consequently lower reliabilities. Furthermore, task performance continued to improve throughout the experiment and did not plateau. In sum, there may be an additional wrinkle in interpreting

both the success and failure of the pressure manipulations as being purely from choking effects (or lack thereof) or potentially also being influenced by speed-accuracy trade-offs.

Choking research may be especially susceptible to the file drawer effect. The file drawer effect refers to a problem in which null findings are often under-reported in science. Such an effect could be hiding the fact that experimentally inducing choking under pressure in novice participants, particularly university undergraduates as alluded to earlier, may be more difficult than might initially appear from the literature. Also, choking studies often study more than just the pressure manipulations per se and justifying publication of a null result stemming from a lack of a choking effect may be even more difficult if the “true” research questions could not even be addressed. Relatedly, even published studies that report significant findings may warrant additional scrutiny in the form of power analysis to ensure that the findings in the current literature stand on solid statistical grounding.

Choking under pressure is common in the everyday world. This project revealed some challenges in studying this phenomenon in experimental settings but also some potential explanations for understanding how some studies succeed in observing a choking effect while others do not. This project provides some basis for future developments, including testing a more robust pressure manipulation, increasing training duration, and recruiting expert participants. One possibility has been explored and is currently under analysis. Participants completed the MA task over the OSA platform and were shown a countdown timer based on their baseline performance during their pressure trials in addition to the team-based monetary incentive and screen recording manipulations from Study 1. It remains to be seen if this will lead to choking under pressure and help answer some of the questions raised in the current project.

Ultimately, the pursuit of perfecting experimental paradigms that can consistently generate a choking effect in a variety of domains may lead to testing for the domain generality of choking. However, it appears at least for the moment, it is important to re-consider how choking under pressure is studied within individual tasks before addressing this over-arching possibility. The heterogenous design of this project (i.e., online, self-administered platform using a modular arithmetic task and an in-person platform using a golf putting task) was instrumental in both serving as a robust test of choking as an individual differences construct and to serve as an important first step of replicating the choking effect independently in different domains. In the future, efforts like this may pave the way to then test individual participants across the multiple different domains.

FOOTNOTES

¹As a result of five MA problems being inadvertently duplicated, participants solved 91 unique MA problems, which were unevenly distributed across the baseline and pressure trials given that the problems were presented in random order for each participant. The second occurrence of each duplicated problem was thrown out from analysis. The duplicated problems were noted by some participants in the post-experiment comments. Duplicate problems were statistically controlled for and did not fundamentally alter any results.²

² The pressure manipulations were assessed by computing a 2 x 3 RM-ANOVA for accuracy in baseline and pressure conditions across 3 baseline-pressure blocks with the first trial a duplicate item appeared during the task as a covariate. This accounted for participants potentially believing that the task was about memorizing answers to identical problems and not about doing mental arithmetic, which would have fundamentally been different from the target cognitive mechanisms that are associated with choking. There was a significant effect of the pressure manipulations on accuracy for block and an interaction ($F(1,1) = 4.775, p = .031$; $F(1,1) = 5.593, p = .038$). The duplicate item covariate marginally missed significance for block and interaction effects ($F(1,1) = 3.862, p = .051$; $F(1,1) = 3.667, p = .057$). These results did not contradict the pattern of results observed earlier, indicating that the pressure manipulations were not successful in inducing performance decrements.

REFERENCES

- Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education.
- Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of experimental psychology: General*, 130(2), 224.
- Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, 44(1-2), 75-106.
- Balk, Y. A., Adriaanse, M. A., Ridder, D. T. D. de, & Evers, C. (2013). Coping under pressure: Employing emotion regulation strategies to enhance performance under pressure. *Journal of Sport and Exercise Psychology*, 35(4), 408–418. <https://doi.org/10.1123/jsep.35.4.408>
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3), 610–620. <https://dx.doi.org.proxy1.cl.msu.edu/10.1037/0022-3514.46.3.610>
- Baumeister, R. F., & Showers, C. J. (1986). A review of paradoxical performance effects: Choking under pressure in sports and mental tests. *European Journal of Social Psychology*, 16(4), 361–383. <https://doi.org/10.1002/ejsp.2420160405>
- Beilock, S. L. (2007). Understanding skilled performance: Memory, attention, and ‘choking under pressure’. *Sport & exercise psychology: International perspectives*, 153-166.
- Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, 130(4), 701.
- Beilock, S. L., & Carr, T. H. (2005). When High-Powered People Fail: Working Memory and “Choking Under Pressure” in Math. *Psychological Science*, 16(2), 101–105. <https://doi.org/10.1111/j.0956-7976.2005.00789.x>
- Beilock, S. L., & DeCaro, M. S. (2007). From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), 983.
- Beilock, S. L., & Gray, R. (2012). Why do athletes choke under pressure? In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology*, 425–444. <https://doi.org/10.1002/9781118270011.ch19>
- Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General*, 133(4), 584.
- Case, R., & Globerson, T. (1974). Field independence and central computing space. *Child development*, 772-778.

