8-2-2018

Measuring Time Perception: A Psychometric Analysis

Patrick Joseph McNicholas

MEASURING TIME PERCEPTION: A PSYCHOMETRIC ANALYSIS

by

Patrick J. McNicholas

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: School Psychology

The University of Memphis

August 2018

Abstract

In the literature, many methods have been used to assess the perception of time in individuals. The purpose of this study was to examine the psychometric properties of time estimation, production, reproduction, and discrimination tasks. Using psychometric terminology consistent with the Standards for Educational and Psychological Testing, this study utilized classical test theory (CTT) and item response theory (IRT) to assess the accuracy, consistency, stability, difficulty, and discrimination of the four aforementioned methods of estimating time perception in college students (*N* = 136). At present, the tasks were able to demonstrate both validity and reliability through the analyses conducted and meet these psychometric standards. By meeting these standards, the evidence suggests that these tasks are tapping into the true construct of time itself and doing so in a relatively consistent manner. This research could inform future studies that wish to include tasks that assess perception of time in individuals.

TABLE OF CONTENTS

List of Tables

Introduction

Experience of Time

As the Earth both rotates and circles the Sun, animals and plants alike adjust their routine based on this pattern. This interplay between astrological entities has led to a parallel biological process in organisms, *circadian rhythm*, a patterned oscillation of approximately 24 hours. Unsurprisingly, these solar patterns are also the basis of modern day timing and dating systems (i.e., 24-hour days and 365-day calendar years), represented by time-keeping mechanisms (e.g., watches and calendars).

Time affects all life forms in various ways. For humans, timing is an important piece of daily living. From smaller intervals of time, such as a sitting in a waiting room for a doctor's appointment, to longer intervals, such as scheduling your next check-up a year away, time is omnipresent. Due to the pertinent nature of timing, it is essential that humans have an adequate way to perceive time, also known as *temporal perception*. Although there is no known internal organ that senses time, per se, many theories have been proposed to explain temporal perception. Amidst the many theories to emerge in the literature, the majority follow a general theme and encompass common components.

More often observed in non-human animals, a model of an internal clock derived from the aforementioned cyclical nature of circadian rhythms has been proposed as a possible mechanism for perceiving time. However, when circadian timing was investigated more thoroughly in humans, there was a strong correlation only for hourly estimates but not for minute intervals (Aschoff, 1985). These findings not only suggest that there are different mechanisms underlying temporal perception, but the outcomes also act as an indication that a circadian clock is more useful for sleep-wake cycles (Wittman, 1999).

Circadian timing acts as a parsimonious theory for most animals, but scalar expectancy theory (SET), also referred to as scalar timing theory, also explains timing mechanisms in animals and humans. SET proposes that internal timing mechanisms play a role in relation to a reinforcement schedule (i.e., animals estimate time based on stimuli and their subsequent reinforcers; Gibbon, 1977; Gibbon, Church, & Meck, 1984). This expectancy of reward and the resulting errors in timing tend to follow a linear trend closely following Weber's law where the variability of the response will grow proportionally to the magnitude of the interval used (Weber, 1933). The linear relationship broadly accounted for by Weber's law has been shown to have mixed findings with humans, so although SET is applicable to humans at much smaller intervals of time, it is usually viewed as a supplement to other more complex theories (Allan, 1979; Allan, 1998; Hancock & Block, 2012).

One of these more complex theories is Treisman's (1963) internal clock model. This model includes four interrelated components: a pacemaker that produces pulses (sometimes referred to more tangibly as "clicks"), a counter that measures the pulses in length and position, a store that holds this accrued measurement, and a comparator that can retrieve the measured pulses to compare with current pulses and make a determination by relating this measurement to relevant responses. Treisman stated that the pacemaker's pulses speed up or slow down due to arousal. This statement was investigated and confirmed by later researchers, paving the way to other modifications of the model and how the components vary accordingly (Block & Zakay, 2006).

A fifth, more recently added component of Treisman's (1963) internal clock model is the attentional switch (Thomas & Brown, 1974). This attentional switch acts as a gate that opens or closes depending on the attention given to time; thus, this enhancement of the internal clock model has been labeled the attentional-gate model (AGM; Block & Zakay, 1996; Zakay & Block, 1995, 1996). The attentional gate's involvement in the model is perhaps best summed up with the adage that "a watched pot never boils" (Block, George, & Reed, 1980), as when one attends intently to time (watching a pot and waiting for it to boil), this duration is perceived as longer. Alternatively, those who briefly turn away—attention or otherwise—often find that they return to a boiling pot in seemingly no time. In the latter case, the attentional gate closes due to the lack of attending to time, preventing as many pulses to pass through to the accumulator (i.e., Treisman's counter), so a shorter duration is perceived. In contrast, when staring at the pot waiting for time to pass, the attentional gate is wide open with full attention to time, and a longer duration is perceived (Wittmann, 2013; Zakay, 1990, 1993).

In addition to attention, mental workload (viz. working memory) may play a role in how the time is perceived. Relating back to the attentional gate, when working memory is being exhausted on another concurrent task or the current task is more mentally demanding, there is less mental ability available to be directed to the attending of time. Additionally, the component that acts as a store of these pulses is synonymous to memory and is used either immediately in working memory or coded to long-term memory. The involvement of memory (specifically, retrieving from long-term memory) leads to another prominent model originally proposed by Ornstein (1969), the contextual-change model. In this model, it is environmental changes that lead to the perceived time that pass. For example, if many events happen within an interval, it perceived as a longer passage of time than the same interval with only one or two events (Block, 1979; Block & Gruber, 2014).

These theories and proposed models attempt to explain how time is perceived by humans from a cognitive perspective. Across them, time is treated as information that must be encoded and decoded using various components within the models. Largely, time perception is observed as a processing of this temporal information—in general terms, gathered from external stimuli and compared to internal working models. There is reason to believe that there is not just one mechanism for perceiving time, as many mechanisms are contingent on variables such as interval length and conflicting processing of nontemporal information (Zakay, 1990). This notion suggests that although these models are useful to conceptualize temporal information processing, there may not be one single model or mechanism to turn to (Mangels & Irvy, 2001). As theories develop and models are modified, concurrent research must facilitate these advances. Much is still to be unlocked regarding the phenomenon of time.

Interest of Perception of Time

What grounds interest in perception of time is its relevance to everyday life. As mentioned previously, many important activities are strongly linked to accurate timing, from daily bell schedules to that coveted yearly vacation. What accrues even more interest in studying the phenomenon of time is the interaction between perception of time and multiple personal characteristics. Thus, psychologists have been intrigued by perception of time and its effects on attitudes (DeWall, Visser, & Levitan, 2006), decisions (Wittmann & Paulus, 2009), and behavior (Perret-Clermont & Lambolez, 2005; Strathman, 2005). Parallel to this trend, researchers have also been interested in how perception of time is affected by variables such as individuals' body temperature (Wearden & Penton-Voak, 1995), routine (Avni-Babad, & Ritov, 2003), mental workload (Brown & Boltz, 2002), fear (Fayolle, Gil, & Droit-Volet, 2015), and emotion (Droit-Volet, Fayolle, Lamotte, & Gil, 2013). Emotion is a broad term that may refer to general mood or a specific state (e.g., depression). This continuum naturally guides researchers to examine individuals with extreme cases of these states, as that they are notably impaired.

Further, there is evidence that some individuals exhibit severe deficits in perceiving time relative to their peers (Allman & Meck, 2012). Consistent with the earlier reference to extreme cases of emotional states, people with mood disorders (including depression and anxiety; Bar-Haim, Kerem, Lamy, & Zakay, 2010; Droit-Volet et al., 2013) as well as schizophrenia (Carroll, Boggs, O'Donnell, Shekhar, & Hetrick, 2008; Clausen, 1950; Davalos & Opper, 2015) often demonstrate deficits in perception of time. In addition, there has been substantial research on perception of time in individuals with neurodevelopmental deficits and disorders such as Parkinson's disease (Buhusi & Meck 2005; Gu, Jurkowski, Lake, Malapani, & Meck, 2015), autism spectrum disorder (Allman & Falter, 2015; Szelag, Kowalska, Galkowski, & Pöppel, 2004), and attention-deficit hyperactivity disorder (ADHD; Meaux & Chelonis, 2003; Pollak, Kroyzer, Yakir, & Friedler, 2009; Toplak et al., 2006; Toplak & Tannock 2005).

In viewing ADHD, one might ascertain that those with ADHD have deficits in perceiving time, considering one of the criteria for the disorder is having poor time management as listed in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM–5*; APA, 2013). This connection is strengthened within the framework of the AGM (Zakay & Block, 1995). If an individual with ADHD has trouble focusing in general, surely this difficulty would translate to attention to time and underestimating how long a task might take along with passage of time. Thus, noting the timing deficits in these populations and potential implications of the outcomes (e.g., incomplete tasks due to poor time management, poor performance on timed tests, or tardiness to scheduled events), it is important to be able to assess for these deficits in a uniform way, in order to ultimately improve the quality of life among all populations most affected by said deficits.

Methods of Measuring Time Perception

To effectively assess deficits or differences in an individual's ability to process temporal information, a system to efficiently and accurately assess their perception of time is needed. Arguably, one of most well-cited publications focusing temporal perception from the early literature comes from William James (1890). Here, James makes an important distinction in temporal perception that has laid the groundwork for many of the time perception theories mentioned earlier as well as guided usage of the tasks developed in sections that follow.

There are two overarching categories of temporal perception that James (1890) referred to as a prospective and retrospective senses of time—experienced duration of time versus remembered duration of time. These concepts are explicitly linked to experimental paradigms in the literature: the *retrospective paradigm* and the *prospective paradigm* (James, 1890; Block, 1992, 1979; Block & Zakay, 2006; Zakay & Block, 1997). The latter paradigm is rooted in the participant's prior knowledge that they are going to judge a duration of time beforehand. A real-life example of this paradigm might be a professor asking a student to come back to their office in five minutes. At this juncture (without the aid of timekeeping mechanisms), the student must determine approximately when that duration has passed as to not bother the professor too quickly or miss the appointment. Conversely, the former paradigm requires that participants experience an elapsed time and be unaware that they will be asked to subsequently judge that duration. An example of this paradigm is asking a visitor how long a drive they recently completed was. Here, the visitor must think back and approximate the duration of their trip.

The most substantial problem with experiments that follow the retrospective paradigm is that once a participant is asked to judge one duration of time, the participant is then privy to the fact they may have to do so again, which negates the goal of the paradigm as they are now expecting to estimate time. Although the retrospective paradigm gives insight to unconscious perception of time—which may or may not lead to participants encoding the duration as time passing—it is much more difficult to complete elaborate experiments to assess multiple increments of time. Taking the delivery of the paradigm into context, if the experimenter is not actively controlling for this awareness of time (e.g., prompting participants to respond in seconds), the unspecified encoding of duration could result in an unclear relation to specific pathways. Additionally, results based on the retrospective paradigm may be less applicable in the real world, as time is typically asked to be estimated prior to the event rather than afterwards. For example, when planning for a trip, the visitor often preemptively judges the travel time as to have an accurate estimated time of arrival and not upset their hosts. With the retrospective paradigm giving way to various confounds, it adds to the difficulty of assessing perception of time. Thus, it has been employed far less often in the literature (Mangels & Irvy, 2001; Zakay & Block, 1997). For these reasons, this project primarily focused on measures consistent with the prospective paradigm (Block & Gruber, 2014).

There have been many methods of measuring time perception consistent with the prospective paradigm. Alan (1979) divided these methods into two categories: *duration scaling* (which includes production and estimation tasks) and *duration discrimination* (which contains comparison tasks). Some examples of the methods encompassed by duration scaling have been described as magnitude estimation, synchronization, ratio-setting, and category rating. Duration discrimination are methods of comparison that use single or dual stimuli comparison, as well as variations called forced-choice comparison and identification (Alan, 1979; Nichelli, 1996). Although Alan made these distinctions among the different methods within the two categories, Zakay (1990) argued that "this differentiation is justified only from a theoretical point of view for analyzing the different cognitive processes that are involved in each case. This differentiation, however, is a posteriori and if not justified from a practical point of view of an experimenter who is designing his or her experiment" (pp. 63-64).

Rather than elaborate and specify the many methods of temporal estimation, Zakay (1990) encouraged a parsimonious approach, alluding to the variation in experiments conducted in past research. Following this charge, his recommendation was to follow the literature's more basic categorization that was proposed by earlier studies (Bindra, 1956; Clausen, 1950; Eisler, 1976; Hicks, Miller, & Kinsbourne, 1976; Hornstein & Rotter, 1969; Siegman, 1962; Thomas & Brown, 1974; Treisman, 1963; Wallace & Rabin, 1960). These early studies tend to group the methods into four major methodological categories of (1) verbal (time) estimation, (2) time production, (3) reproduction (of time), and (4) discrimination (also referred to as comparison). These primarily psychophysical methods are typically represented by tasks considered to be temporal information processing tasks (TIPT; Toplak et al., 2006). These methods and examples of common tasks used are described in detail below.

Verbal estimation. Methods of verbal estimation—previously referred to as magnitude estimation (Nichelli, 1996)—are perhaps the most well-known and widely used tasks of assessing temporal perception (Block, Hancock, & Zakay, 2000, 2010; Block & Zakay, 1997; Thones, 2015). For these tasks, participants are exposed to a stimulus and explicitly asked to indicate how long they thought that stimulus was present, typically using a numbered time guess (e.g., 1 minute, 20 seconds, 500 milliseconds). Due to the estimation being originally offered verbally, the title “verbal estimation” stuck. In more recent studies, participants have provided estimates of the length of exposure to stimuli using pen and paper and electronic media.

Clausen (1950) used the verbal estimation method and asked patients with schizophrenia to estimate the duration that a lamp was lighted across varying intervals. Similarly, Hornstein and Rotter (1969) used a similar task that employed a small red light as the stimulus rather than a lamp. Across both studies, the time intervals were 2, 5, 8, 10, 11, 14, 5, 17, 20, 23, 26, and 29 seconds. In a more recent study that utilized the method of verbal estimation, Bauermeister et al. (2005) administered a verbal estimation task to children with and without ADHD in which they were instructed to report how long a flashlight was illuminated. The intervals used in this study were 6, 10, 13, 18, 25, and 33 seconds.

Time production. Time production methods require participants to generate a duration of time based on a given interval. Thus, these tasks provide prompts in the form of temporal intervals in numerical form that should be produced. For example, the participant is told to generate time for 5 seconds and then must physically generate a stimulus representing their perception of 5 seconds passing. This representation may be done by having them hold a button for that specified amount of time or by indicating when the time interval starts and stops with button pushes.

In Clausen’s (1950) study discussed in the previous section, methods of production were also used in a task for the same intervals as before, where the participant was instructed to press a button for an amount of time verbally stated by the experimenter. In a more recent study that utilized the method of time production, Mioni, Stablum, Prunetti, and Grondin (2016) administered a time production task to participants with depression and anxiety, as well as control participants. The time production task that this study implemented instructed participants to produce durations of .5, 1, and 1.5 seconds as prompted by a prompt in the middle of a computer screen for the associated amount of time. During this production period as participants pressed and held the space bar, a grey circle appeared on a white background at the center of the screen and remained there until the participant released the space bar. Additionally, another study that employed methods of reproduction using the same procedure as above with patients with traumatic brain injury, except that the stimulus presented when participants were producing time was a smiley face rather than a grey circle (Mioni, Mattalia, & Stablum, 2013).

Reproduction. Methods of reproduction are, in a way, a combination of verbal estimation and production methods. For these tasks, participants are presented a stimulus for a set amount of time (like verbal estimation) and then asked to generate another stimulus for the same amount of time (like production). This task introduces additional complexity in that participants must employ additional working memory resources to not only perceive the initial stimulus but also remember how long that time was while concurrently producing it.

In Hornstein and Rotter's (1969) earlier study, methods of reproduction were also used in a task for the same intervals as before, where the experimenter displayed the red light for the appropriate duration and the participant was instructed to allow the light to remain on for what the same amount of time. Bauermeister et al. (2005) also included a reproduction task using the same procedure and intervals as their verbal estimation task. The slight variation here was that instead of the child responding verbally, they had their own flashlight to replicate the experimenter's demonstrated duration. Along with methods of production in used by Mioni et al. (2016) to examine populations with anxiety and depression, they also used methods of reproduction with this population utilizing a similar procedure. The variation was that instead of the visual prompt in units of time (e.g., "produce 1 second"), the stimulus was a grey circle presented in the center of the screen acting as the target interval. After this presentation and a delay of 1 second, a question mark prompted participants to press and hold the space bar for an equal duration.