- Cowan, Nelson, et al. "On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes." *Cognitive Psychology* 51.1 (2005): 42-100.
- DeCaro, M. S., Rotar, K. E., Kendra, M. S., & Beilock, S. L. (2010). Diagnosing and alleviating the impact of performance pressure on mathematical problem solving. *Quarterly Journal of Experimental Psychology*, 63(8), 1619-1630.
- DeCaro, M. S., Thomas, R. D., Albert, N. B., & Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. *Journal of Experimental Psychology: General*, 140(3), 390–406. <http://dx.doi.org.proxy2.cl.msu.edu/10.1037/a0023466>
- Eysenck, M. W. (1992). *Anxiety: The cognitive perspective*. Hillsdale, NJ: Erlbaum.
- Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 43(4), 522.
- Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns?. *Organizational behavior and human decision processes*, 90(1), 148-164.
- Gröpel, P., & Mesagno, C. (2017). Choking interventions in sports: A systematic review. *International Review of Sport and Exercise Psychology*. <https://doi.org/10.1080/1750984X.2017.1408134>
- Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert?. *Intelligence*, 45, 34-45.
- Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab.js: A free, open, online study builder. *Behavior Research Methods*, 1-18.
- Hill, D. M., Hanton, S., Fleming, S., & Matthews, N. (2009). A re-examination of choking in sport. *European Journal of Sport Science*, 9(4), 203-212.
- Hill, D., Hanton, S., Matthews, N., & Fleming, S. (2010). Choking in sport: A review. *International Review of Sport and Exercise Psychology*, 3, 24–39. <https://doi.org/10.1080/17509840903301199>
- Hill, D. M., Carvell, S., Matthews, N., Weston, N. J., & Thelwell, R. R. (2017). Exploring choking experiences in elite sport: The role of self-presentation. *Psychology of Sport and Exercise*, 33, 141-149.
- Jackson, R. C. (2013). Babies and bathwater: commentary on Mesagno and Hill’s proposed re definition of ‘choking’. *International Journal of Sport Psychology*, 44(4), 281-284.
- Kayikcioglu, O., Bilgin, S., Seymenoglu, G., & Deveci, A. (2017). State and trait anxiety scores of patients receiving intravitreal injections. *Biomedicine hub*, 2(2), 1-5.

- Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2003, September 30). *Psychological research online: Opportunities and challenges*. American Psychological Association. <http://www.apa.org/science/leadership/bsa/internet/internet-report>
- Lewis, B. P., & Linder, D. E. (1997). Thinking about Choking? Attentional Processes and Paradoxical Performance. *Personality and Social Psychology Bulletin*, 23(9), 937–944. <https://doi.org/10.1177/0146167297239003>
- Lyons, I. M., & Beilock, S. L. (2012). When math hurts: math anxiety predicts pain network activation in anticipation of doing math. *PloS one*, 7(10), e48076.
- Masaki, H., Maruo, Y., Meyer, A., & Hajcak, G. (2017). Neural Correlates of Choking Under Pressure: Athletes High in Sports Anxiety Monitor Errors More When Performance Is Being Evaluated. *Developmental Neuropsychology*, 42(2), 104–112. <https://doi.org/10.1080/87565641.2016.1274314>
- Mascarenhas, D. R., & Smith, N. C. (2011). Developing the performance brain: Decision making under pressure. *Performance psychology—A practitioner's guide*, 245–267.
- Masters, R. S. W., Polman, R. C. J., & Hammond, N. V. (1993). ‘Reinvestment’: A dimension of personality implicated in skill breakdown under pressure. *Personality and Individual Differences*, 14(5), 655–666. [https://doi.org/10.1016/0191-8869\(93\)90113-H](https://doi.org/10.1016/0191-8869(93)90113-H)
- Mesagno, C., Geukes, K., & Larkin, P. (2015). Choking under pressure: A Review of current debates, literature, and interventions. In *Contemporary Advances in Sport Psychology: A Review* (pp. 148–174).
- Mesagno, C., Harvey, J., & Janelle, C. (2011). Self-presentation origins of choking: Evidence from separate pressure manipulations. *Journal of Sport & Exercise Psychology*, 33, 441–459. <https://doi.org/10.1123/jsep.33.3.441>
- Mesagno, C., & Hill, D. (2013). Definition of choking in sport: Re-conceptualization and debate. *International journal of sport psychology*, 44, 267.
- Miller, J., & Schwarz, W. (2018). Implications of individual differences in on-average null effects. *Journal of Experimental Psychology: General*, 147(3), 377.
- Omoregie, P. O., & Adegbesan, O. A. (2011). Effects of extraneous variables on performance of choking-susceptible university athletes. *IFE Psychologia: An International Journal*, 19(2), 75–91. <http://dx.doi.org.proxy2.cl.msu.edu/10.4314/ifep.v19i2.69513>
- Oudejans, R. R., & Pijpers, J. R. (2009). Training with anxiety has a positive effect on expert perceptual–motor performance under pressure. *Quarterly Journal of Experimental Psychology*, 62(8), 1631–1647.
- Oudejans, R. R., & Pijpers, J. R. (2010). Training with mild anxiety may prevent choking under

higher levels of anxiety. *Psychology of Sport and Exercise*, 11(1), 44-50.

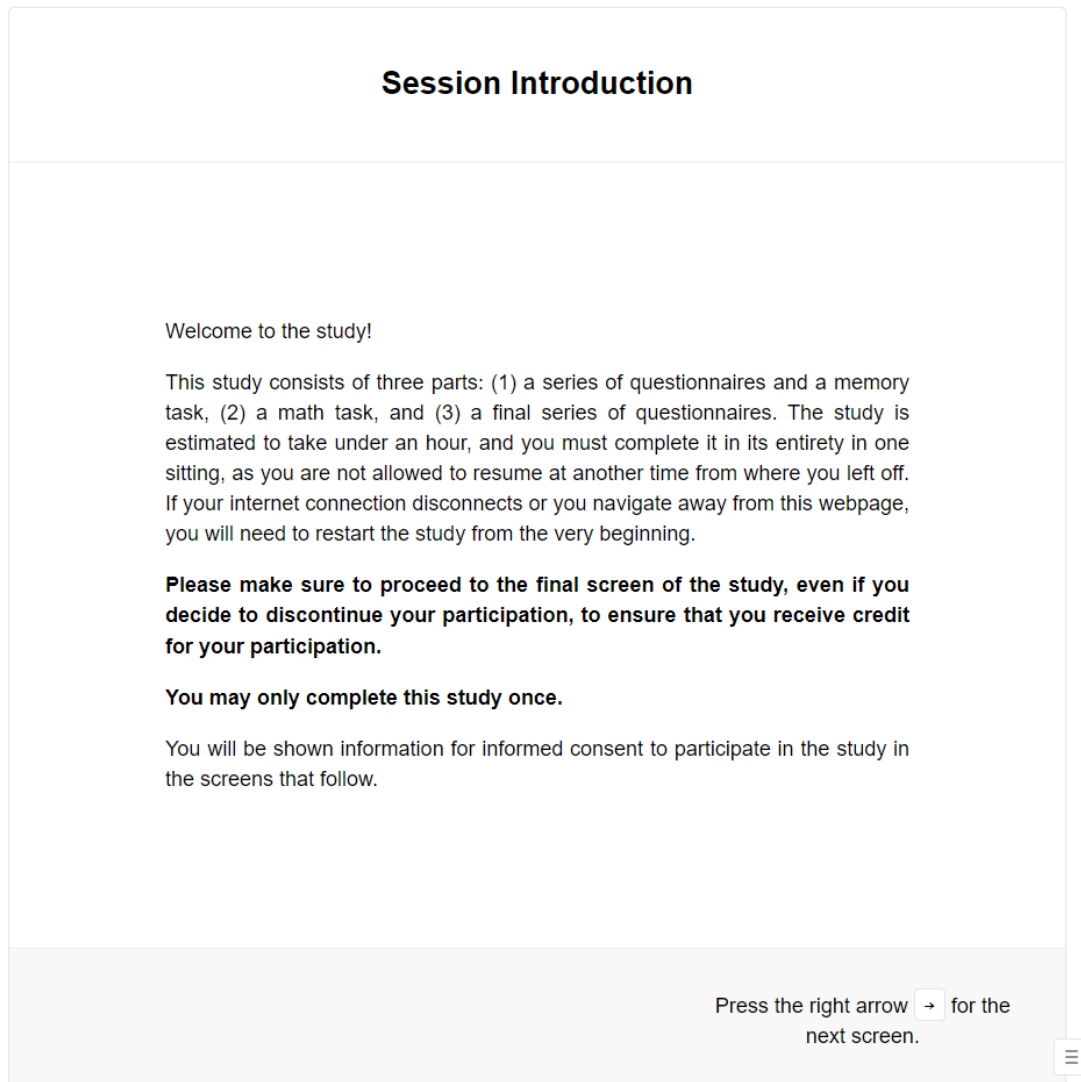
- Read, A., & Wert, K. (2021, December 6). *How states are using pandemic relief funds to boost Broadband Access*. The Pew Charitable Trusts. Retrieved May 4, 2022, from <https://www.pewtrusts.org/en/research-and-analysis/articles/2021/12/06/how-states-are-using-pandemic-relief-funds-to-boost-broadband-access>
- Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 43(3), 450–461. <http://dx.doi.org.proxy1.cl.msu.edu/10.1037/0022-3514.43.3.450>
- Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology*, 55(4), 807-825.
- Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, 27(3), 183-198.
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Smith, R. E., Smoll, F. L., Cumming, S. P., & Grossbard, J. R. (2006). Measurement of Multidimensional Sport Performance Anxiety in Children and Adults: The Sport Anxiety Scale 2. *Journal of Sport and Exercise Psychology*, 28(4), 479–501. <https://doi.org/10.1123/jsep.28.4.479>
- Spielberger, C. D. (1983). State-trait anxiety inventory for adults.
- Veldhuijzen Van Zanten, J. J., De Boer, D., Harrison, L. K., Ring, C., Carroll, D., Willemsen, G., & De Geus, E. J. (2002). Competitiveness and hemodynamic reactions to competition. *Psychophysiology*, 39(6), 759-766.
- Vock, M., & Holling, H. (2008). The measurement of visuo-spatial and verbal-numerical working memory: Development of IRT-based scales. *Intelligence*, 36(2), 161-182.
- Wang, J., Marchant, D., Morris, T., & Gibbs, P. (2004). Self-consciousness and trait anxiety as predictors of choking in sport. *Journal of Science and Medicine in Sport*, 7(2), 174–185. [https://doi.org/10.1016/S1440-2440\(04\)80007-0](https://doi.org/10.1016/S1440-2440(04)80007-0)
- Wang, Z., & Shah, P. (2014). The effect of pressure on high- and low-working-memory students: An elaboration of the choking under pressure hypothesis. *British Journal of Educational Psychology*, 84(2), 226–238. <https://doi.org/10.1111/bjep.1202>

APPENDIX A: Modular Arithmetic Task Instructions (OSA)

The following are screenshots of the instructions provided throughout the modular arithmetic task, taken directly from the OSA platform.

Figure A1

Instructions for Modular Arithmetic Task (OSA)



Note. Part 1 of 4 for instructions provided at the beginning of the modular arithmetic task to obtain participant consent.

Figure A2

Instructions for Modular Arithmetic Task (OSA)

Participant Consent

The purpose of this research study is to better understand how people learn a novel math task and how it may relate to certain psychological traits. You may be asked to (a) complete tests designed to measure psychological traits; (b) complete tests designed to assess your prior experience with modular arithmetic and alphabet arithmetic; (c) complete a brief modular arithmetic task; (d) complete a brief alphabet arithmetic task; (e) improve your performance on these tasks with the possibility of monetary rewards for superior performance. The research study will occur in a single session lasting under an hour.

1. Your participation in this research study is voluntary. You may choose not to participate at all, or you may refuse to complete certain tests or answer certain questions, or to discontinue your participation at any time without penalty or loss of benefits.
2. You will be compensated with course credit (1 credit per half hour of participation) for participating in this research study. You will also be eligible to earn additional financial rewards for sufficiently improving your performance on the tasks. Under no circumstances are you required to participate in research to fulfill a course requirement. You may complete the alternative assignment that your instructor has made available to you.