Discrimination. Discrimination methods broadly require participants to offer a dichotomous response following exposure to stimuli. Again, these methods are considered methods of comparison, as participants are often presented with two sequential stimuli: the first being a standard or target stimulus and the latter being the comparison stimulus that participants will judge in relation to the target stimulus. There are variations of comparison tasks based on variables such as number of presented stimuli, the order they are presented, if the intervals are fixed, and what is being asked of the participant.

The two most common and relevant differences among comparison tasks involve variation of intervals and number of stimuli. Macmillan and Creelman (1991) discerned that if the target stimulus is first presented each trial, it is presented as fixed or called the *reminder* variation. Alternatively, if the target stimulus varies across trials it is considered the *roving* variation. This distinction is relevant regardless of the length of the interval. Additionally, participants may be asked to compare the latter stimulus to the first or be asked which of the two stimuli was longer (or shorter). When the stimuli are presented in this manner, it is considered a *forced choice* (FC) variation. In other cases, roving variations refer to the placement of the standard stimulus (i.e., first or second). Further, participants are usually tasked with the objective of determining one of two things. This objective is to note whether the stimuli's intervals are the same duration or not (the *equality task)*, or if the comparison stimulus is longer or shorter than the target stimulus (the *comparative task)*. There is also a *ternary task* that combines these tasks and requires participants to determine if the comparison stimulus is longer, shorter, or not different than the target stimulus (García-Pérez, 2014; Grondin, 2010; Mioni et al., 2013; Treisman, 1963).

In Mioni et al.'s (2013) study focusing on a population of individuals with traumatic brain injury, methods of discrimination were also used with a roving duration, FC comparison task of similar intervals as reported previously. The stimuli to be compared were black and white smiley faces, presented in the order with the standard stimulus first and the comparison second, where participants were then tasked to determine of the second stimulus was longer or shorter than the first. In another study that utilized the method of discrimination, Rubia, Noorloos, Smith, Gunning, and Sergeant (2003) focused on children with ADHD and administered a task where pairs of airplanes were presented in fixed duration of 5 seconds for the standard stimulus and either 3 or 5 seconds as the comparison.

Criticism of Prior Research

To ensure any task is practical and useful, in terms of trustworthiness that it is assessing the correct construct and doing so in a consistent manner, it must be evaluated to determine evidence of its validity and reliability. Notably, other than a few of the earliest studies designed to evaluate the reliability of time perception tasks, there has been little emphasis on their psychometric evaluation (Wittman, 1999). When the measurement properties of time perception tasks have been considered, a variety of idiosyncratic terms have been employed by researchers. Following this trend, there has been a lack of uniformity across and within these methods, as pointed out by many reviews (Allman & Meck, 2012; Grondin, 2010; Toplak et al., 2006). Although there have been studies that investigate various methodology of one task (Mioni, Stablum, McClintock, & Grondin, 2014) as well as studies that have investigated the accuracy, consistency, and stability of multiple measures (Asaoka & Wantanabe, 2015), there has not been one fully encompassing study that examines the psychometric properties of the most common methods of estimating time perception.

Purpose of the Study

The purpose of this study is to examine the various methods of time perception and determine the most reliable and valid tasks designed to assess it. Another goal of this study is to consolidate and define the terminology of measurement of time perception in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and when this goal cannot be achieved, to assert that another term should be used. One term commonly implemented across the time perception literature is *accuracy*. In the realm of research on time perception, this term is pretty straightforward. The accuracy of a task would demonstrate evidence of criterion-related validity from the perspective of *Standards*. For the case of the TIPT, this property would entail concurrent validity evidence comparing the participants answer of subjective time to the incontrovertible index of objective time itself using, sometimes considered the estimated-to-target-duration ratio (RATIO; Glicksohn & Hadad, 2012; Hornstein & Rotter, 1969; Mioni et al., 2014, Nichelli, 1969). Thus, the accuracy of a measure would be how closely to the objective time it allows a participant to produce or estimate a subjective time. Other terms that are widely used across the time perception literature are *consistency* and *precision* (Allan, 1979; Nichelli, 1996). As related to TIPTs, this property would represent test-retest reliability from the perspective of *Standards*; this property would be demonstrated by a portion of participants who would complete some of the TIPT multiple times through design. Another way the literature refers to consistency is through coefficient of variation (CV), which is the standard deviation of a particular response divided by the mean of said response. This calculation is also known as the *Weber ratio*, as it pertains to Weber's law through SET as well as many other models (Mangels & Irvy, 2001). Another term used by Asoaka & Watanabe (2015) is *stability*, which refers to the standard error of accuracy across methods. To assess the stability of a method, the authors take CV across the methods by looking at the same intervals using different methods (compared to precision when looking at same methods with different intervals or individuals who respond to the same interval multiple times using the same method). It is possible that this property best reflects internal consistency reliability from the perspective of *Standards.*

Two final properties, *difficulty* and *discriminability*, are considered in this study to represent the characteristics and scaling of items included on each time perception task. For this study, the discriminability of the item indicates how well a task differentiates between participants who have higher and lower abilities of perceiving time. Difficulty, which is the ease of the task (i.e., what degree a response is considered "wrong"), is computed by calculating z-scores stemming from the standard deviation of the item—just as stability was described earlier. For this study, utilizing the incontrovertible index of objective time, responses exceeding a range of .10 above or below the target objective time as incorrect—a method that produces more incorrect answers would be considered more difficult. Additionally, a method that is highly discriminable would produce accurate results from those who are, overall, highly accurate at perceiving time. Likewise, those who experience may experience deficits in perceiving time would do so more clearly when completing methods that have evidence of high discriminability. Conversely, methods with the lowest discriminability would not be able to determine differences of time perception ability.

This study sheds light on the best methods for assessing time estimation and provides a clear example of consistent measures, as well as terminology, to be utilized in future research. Evidence of reliability and validity demonstrated by these measures not only opens the door for finding a more direct representation of time perception but also allows for future research to investigate various components of time perception in specific populations (e.g., individuals with ADHD). Understanding the underlying components of time perception also provides more insight on (a) the mechanisms involved and (b) what other aspects of daily living are affected by variability in these mechanisms seen between individuals.

Method

Participants

This investigation recruited 136 participants from the University of Memphis' SONA-System—a subject pool consisting of primarily undergraduate psychology students. This number of participants exceeds the number that the previous literature recommends for analyses using item response theory (Foley 2010; He & Wheadon, 2013; Jiang, Wang, & Weiss, 2016; Thorpe & Favia, 2012). The majority of the participants identified as female (65.4%). Additionally, most participants described themselves as either White or Caucasian (48.6%) or Black or African American (42.0%), with a much smaller number as Asian (3.6%), Other (3.6%), Native Hawaiian or Other Pacific Islander (0.7%), or preferring not to answer (1.5%). The age range of participants was from 18 to 53 years (*M* = 21.24, *SD* = 4.89). The only exclusionary criterion was if the participant lacked adequate vision to register the stimuli presented in the tasks or motor dexterity to respond to the tasks (as described in sections that follow).

Measures

Time perception tasks. Participants completed the four tasks below on a computer screen with the clock covered up on the desktop and all wall mounted clocks removed from the room. The tasks were run on Superlab 5 (Heller et al., 2015). All tasks included instructions and one sample item prior to initiating the task.

Time estimation. For the verbal estimation task (Clausen, 1950; Hornstein & Rotter, 1969), participants were exposed to a neutral visual stimulus (a blue circle) and asked to report the length of its exposure. Stimuli were exposed, in random order, for 1, 2, 3, 5, 7, 11, 18, 28, 30, 43, 45, 49, and 60 seconds. Following each exposure, participants were asked to estimate how long the stimulus was on the screen using the number keys on the keyboard. Participants were prompted to give their response in seconds (e.g., entering the number 3 for 3 seconds). Accuracy of response is calculated by the RATIO (i.e., a participant's response divided by the actual duration). For example, if a participant response is "3" for a duration that was actually 5 seconds, the RATIO would be .6, indicating a stark underestimation of time. Further, the WR was calculated by taking the standard deviation divided by participants' mean response, within task (for precision) or across task (for stability). For the CV, the larger the fraction is, the less consistent the participant's response is.