Press the left arrow  for the previous screen.

Press the right arrow  for the next screen.



Note. Part 2 of 4 for instructions provided at the beginning of the modular arithmetic task to obtain participant consent.

Figure A3

Instructions for Modular Arithmetic Task (OSA)

Participant Consent

3. You may experience mild levels of mental fatigue while participating in this research study. Risks associated with participation in this research study will not exceed those of typical classroom instruction.

4. Your privacy will be protected to the maximum extent allowable by law, and data that you contribute in the research study will be kept strictly confidential. Only David Z. Hambrick, Ph.D., a limited research staff, and the Michigan State University Institutional Review Board will have access to your test forms. Data that you contribute will be stored on password-protected computers.

5. You will not directly benefit from your participation in this research study beyond the possibility of the monetary rewards. However, your participation in this research study may contribute to the understanding of factors that influence success in the performance of complex tasks.

6. The true nature of this research will be provided after your completion of the task. You may also ask any question you have at that time.

Press the left arrow  for the previous screen.

Press the right arrow  for the next screen.



Note. Part 3 of 4 for instructions provided at the beginning of the modular arithmetic task to obtain participant consent.

Figure A4

Instructions for Modular Arithmetic Task (OSA)

Participant Consent

If you have questions or concerns regarding this research study, please contact:

David Z. Hambrick, Ph.D.
Professor of Psychology
Psychology Department, Room 298B
Michigan State University
East Lansing, MI 48824
Phone: (517) 377-2195
e-mail: hambric3@msu.edu

You indicate your voluntary agreement to participate by beginning this research.

You will be given a copy of this consent form for your records if you wish.

Press the left arrow for the previous screen.

Press the to continue.

Note. Part 4 of 4 for instructions provided at the beginning of the modular arithmetic task to obtain participant consent.

Figure A5

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

You will now be asked to learn how to solve an unusual type of math problem. Indicate if the following math equations involving modular arithmetic are *True* by pressing the key or *False* by pressing the key.

Please complete all problems in the experiment on your own and without any outside help, including the use of your phone, calculators, scrap paper, etc.

Press the right arrow for the next screen.

Note. Part 1 of 3 for instructions provided at the beginning of the practice block for the modular arithmetic task.

Figure A6

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

Moduar arithmetic problems take the following form and require a multi-step approach:

$$7 \equiv 4 \pmod{3}$$

First, you must take the first value, 7, and subtract from it the second value, 4, which comes after the operator, $\equiv$. Then, you must divide that number by the mod value, $\pmod{3}$.

This would lead us to solve the problem in the following way:

$$7 - 4 = 3$$
$$3 \div 3 = 1$$

Because 1 is a whole number, you would press z to indicate that the problem is true.

Press the left arrow ← for the previous screen.

Press the right arrow → for the next screen.

Note. Part 2 of 3 for instructions provided at the beginning of the practice block for the modular arithmetic task.

Figure A7

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

Here is another example:

$5 \equiv 3 \pmod{3}$

For this problem, you would press **m** to indicate that the problem is false. This is because:

$5 - 3 = 2$

$2 \div 3 = 0.67$

Because **0.67** is **not** a whole number, this problem is false.

You will now solve practice problems. Feedback will be provided to indicate whether you solved the problem correctly or incorrectly. These problems will be harder than the examples.

There are four different problems. You must solve all of them correctly in a row to proceed to the next section.

Press the left arrow **<** for the previous screen.

Press the **space bar** to continue.

Note. Part 3 of 3 for instructions provided at the beginning of the practice block for the modular arithmetic task.

Figure A8

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

You will now receive an additional set of practice problems to help you learn how to solve modular arithmetic problems to the best of your ability. Try to solve the problems as quickly as possible without sacrificing accuracy.

To measure your progress, you will be tested twice during the course of the experiment. Based on your performance in these tests, you will have the opportunity to earn a reward based on your speed and accuracy. The details of the tests and the exact criteria for earning the reward will be explained to you after you complete this set of practice problems.

Press the space bar to continue.

Note. Instructions provided at the beginning of the baseline trials in the first block for the modular arithmetic task.

Figure A9

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

Now that you have completed a set of practice problems, you will take the first of two tests to measure your progress in learning how to solve modular arithmetic problems. In each of these tests, you will receive a set of problems similar to those in the practice set you just completed. However, unlike in the practice set, a **performance score** will be calculated using a special formula that equally takes into account both your speed and accuracy in solving the problems. Therefore, to achieve the highest performance score, you must solve these problems as quickly as possible without sacrificing accuracy. Your final performance score will be based on both tests.

You will have an opportunity to earn \$5 based on your final performance score. However, all participants in this study have been randomly paired up into two-person teams. To receive the \$5 reward, the **average** of your final performance score and your partner's final performance score must be within the top 20% of all teams in this study.

Your partner has already completed the study and their performance score is currently within the **top 18%**. Now it is up to you to secure a top 20% placement so that you both will receive the \$5 reward. If you fail to finish within the top 20%, neither you nor your partner will receive the reward.

Press the right arrow → for the next screen.

Note. Part 1 of 3 for instructions provided at the beginning of the pressure trials in the first block for the modular arithmetic task.

Figure A10

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

During the tests, your performance will also be recorded via a screen-capturing software as part of the MSU Psychology Department's new initiative on re-assessing how math skills are taught to undergraduates. This software will only capture the experiment. Your data will be analyzed by psychologists in consultation with experts in math education to assess the abilities of current students and decide whether adjustments need to be made to the curriculum.

You will see that a red dot is present on the top right corner of the screen. This red dot indicates that your performance is being recorded for analysis. When the dot is no longer present, that indicates the recording software is not activated and you are not being recorded. You will see an example on the next screen.

As before, try to solve these problems as quickly as possible without sacrificing accuracy.

Press the left arrow  for the previous screen.

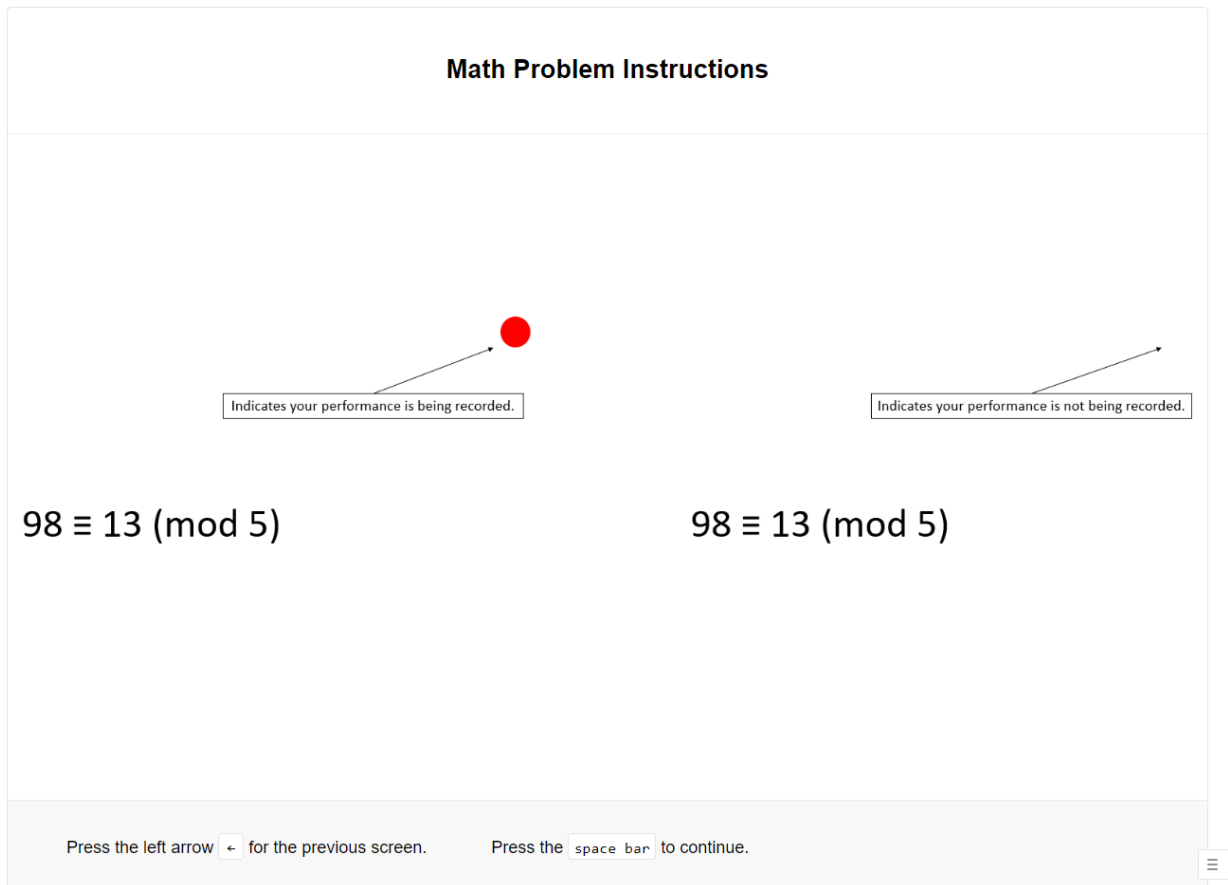
Press the right arrow  for the next screen.