Time production. Based on the time production task used by Mioni et al. (2016), participants were given a chronometric prompt on screen and asked to generate a stimulus (a green circle) for that amount of time. The visual prompt is text that instructs participants to generate a stimulus, in random order, for 1, 2, 3, 5, 7, 11, 18, 28, 30, 43, 45, 49, and 60 seconds. To accomplish this, participants would push and hold the space bar on a keyboard for the prompted amount of time. Again, as they were holding the space bar, the neutral stimulus appeared on the computer screen. Their response is calculated in the same manner as time estimation, except using their space push duration as their response. Thus, if they hold down the space bar for 10 seconds when the prompt was for 11 seconds, the RATIO would be .91.

Time reproduction. For the time reproduction task (Bauermeister et al., 2005; Mioni et. al., 2016), participants were presented with a neutral visual stimulus (a blue circle) and asked to recreate the perceived amount of time that was presented. Stimuli were exposed, in random order, for 1, 2, 3, 5, 7, 11, 18, 28, 30, 43, 45, 49, and 60 seconds. The participant was then instructed to push and hold down the space bar on the keyboard and generate a stimulus (a green circle), with an attempt to recreate the amount of time that the previous stimulus was presented. Same as the previous two tasks, the RATIO and CV was calculated based on the participant's perceived duration of the target stimulus.

Time discrimination. For the time discrimination tasks, there were two types of neutral visual stimuli (blue and green circles) that participants were asked to compare the durations of using the variation indicated as forced choice (FC). The blue circle is the standard stimulus for which participants based their answers on and the green circle was the comparison stimulus. The stimuli were presented together for random intervals of time between 1-4 seconds in quarter-second intervals. Further, there were three versions of this task that followed either the fixed method of stimulus presentation, given that the standard always comes first, and the comparison follows.

In the comparative variation of this this FC task, the stimuli were presented following method of constant stimuli (MCS; Treisman, 1963), where the stimuli were given and the participant was instructed to indicate if the comparison stimulus was longer or shorter than the standard stimulus. Participants were prompted to provide their response by pressing corresponding keys on the keyboard: the “L” key indicating that the comparison stimulus was longer or the “S” key indicating the comparison stimulus was shorter.

In the equality variation of this FC task, participants were given the two stimuli and instructed to indicate whether or not the two stimuli were the same duration. Participants were prompted to provide their response by pressing corresponding keys on the keyboard: the “N” key indicating that the stimuli were not different intervals or the “D” key indicating the stimuli were different durations.

In the ternary variation of the FC task, participants were given the two stimuli and instructed to indicate whether the comparison stimulus was not different from, longer, or shorter than the standard stimulus. It is argued that this variation of the task may reduce some response bias, as the “no difference” response acts as a choice not to respond. That is to say, if the participant does not discern if the comparison is longer or shorter, they do not necessarily have to guess (García-Pérez, 2014). Participants were prompted to provide their response by pressing corresponding keys on the keyboard: the “L” key indicating that the comparison stimulus was longer, the “S” key indicating the comparison stimulus was shorter, or the “N” key indicating that there was no difference. For all discrimination tasks, items were scored dichotomously based on a correct response, yielding a percentage correct that represents a participant’s accuracy at different intervals, as well as the task overall[1].

Post-task questionnaire. Participants completed a questionnaire on Qualtrics following the presentation of the four tasks.

[1] In other studies, assessing discrimination tasks the *point of subjective equality* (PSE) is often calculated. The PSE, or *just noticeable difference*, is when a participant classifies a duration as “longer” for half of the intervals, which was not able to be computed in the current study (Allman & Meck, 2012; Lapid, Ulrich, & Rammsayer, 2008; Mioni et al., 2016; Nichelli, 1996).

Self-report measures. There were four questions regarding the completed tasks. As follows they asked about participant's overall estimation of the entire study, their perceived performance on the tasks, what strategies they might have employed during the tasks, and a manipulation check to assure they removed timing devices (see Appendix A). The current study only examined the responses regarding perception strategies and the manipulation check.

Demographics survey. Participants were asked to answer various questions regarding demographic information such as age, gender, race and ethnic background, and presence of diagnosis of a psychological disorder (see Appendix B).

Procedure

The study was available on the SONA-System website. In addition, there were recruitment flyers posted around the Psychology Building and brief recruitment announcements given in various undergraduate psychology classes.

Participants were instructed to come to a room equipped with 20 computers in the Psychology Building. This room contained no clocks. They were presented with an informed consent and signed if they understand and agree (see Appendix A). Following this interaction with the primary researcher or trained undergraduate research assistant, participants were instructed to remove their watches and any other time-telling devices such as cellular phones and place the items in a locker at the front of the room. They were guided to the computer where they completed the four tasks; prior to each of the tasks, the participants were presented with instructions and a sample item for each task to ensure understanding and to assure accuracy. The four tasks were counterbalanced using a balanced Latin square to reduce ordering as well as carryover effects. The total duration of the study lasted between 35 and 60 minutes. Participants were awarded 1 hour of credit for their time with an additional half credit if the time ran over.

Additionally, approximately half of the participants completed half of the tasks two times. This portion was to satisfy the test-retest condition as well as to allow the data from the initial completion of the tasks to be analyzed.

Following the completion of the tasks, participants completed a brief survey on Qualtrics asking a few various qualitative questions asking about personal metacognitive methods of estimating time, their perceived performance, grade point average, and demographic information. Questions about any formal psychological diagnosis were included in the survey, but participants were not required to answer them (see Appendix B). Following participation in the study, participants were debriefed regarding the nature of the study and given an opportunity to ask questions.

Results

Pilot Data

In a small pilot study ($N = 15$), a portion of the proposed tasks were administered to determine if the current methodology would be feasible and fine tune any tasks based on objective outcomes and informal feedback (McNicholas & Floyd, 2017). In this pilot study, verbal estimation, production, reproduction, and the ternary comparison tasks were completed. The data were analyzed using SPSS to achieve the means, standard deviations, absolute difference, and RATIO (i.e., accuracy). For the current analysis, there will be a few additional outcomes calculated as well as incorporation of item response theory (IRT) analysis.

Data Screening

Responses on the manipulation check item were reviewed and all participants reported removing timing devices. Following recommendations of Tabachnick and Fidell (2013), data were screened to determine if values of continuous and discrete variables were within range, if means and standard deviations were realistic, and if any outliers existed. Although all variables were technically in range, there were 16 outliers in which the response was removed due to an error (e.g., the user did not understand the task and incorrectly produced stimuli—the proctor noticed and explained the task again, so the remaining responses were valid). These outliers were coded as missing data as to not be included in analyses and skew the means and standard deviations. Further, production and reproduction responses that were considered a system error were removed on the lower end of responses (i.e., if a response was less than .10 of the target interval, it was due to an erroneous keystroke rather than pushing and holding the space bar), which resulted in the removal of 122 responses and no variable missing more than 10.3% of the cases. Thus, the valid individual responses for all estimation, production, and reproduction tasks totaled 4938 (97.28% of all responses for those tasks).

Analysis of Outcomes by Task Trial

For the estimation, production, and reproduction tasks, participants' raw score responses were recorded in seconds. For the comparison tasks, participants' responses are recorded as their discrimination in response to the comparison stimulus: longer, shorter, no difference, or difference (for the equality task). For the tasks, six different methods were employed to yield total scores for each. In addition, corrected item-total correlations (i.e., correlation of all items if item is removed) and coefficient alpha values were calculated based on correctness scores.

For estimation, production, and reproduction tasks, the *absolute difference* (AD), the *absolute error* (AE), the RATIO, the CV, and correctness were calculated based on the scores from each item. The AD is computed by taking the absolute value of the difference between the subjective an objective time, whereas the AE is computed by taking the absolute value of the difference between the subjective time and objective time, divided by the objective time. The

RATIO is the subjective time divided by the objective time, and the CV is determined by taking the standard deviation and dividing it by the average subjective response (Gibbon, 1977; Mioni et al., 2014). As mentioned, the CV is also known for representing Weber's law through the WR. For the discrimination tasks, the correctness (i.e., if the comparison stimulus was identified accurately in comparison to the target stimulus) was calculated for each of the three overall versions of the task, as well as the interval difference between the two stimuli within each task.