Note. Part 2 of 3 for instructions provided at the beginning of the pressure trials in the first block for the modular arithmetic task.

Figure A11

Instructions for Modular Arithmetic Task (OSA)



Note. Part 3 of 3 for instructions provided at the beginning of the pressure trials in the first block for the modular arithmetic task.

Figure A12

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

You will now solve one more set of practice problems to help you continue to learn how to solve modular arithmetic problems to the best of your ability. Try to solve the problems as quickly as possible without sacrificing accuracy.

To measure your progress, you will be tested for a final time after you complete this set of practice problems.

Press the space bar to continue.

Note. Instructions provided at the beginning of the baseline trials in the second block for the modular arithmetic task.

Figure A13

Instructions for Modular Arithmetic Task (OSA)

Math Problem Instructions

You will now be tested for a final time to measure your progress in learning how to solve modular arithmetic problems. Your performance is based on both the speed and accuracy at which you solve these problems.

As explained before, you will earn a \$5 reward if the average of the final performance scores of both you and your partner fall within the top 20% of all teams tested. Again, your partner has already completed the experiment and their performance score is currently in the **top 18%** of all participants. Now it is up to you to secure a top 20% finish for your team. If your team's average performance score is not within the top 20%, neither you nor your partner will receive the \$5 reward.

As indicated by the red dot, your performance is being recorded for analysis by the MSU Psychology Department and by outside experts to assess the abilities of current students to decide whether adjustments need to be made to the curriculum.

Press the **space bar** to continue.

Note. Instructions provided at the beginning of the pressure trials in the second block for the modular arithmetic task.

Appendix B: Golf Putting Task Instructions (In-Person)

The following are excerpts from the Experimenter's Handbook that includes the scripts that Experimenter A followed when interacting with participants throughout the study.

Welcome Script

**Learn this and speak to participants conversationally (it does not have to be verbatim) when they come into the lab but before you start anything.*

“Thank you for agreeing to participate in our experiment. Today, we’re going to have you practice about 180 golf putts to see how well you’re able to learn and improve. We’ll break the session up into thirds, so you’ll have a short break after each set of about 60 putts. Just a reminder, your participation is voluntary. If you have any questions or if you no longer want to continue the experiment, please let us know and we can stop.”

Task Instruction Script: Pre-Training Phase

**Learn this and speak to participants conversationally to begin the experiment.*

“To start, I need you to watch a brief YouTube video on how to putt. [Show YouTube video on laptop.] Now, I’ve got both right-handed and left-handed putters, so please select one that suits you best. [Hand participant the putter that they indicate.] Before we begin the actual practice, please take a few putts to get a feel for things. Your goal is to aim at the target marker and make the ball stop as close to the marker as you can—right on the marker is perfect. These putts won’t count for anything.” [Stand to the side of target marker, away from the participant’s field of view. For more details, see “Golf Putt Measuring Protocol.” Have participant take more putts even if they indicate they are ready if they are bouncing the putts off the wall(s). Participants must make two successful non-bounce out putts in a row before they are considered ready for the training phase. This criterion does not need to be made explicit unless the participant repeatedly fails the task and, in that case, let the participant know in a friendly manner to avoid pressuring them.]