Classical test theory. Item difficulty was assessed for the estimation, production, and reproduction tasks through a dichotomously scored item based on the RATIO; if the participant's response was close to 1 (within ten percent), it was scored as correct. These percentage correct (i.e., correctness) values indicate how easy or hard a given interval might be on a specific task. Further, discriminability was assessed using item-total correlations (corrected through omission of the item in question from the total score) for each interval across tasks based on correctness. Coefficient alpha values were also calculated based on the correctness values for each interval.

Time estimation. For time estimation (see Table 1), as the length of the intervals uniformly increased (see Interval column), the mean AD values increased; alternatively, the mean AE values also tended to decrease with longer intervals, indicating better performance, on average, as the errors were closer to the target duration. The mean RATIO values across intervals demonstrated a similar pattern to the mean AE, with longer interval items having a mean RATIO value closer to 1. Notably, all of the mean RATIO values fell above 1, indicating that participants typically overestimated the intervals across the task. Finally, the mean CV values for time estimation across intervals were smaller as the intervals increased, indicating that responses were less variable, on average, for the longer durations. At a ±.10 cutoff for the RATIO indicating correctness, interestingly enough, smaller intervals were more often estimated correctly, with an average of 33% correct across all intervals for the estimation task. Based on the correctness of intervals, participants more often answered correctly on the shorter intervals (1s to 3s), indicating these intervals were easier to estimate than some of the lengthier intervals. The corrected item-total correlations for the estimation task ranged from .12 to .65, with many of the intermediate intervals yielding the highest values. These are also the intervals that the lowest number of participants estimated correctly, indicating that these arguably more difficult intervals were better at differentiating those who were better at estimating time. The coefficient alpha value was .82 for the estimation task including all intervals.

Time production. For the time production task (see Table 2), similar to the estimation task, the mean AD values increased, and mean AE values decreased as the length of the intervals increased. Notably, the mean AE across all intervals were consistently lower (with smaller standard deviations) than the estimation task. The mean RATIO values all remained less than 1.0, which for production tasks, indicates an overestimation of time—the inverse of the reproduction and estimation tasks—because for methods that involve production of time, this value would actually be considered an overestimation of time as the experimenter defines the interval, whereas for production tasks the participant generates the interval themselves (Brown, 1997; Nichelli, 1969). Additionally, compared to the estimation task, the mean RATIO values tended to be further away from 1 across all intervals. Contrasting the estimation task, the mean CV values for the production task were generally smaller, indicating less variability across intervals, and with the variability decreasing as the length of the interval increased. Finally, using the same criteria as the estimation tasks, participants completed fewer items correctly across the intervals, with an average of 23% correct. For the production task, contrary to the estimation task, fewer participants correctly perceived time at the shorter interval, specifically the 1 second interval only 6% of participants able to estimate this interval correctly, proving to be the highest difficult among all intervals for every task. The corrected item-total correlations for the production task ranged from .05 to .60, noting that generally intervals across this task have a higher discriminability compared to the estimation and reproduction tasks. These intervals values also yielded the highest coefficient alpha (.86) for this task.

Time reproduction. For the time reproduction task (see Table 3), although the mean AD values and mean AE values followed similar patterns as the previous two tasks, both sets of values were lower than the previously mentioned tasks. This finding means a smaller absolute difference across all intervals (even the longest ones) as well as better performance with the errors closest to the target interval. Additionally, the mean RATIO values for this task all fell below 1, indicating an underestimation of time for all intervals. Notably, compared to the other two tasks, the mean RATIO values were closer to 1, with smaller standard deviations as well. Although the mean CV values for time reproduction were similar to production (decreasing as intervals increased, with an average mean CV of 0.38), the reproduction task had a greater range (.21 to .79) than the production task (.32 to .53). As expected from the RATIO values for this task, the reproduction task also yielded the highest percent of correct responses, again using the same cutoff as the previous two tasks. The average amount of correct responses was 45% across all intervals, with shorter intervals (1s to 3 s) more commonly missed by participants indicating greater difficulty with these shorter intervals. Examining the corrected item-total correlations for the reproduction task, these values were generally lower across majority of the intervals with a range of .14 to .53. Coefficient alpha was .72, which is lower than comparable values for the estimation and production tasks.

When comparing all three tasks (see Table 4), based on the averages of all intervals, participants tended to overestimate time with the estimation and production tasks while underestimating on the reproduction task. Further, there was the least amount of error and highest accuracy on the reproduction tasks yet similar variability across tasks (averaged across intervals). Additionally, the reproduction task yielded the highest percentage of correct responses across intervals, on average. Interestingly enough, the internal consistency reliability of the reproduction task was the lowest, whereas the reliability was highest for the production task—which was the task with the lowest average for overall percentage correct.

Time discrimination. For the time discrimination tasks, considering they were forced choice (FC) tasks, there was an objective correct and incorrect answer based on the participant's response. These tasks were designed using three different prompt options described as comparative, equality, and ternary. The overall mean percentage correct for those versions of the tasks were 81%, 66%, and 68%, respectively (see Table 5). Across tasks, the comparative task had a higher overall percentage correct at shared interval differences than the other two tasks. The ternary task had slightly higher percentage correct across all intervals except for the no difference interval, which the equality task had a slightly higher overall percentage correct. Further, across all three tasks, more items were generally answered correctly the longer the difference between intervals. There were a few exceptions at the longest interval and at the no difference interval (as mentioned), which is possible to examine in more detail after looking at the correctness across these interval differences (as described in the paragraph that follows).

When looking at the 250ms difference on the comparative task (see Table 6), the mean percentage of correct responses range from 40% to 84%. The mean percentage of correct responses ranged from 68% to 95% at a 500ms difference, 70% to 90% at a 750ms difference, 73% to 93% at a 1250ms difference, 90% to 91% at a 1500ms difference, 89% to 96% at a 1750ms difference, 87% to 92% at a 2250ms difference, and 88% at the 2500 difference. For the equality version of the discrimination task (see Table 7), average correct responses for intervals that had no difference ranged from 66% to 79%. The mean percentage of correct responses ranged from 32% to 59% at a 500ms difference, 51% to 78% at the 1000ms difference, 71% to 82% at the 1500ms difference, 72% to 86% at the 2000ms range, and 80% to 85% at the 2500ms difference. For the ternary version of the discrimination task (see Table 8), average percentage of correct responses for when stimuli were the same interval (no difference) ranged from 60% to 77%. The mean percentage of correct responses ranged from 32% to 62% at a 500ms difference, 63% to 85% at the 1000ms difference, 67% to 89% at the 1500ms difference, 80% to 88% at the 2000ms range, and 85% at the 2500ms difference.

Item response theory. A psychometric analysis was conducted using item response models that investigated the relationships between intervals and task outcomes with jMetrik (Meyer, 2014). Given the correctness of the aforementioned items, each individual interval was scored as a dichotomous item and analyzed across estimation, production, and reproduction tasks for all 13 intervals. For time estimation (see Table 1), item difficulty was obtained for each interval (see b column), with higher positive values representing more difficult items and more negative values representing easier items. Observably, these values are comparable to the percentage correct, indicating that intervals where more participants correctly identified the target duration (i.e., 1s to 3s) are those with the lowest negative b value. For time production (see Table 2, b column), this task appeared to have more difficult items across all intervals, as all values of b were positive (range 0.15 to 2.38). For time reproduction (see Table 3), b values for intervals 5 seconds or more were negative (range -1.38 to -0.51), indicating the majority of these intervals on this task were less difficult to correctly perceive and reproduce. Interestingly enough, the shortest three intervals—the least difficult items on the estimation task—were the most difficult on the reproduction task yielding b values of 1.52, 0.66, and 0.89 for 1s, 2s, and 3s intervals, respectively.

Test-Retest Reliability

A portion of participants (*N* = 69) were presented two of the tasks twice, following the counterbalanced order, to gather information regarding test-retest reliability of the tasks. There were no significant differences in means between time one and time two for all tasks (see Table 9). For the estimation task, the test-retest reliability coefficient was .71 for the overall average AE and .85 for the average RATIO. The correlation for average correctness across times was .58. For the production task, the test-retest reliability coefficients for the average AE and RATIO were .85 and .83, respectively. The correlation of average correctness was .62. For the reproduction task, the test-retest reliability coefficients for the AE was .64, and the RATIO was .76. In contrast, the test-retest reliability coefficient for the average correctness on the reproduction task was .64. Finally, for the discrimination tasks, for the test-retest reliability coefficients for average correctness were .39, .54, and .60 for the comparative, equality, and ternary tasks, respectively. Cohen's *d* was calculated based on the sample size to determine the effect sizes for the differences. The majority of the differences were negligible (less than 0.20), and only small effect sizes (between 0.20 and 0.50) for the correctness outcome on the estimation task (-.038), and the AE (-0.25) and RATIO (0.47) on the production task (Cohen, 1977).

Relations between Scores within Tasks and Across Tasks

The total outcome scores for all tasks were correlated (see Table 10). As anticipated, the RATIO and AE for the estimation task were highly correlated (.86). When examining the same correlation between RATIO and AE on the production task, the correlation was negative and only moderate (-.43). Additionally, strong negative correlations between the AE and correctness were observed within all tasks for estimation (-.66), production (-.70), and reproduction (-.74). Although the within-task correlations are quite strong, the outcomes across the tasks provide more insight.

Upon inspecting correlations across tasks, the AE for the estimation task was highly correlated with the AE for both the production task (.73) and the reproduction task (.54). There was a strong negative correlation between RATIO for the estimation task and the RATIO for the production task (-.77), which could be explained by the inverse nature of production scoring (i.e., a RATIO below 1 is an overestimation for production, contrary to estimation and reproduction). Interestingly, however, is the weak positive correlation between RATIOS on the production and reproduction tasks (.27). The correlation between the overall correctness on the estimation task was more highly correlated with the production task (.63) than the reproduction task (.34), with the latter being commensurate with the correctness correlation between the production and reproduction tasks (.33).

When examining the correlations between the overall correctness value for the discrimination tasks, there were significant correlations between all tasks including ternary and comparative (.67), ternary and equality (.65), and comparative and equality (.42). When comparing the correctness values from the estimation, production, and reproduction tasks to the same values from the discrimination tasks, the strongest correlations evidenced from the reproduction task; correlations were .51, .46, and .32 with the ternary, comparative, and equality tasks, respectively.

Relations with Demographic Variables

A one-way ANOVA examining the relations between the total scores from each task and gender revealed only nonsignificant differences. Additionally, the total scores from each task were correlated with participants' age in months and GPA (which were, of note, negatively correlated at -.26). All significant correlations were relatively weak (see Table 10), with the strongest being .29 between age and the overall average RATIO for the time reproduction.

When investigating the different strategies reported by participants, the three most common themes were counting (55.88%), tapping to a tempo or using a song (38.97%), and visualizing a clock (9.56%). The majority of participants reported only one method (78.68%), whereas a smaller amount used multiple methods (15.44%) or no strategy at all (5.88%). One-way ANOVAs were used to compare the total scores from each task for (a) participants who used counting strategy versus participants who used the other methods as well as (b) participants who used one strategy versus multiple strategies. The analyses yielded nonsignificant differences across groups.

For group difference and discriminative relations, participants were selected if they report some diagnosis of a psychological disorder (*N* = 18) and within this sub-sample, eight reported having ADHD (44.44%), half reported having some form of depression (50%), half reported having generalized anxiety disorder or some form of anxiety (50%), and one each reported having obsessive compulsive disorder, borderline personality disorder, a specific phobia, or chose not to answer (0.06%). Additionally, among those reporting having a diagnosis of a psychological disorder, the majority reported multiple disorders (67%). These participants were matched by age, gender, and race to those without, also using a one-way ANOVA to compare the total scores; this analysis yielded nonsignificant differences between the groups.

Discussion

The purpose of the current study was to examine the psychometric properties of the most widely used methods of assessing perception of time. Previous studies have investigated different methods of the same tasks (Mioni, Stablum, McClintock, & Grondin, 2014), as well as different intervals across tasks (Asaoka & Wantanabe, 2015); however there has not been a comprehensive study that integrates multiple methods and intervals to a similar degree. The current study aimed to bridge this gap in the literature, completing this comprehensive examination of four of the major methods, represented by a total of 6 different tasks (or task variations). Some of the major terminology used in the past was represented in this study in an attempt to consolidate the various terms used and define them in a manner that is in accordance with the *Standards for Educational and Psychological Testing*. In summary, the terms used were: *accuracy*, represented by the AE and RATIO values and demonstrating evidence of criterion-related validity; *consistency*, represented by the test-retest condition and demonstrating evidence of test-retest reliability; *stability*, represented by the CV value demonstrating internal consistency reliability; and *difficulty* and *discriminability*, represented by the correctness of a participant’s response.

A consistent pattern across all tasks was that as the length of the interval increased, the error and variability tended to decrease, which is in contrast to Weber’s law in general yet commensurate with other findings in humans (Allan, 1979; Allan 1998; Hancock & Block, 2012; Weber, 1933). One explanation of this finding may be due to the smaller intervals allowing for participants to proportionally be very inaccurate, leading to a larger absolute error. Further, with the larger intervals, there was a larger window of correctness due to the RATIO. Overall, for time estimation and production, participants generally overestimated, whereas time reproduction led to underestimation, yet less error. Interestingly, this task differs in that it does not explicitly use traditional timing of "seconds" to generate a response. For example, estimation prompts a numbered response in seconds and production prompts numbered seconds as the target. It is possible that participants who may have been very incorrect on the production and estimation tasks were very accurate on the reproduction tasks, as they are defining their own units of time (e.g., they may have counted twice as fast, thinking that a 30-second stimulus was 60 seconds, but as long as they reproduced their stimulus for a similarly incorrect 60 seconds (that was actually 30 seconds in length), they would be perfectly correct).

The prior observation parallels the results obtained with both CTT and IRT when considering the difficulty of intervals and tasks. As mentioned, participants are defining their own time on the reproduction task, which may have led to the majority of the items yielding both a lower reliability and lower difficulty, compared to the other tasks. Thus, it was easier for participants to correctly reproduce the given interval, although the response patterns were inconsistent across participants. For all tasks, the results from the CTT and IRT analyses were similar—intervals that had a lower percentage correct were indicated to be more difficult and also highly discriminable.

The test-retest reliability analysis was employed to demonstrate the consistency of these tasks across time points. The lowest correctness across times was demonstrated by the comparative discrimination task (.39), and the highest was demonstrated by the reproduction task (.64). Further, looking at both the AE and RATIO across times, in whole, the production task demonstrated steadily high reliability (.83 for the RATIO and .85 for the AE). These findings suggest that production may have the strongest consistency across time.

The correlations of the total outcome scores within tasks were fairly strong, ranging .54 to .86 and demonstrating convincing stability of the tasks (i.e., internal consistency reliability); however, the correlations of these outcomes across tasks were just as strong, if not stronger. Namely, both production and reproduction tasks evidenced solid negative correlations (-.70 and -.74, respectively) between AE and overall correctness. This result demonstrates that when a participant's absolute error increased, the total of items that yielded a correct score decreased, with especially high reliability on both of these tasks. In contrast, and compared to previous literature (Block, Hancock, & Zakay, 2000; Hancock & Block, 2012), relations with temporal processing tasks and external correlates did not produce consistent findings. Specifically, there were no notable significant differences for age, gender, GPA, or psychological diagnosis. Although the investigation of relations to demographic data was not the primary focus of the study, the lack of findings in this area is somewhat surprising, considering the degree that individual differences might otherwise play on various abilities.

Limitations and Future Directions

Some limitations in this project exist, such as utilizing a convenience sample of college students that may not adequately yield a true representation of the general population. Although in examining some characteristics of the sample this sample may adequately represent or exceed the make-up of the general population (e.g., a higher percentage of individuals who identify as Black or African American compared to the US population), there was certainly an underrepresentation of males compared to females (2/3 of the sample consisting of the latter). Future studies may attempt to acquire a more representative sample, perhaps reaching out in the community to obtain non-college-student participation. Considering said college sample, the age range was relatively narrow, with the sample of the current study representing only adults. Future studies may attempt to incorporate children and adolescents to better understand time perception across the lifespan and demonstrate that these methods are equally as reliable and valid at much younger age ranges. Furthermore, despite attempts to control all confounding variables and ensure validity, there were some instances where participants' response pattern characterized a lack of effort of misunderstanding of the task. Obtaining a larger, more general sample of the population (with added incentives to participate), as well as requiring less of each participant (in terms of time and effort), may reduce fatigue and increase motivation to complete the tasks to the best of their ability. More frequent manipulation checks throughout the study (rather than only at the end) may also assist in screening for less apparent noncompliant participants.