Task Instruction Script: Training Phase

**Learn this and speak to participants conversationally to begin the experiment.*

“Now we’re going to have you take 40 practice putts. Remember, your goal is to aim at the target marker and make the ball stop as close to the marker as you can—right on the marker is perfect. We’ll be measuring your putts, to gauge your performance. To measure your progress, you’ll be tested three times

during the experiment. You'll have a chance to earn a reward if you do well during these tests, which we'll explain later. Just focus on getting as good as you can on this task for now." [Measure the putt distance (see "Golf Putt Measuring Protocol").]

Task Instruction Script: Pressure Manipulation Phase

**Learn this and speak to participants conversationally after they have completed the first block of practice putts but before they begin their first block of evaluation putts.*

"Now that you've practiced some, we're going to have you take 20 putts to evaluate your current skill. We'll be using these 'money balls.' [Swap out the plain 'practice' balls for the yellow 'evaluation' balls.] As part of evaluating your performance, we're going to capture these 'evaluation putts' on camera for later analysis at an upcoming sports psychology meeting to be sponsored by the Kinesiology Department. We work with a golf coach and a golf pro to do this. [Prepare the camera (see "Camera Preparation Protocol").] As an incentive during these putts, you'll have the opportunity to earn extra rewards. There will be three ways to earn a reward. Two are based on teamwork; we've randomly partnered you with another participant for this. If the average performance between the two of you during these evaluation putts is in the top 20% of all the participants we've tested, then you'll both be contacted later and receive \$5 each. You'll also get a chance to earn an additional \$10 if you or your partner managed to get the closest putt to the target. And finally, our golf coach and golf pro will select up to 5 participants who they deem have made the most improvements from baseline. Your performance will obviously be considered but they will also consider things like form and your concentration. Don't worry, they know that most people are inexperienced or novices. If you're one of the lucky few to be selected, then you'll also win \$5. You'll have two more blocks of practice and evaluation putts, so you'll have more opportunities. The good news is that your partner has already completed this experiment before you and currently is sitting inside the top 20%. Of course, that does mean now it's all up to you!" [Measure the putt distance (see "Golf Putt Measuring Protocol"), if the participant asks how the other participant did, then tell them that they did well but we can't say how well, just focus on your own performance.]

Between-Tasks Script: Break Phase

**Learn this and speak to participants conversationally each time they complete a practice and evaluation block (40 practice putts + 20 evaluation putts = 60 putts).*

"You've completed a portion of the experiment. You are free to take up to a 5-minute break before moving on to the next set of practice putts. Please let us know when you are ready to continue."

Task Instruction Script: Repeat-Training Phase

**Learn this and speak to participants conversationally after they have completed the first block of evaluation putts but before they begin their second block of practice putts.*

“Now we’re going to have you take practice putts again.” [Measure the putt distance (see “Golf Putt Measuring Protocol”). You may indicate which practice block they are on (e.g., 2 of 3).]

Task Instruction Script: Repeat-Pressure Manipulation Phase

**Learn this and speak to participants conversationally after they have completed the second block of (40) practice putts but before they begin their second block of evaluation putts.*

“Now we’re going to have you take evaluation putts again to gauge your current skill. [Swap out the white ‘practice’ balls for the yellow ‘evaluation’ balls.] Remember, we’ll be recording your performance for later analysis [prepare the camera (see “Camera Preparation Protocol”)] and you have the opportunity to earn extra money for you and your partner if you putt well!” [Measure the putt distance (see “Golf Putt Measuring Protocol”). You may indicate which evaluation block they are on (e.g., 2 of 3).]

End of Experiment Script: Manipulation Check and Debrief

**Learn this and speak to participants conversationally after they have completed all practice and evaluation putts.*

“Thank you for participating in our experiment. I’m now going to ask you a few questions about your experience during the experiment.” [Have the participants fill out the manipulation check on a laptop. Fill in information via Qualtrics survey available [here](#).]

“Now that the experiment has officially concluded, we want to inform you about the true purpose of this experiment. The experiment was designed to understand how people perform under pressure and if those differences were related to certain psychological traits. To this end, it was critical that we made you feel that there was something real at stake. In reality, the things we told you during the evaluation putts were not true—the camera was not recording, and you did not have a partner who was relying on you to improve your performance. That said, you will still receive a \$5 reward as part of participating in our study. It is important that if other people knew the true purpose of the study, it might affect how they perform, so we are asking you not to share the information we just discussed. I hope you enjoyed your experience today. If you have any questions later please feel free to contact our lab.” [Provide contact sheet]