In addition, although the study was designed to examine all tasks similarly, there is an evident disconnect between the discrimination and all other tasks, with the former only yielding overall correctness. It may be possible to compare these tasks equally to the others, but it would require a different study design (namely, many more trials of the discrimination items).

A final limitation of this study is the lack of design focusing on differences in individuals with diagnosed psychopathologies. Through random sampling, only a small subset of participants identified as being diagnosed with a psychological disorder (with the option to withhold this information). What would be more beneficial for future studies, coinciding with sampling a more diverse pool of participants (i.e., community involvement and extended age ranges), it could be possible to recruit those with documented diagnoses. One example might be to sample from an elementary or middle school and engage students with individual education plans, to better investigate the impact that these differences may cause.

Implications

For many of these tasks to be both practical and useful, assuring that the correct construct is being assessed in a consistent manner, evidence of validity and reliability must be demonstrated. At present, the tasks were able to demonstrate both validity and reliability through the analyses conducted and meet these psychometric standards. By meeting these standards, especially validity, the evidence supports that these tasks are tapping into the true construct of time itself and how well individuals are able to perceive it. One of the major goals of this study was to identify the best method for assessing perception of time. To accomplish this goal, first it is necessary to focus on the outcomes that represent the highest psychometric standards for the tasks. The outcomes that would best represent these standards would be the RATIO (indicative of accuracy and criterion-related validity). Across the majority of intervals, the reproduction task yielded average RATIO values within 0.10 of one, with a value of one representing perfect accuracy. Another outcome that represents accuracy is absolute error, with the reproduction task yielding AE values that were under 0.20 for all but three of the intervals, whereas the other two tasks only had a few intervals under this value. In addition, a lower coefficient of variance value indicates higher stability for a task, where the reproduction task also yielded the lowest average values for the majority of the intervals. Thus, based on the overall outcomes, the reproduction task yielded values that were preferable compared to the estimation and production task. Conversely, the reproduction task yielded the lowest coefficient alpha and test-retest reliability. Despite time estimation yielding a high coefficient alpha, this task evidenced the highest error and variability, as well as the lowest accuracy (paired with the largest standard deviation for all). Alternatively, the production task yielded the highest coefficient alpha and test-retest reliability. There may not be a clear winner between the production and reproduction task, both demonstrating solid psychometric evidence, it is certainly recommended that one of these two tasks be used in future research.

Another interesting implication from this study stems from the inter-task correlations, which reveal that there may be important individual differences that play a role in perception of time, such as working memory (Block, 1992; Block & Gruber, 2014; Block & Zakay, 2006; Brown, 1997). These differences may better be assessed in a further study, perhaps one highlighting the correlation between these individual differences and perception of time through other measures. Not only would this research be informative to future research, it would also assist in solidifying many of the theoretical underpinnings of temporal processing. These future studies could be designed in a manner that focuses on said theories, further bridging the gap between conceptualizing and demonstrating perception of time.

Across the estimation, production, and reproduction tasks, intervals were included up to 60 seconds to see if there were any prominent differences as stimuli length increased. It is worth noting that as intervals increased, there was less variability among all tasks. When considering future studies and considering what intervals ought to be examined, as found in the current study, intervals exceeding 30 seconds did not yield any noteworthy differences, indicating it may be more beneficial for researchers to focus on the shorter intervals (perhaps those under 25 seconds). The use of shorter intervals would allow for more repetitions and less lengthy run-time.

Following the recommendations for future directions within this line of research, one overarching goal of the scientific community should be to definitively determine the ideal method for assessing time perception. Psychometric soundness was demonstrated through this study, but there are many other considerations to consider take into account (i.e., external correlates such as age). It is evident that interest in perception of time has had a strong hold on the scientific community for many years, with this interest linking to the noticeable relevance in everyday life. Prior to integrating any of these measures into the practical realm of assessment, plenty of additional research must be conducted to develop these measures in a way that is in accordance to the high standards of psychological testing. It is encouraged that researchers continue to study and refine these measures so that one day practitioners may be able to implement the assessment of time perception in both diagnosis and recommendations—a lofty goal to be fulfilled in due time.

References

Allan, L. G. (1979). The perception of time. *Perception & Psychophysics, 26*, 340-354.

Allan, L. G. (1998). The influence of the scalar timing model on human timing research. *Behavioural Processes, 44*, 101-117.

Allman, M. J., & Falter, C. M. (2015). Timing processes in autism spectrum disorder. In A. Vatakis & M. J. Allman (Eds.), *Time distortions in mind–temporal processing in clinical populations* (pp. 37–56). Leiden, The Netherlands: Brill.

Allman, M. J., & Meck, W. H. (2012). Pathophysiological distortions in time perception and timed performance. *Brain, 135*, 656-677.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Publishing.

Asaoka, R., & Wantanabe, Y. (2015). Differences between measuring methods of time estimation. *Tohoku Psychologica Folia, 74*, 1-12.

Aschoff, J. (1985). On the perception of time during prolonged temporal isolation. *Human Neurobiology, 4*, 41-52.

Avni-Babad, D., & Ritov, I. (2003). Routine and the perception of time. *Journal of Experimental Psychology: General, 132,* 543.

Bar-Haim, Y., Kerem, A., Lamy, D., & Zakay, D. (2010). When time slows down: The influence of threat on time perception in anxiety. *Cognition and Emotion, 24*, 255-263.

Bauermeister, J. J., Barkley, R. A., Martínez, J. V., Cumba, E., Ramírez, R. R., Reina, G., ... & Salas, C. C. (2005). Time estimation and performance on reproduction tasks in subtypes of children with attention deficit hyperactivity disorder. *Journal of Clinical Child and Adolescent Psychology, 34*, 151-162.

Bindra, D., & Waksberg, H. (1956). Methods and terminology in studies of time estimation. *Psychological Bulletin, 53*, 155.

Block, R. A., (1979). Time and Consciousness. In G. Underwood & R. Stevens (Eds.), *Aspects of consciousness: vol. 1 psychological issues* (pp. 179-217). London: Academic Press.

Block, R. A. (1992). Prospective and retrospective duration judgment: The role of information process and memory. In F. Macar, V. Pouthas, & W. J. Friedman (Eds.), *Time, action and cognition: towards bridging the gap* (pp. 141–152). Dordrecht, the Netherlands: Kluwer.

Block, R. A., & Gruber, R. P. (2014). Time perception, attention, and memory: A selective review. *Acta Psychologica*, *149*, 129-133.

Block, R. A., & Zakay, D. (1996). Models of psychological time revisited. *Time and Mind, 33*, 171-195.

Block, R. A., & Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review*, *4*, 184-197.

Block, R. A., & Zakay, D. (2006). Prospective remembering involves time estimation and memory processes. In J. Glicksohn & M. S. Myslobodsky (Eds.), *Timing the future: the case for a time-based prospective memory* (pp. 25-49). London, England: World Scientific.

Block, R. A., George, E. J., & Reed, M. A. (1980). A watched pot sometimes boils: A study of duration experience. *Acta Psychologica, 46*, 81-94.

Block, R. A., Hancock, P. A., & Zakay, D. (2000). Sex differences in duration judgements: A meta-analytic review. *Memory & Cognition, 28*, 1333-1346.

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica, 134*, 330-343.

Brown, S. W. (1997). Attentional resources in timing: Interference effects in concurrent temporal and nontemporal working memory tasks. *Perception & Psychophysics, 59*, 1118-1140.

Brown, S. W., & Boltz, M. G. (2002). Attentional processes in time perception: Effects of mental workload and event structure. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 600.

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience, 6*, 755-765.

Carroll, C. A., Boggs, J., O'Donnell, B. F., Shekhar, A., & Hetrick, W. P. (2008). Temporal processing dysfunction in schizophrenia. *Brain and Cognition, 67*, 150-161.

Clausen, J. (1950). An evaluation of experimental methods of time judgment. *Journal of Experimental Psychology, 40*, 756.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

Davalos, D. B., & Opper, J. (2015). Time processing in schizophrenia. In A. Vatakis & M. J. Allman (Eds.), *Time distortions in mind–temporal processing in clinical populations* (pp. 93–114). Leiden, The Netherlands: Brill.

DeWall, C. N., Visser, P. S., & Levitan, L. C. (2006). Openness to attitude change as a function of temporal perspective. *Personality and Social Psychology Bulletin, 32*, 1010-1023.

Droit-Volet, S., Fayolle, S., Lamotte, M., & Gil, S. (2013). Time, emotion and the embodiment of timing. *Timing & Time Perception, 1*, 99-126.

Eisler, H. (1976). Experiments on subjective duration 1868-1975: A collection of power function exponents. *Psychological Bulletin, 83*, 1154.

Fayolle, S., Gil, S., & Droit-Volet, S. (2015). Fear and time: Fear speeds up the internal clock. *Behavioural Processes, 120*, 135-140.

Foley, B. P. (2010). *Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique.* (Doctoral Dissertation). University of Nebraska—Lincoln: Open Access Theses and Dissertations from the College of Education and Human Sciences. Paper 75. Retrieved from http://digitalcommons.unl.edu/cehsdiss/75.

García-Pérez, M. A. (2014). Does time ever fly or slow down? The difficult interpretation of psychophysical data on time perception. *Frontiers in Human Neuroscience, 8*, 1-19.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review, 84*, 279.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences, 423*, 52-77.

Glicksohn, J., & Hadad, Y. (2012). Sex differences in time production revisited. *Journal of Individual Differences, 33*, 35-42.

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics, 72*, 561-582.

Gu, B. M., Jurkowski, A. J., Lake, J. I., Malapani, C., & Meck, W. H. (2015). Bayesian models of interval timing and distortions in temporal memory as a function of Parkinson's disease and dopamine-related error processing. In A. Vatakis & M. J. Allman (Eds.), *Time distortions in mind* (pp. 281-327). Boston, MA: Brill.

Hancock, P. A., & Block, R. A. (2012). The psychology of time: a view backward and forward. *The American Journal of Psychology, 125*, 267-274.

He, Q., & Wheadon, C. (2013). The effect of sample size on item parameter estimation for the partial credit model. *International Journal of Quantitative Research in Education, 1*, 297-315.

Heller, K., Matsak, E., Abboud, H., Schultz, H., & Zeitlin, V. (2015). SuperLab 5 (Version 5.0.5) [Computer software]. San Pedro, CA: Cedrus.

Hicks, R. E., Miller, G. W., & Kinsbourne, M. (1976). Prospective and retrospective judgments of time as a function of amount of information processed. *The American Journal of Psychology, 89*, 719-730.

Hornstein, A. D., & Rotter, G. S. (1969). Research methodology in temporal perception. *Journal of Experimental Psychology, 79*, 561.

James, W. (1890). *The principles of psychology*. New York, NY: Henry Holt.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology, 7*, 109.

Lapid, E., Ulrich, R., & Rammsayer, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Attention, Perception, & Psychophysics, 70*, 291-305.

Mangels, J. A., & Ivry, R. B. (2001). Time perception. In B. Repp (Ed.), *The handbook of cognitive neuropsychology: What deficits reveal about the human mind* (pp. 467-493). Philadelphia, PA: Psychology Press.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York, NY: Cambridge University Press

McNicholas, P. J., & Floyd, R. G. (August 2017). *Measuring time perception: A psychometric analysis*. Poster session presented at the annual convention of American Psychological Association, Washington D.C.

Meaux, J. B., & Chelonis, J. J. (2003). Time perception differences in children with and without ADHD. *Journal of Pediatric Health Care, 17*, 64-71.

Meyer, J. P. (Ed.) (2014). *Applied measurement with jMetrik.* New York, NY: Routledge.

Mioni, G., Mattalia, G., & Stablum, F. (2013). Time perception in severe traumatic brain injury patients: a study comparing different methodologies. *Brain and Cognition, 81*, 305-312.

Mioni, G., Stablum, F., McClintock, S. M., & Grondin, S. (2014). Different methods for reproducing time, different results. *Attention, Perception, & Psychophysics, 76*, 675-681.

Mioni, G., Stablum, F., Prunetti, E., & Grondin, S. (2016). Time perception in anxious and depressed patients: A comparison between time reproduction and time production tasks. *Journal of Affective Disorders, 196*, 154-163.

Nichelli, P. (1996). Time perception measurements in neuropsychology. *Advances in Psychology, 115*, 187-204.

Ornstein, R. E. (1969). *On the experience of time.* Baltimore, MD: Penguin Books.

Perret-Clermont, A. N., & Lambolez, S. (2005). Time, Mind, and Otherness. In A.N. Perret-Clermont (Ed.), *Thinking time: A multidisciplinary perspective on time* (pp.1-12). Cambridge, England: Hogrefe & Huber.

Pollak, Y., Kroyzer, N., Yakir, A., & Friedler, M. (2009). Testing possible mechanisms of deficient supra-second time estimation in adults with attention-deficit/hyperactivity disorder. *Neuropsychology, 23*, 679-686.

Rubia, K., Noorloos, J., Smith, A., Gunning, B., & Sergeant, J. (2003). Motor timing deficits in community and clinical boys with hyperactive behavior: The effect of methylphenidate on motor timing. *Journal of Abnormal Child Psychology, 31*, 301-313.

Siegman, A. W. (1962). Intercorrelation of some measures of time estimation. *Perceptual and Motor Skills, 14*, 381-382.

Strathman, A., & Joireman, J. (Eds.). (2005). *Understanding behavior in the context of time: Theory, research, and application.* Mahwah, NJ: Erlbaum.

Szelag, E., Kowalska, J., Galkowski, T., & Pöppel, E. (2004). Temporal processing deficits in high-functioning children with autism. *British Journal of Psychology, 95*, 269-282.

Tibachnick, B. G., & Fidell, L. S. (Eds.). (2013). *Using multivariate statistics* (6th ed.). New York, NY: Harper & Row.

Thönes, S., & Oberfeld, D. (2015). Time perception in depression: A meta-analysis. *Journal of Affective Disorders, 175*, 359-372.

Thomas, E. C., & Brown, I. (1974). Time perception and the filled-duration illusion. *Attention, Perception, & Psychophysics, 16*, 449-458.

Thorpe, G. L., & Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship, 20*, 1-33. Retrieved from http://digitalcommons.library.umaine.edu/psy_facpub/20

Toplak, M. E., & Tannock, R. (2005). Time perception: modality and duration effects in attention-deficit/hyperactivity disorder (ADHD). *Journal of Abnormal Child Psychology, 33*, 639-654.

Toplak, M. E., Dockstader, C., & Tannock, R. (2006). Temporal information processing in ADHD: Findings to date and new methods. *Journal of Neuroscience Methods, 151*, 15-29.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the "internal clock". *Psychological Monographs: General and Applied, 77*, 1-31.

Wallace, M., & Rabin, A. I. (1960). Temporal experience. *Psychological Bulletin, 57*, 213-236.

Wittmann, M. (1999). Time perception and temporal processing levels of the brain. *Chronobiology International, 16*, 17-32.

Wittmann, M., & Paulus, M. P. (2009). Temporal horizons in decision making. *Journal of Neuroscience, Psychology, and Economics, 2*, 1-11.

Wittmann, M. (2013). The inner sense of time: How the brain creates a representation of duration. *Nature Reviews Neuroscience*, *14*, 217-223.

Wearden, J. H., & Penton-Voak, I. S. (1995). Feeling the heat: Body temperature and the rate of subjective time, revisited. *The Quarterly Journal of Experimental Psychology, 48*, 129-141.

Weber, A. O. (1933). Estimation of time. *Psychological Bulletin, 30*, 233-252.

Zakay, D. (1990). The evasive art of subjective time measurement: Some methodological dilemmas. In R. A. Block (Ed.), *Cognitive models of psychological time* (pp. 59–84). Hillsdale, NJ: Erlbaum.

Zakay, D. (1993). Relative and absolute duration judgments under prospective and retrospective paradigms. *Perception & Psychophysics, 54*, 656-664.

Zakay, D., & Block, R. A. (1995). An attentional-gate model of prospective time estimation. In M. Richelle, V.D. Keyser, G. d'Ydewalle, A. Vandierendonck (Eds.), *Time and the dynamic control of behavior* (pp. 167-178). Liège, Belgium: Universite de Liege.

Zakay, D., & Block, R. A. (1996). The role of attention in time estimation processes. *Advances in Psychology, 115*, 143-164.

Zakay, D., & Block, R. A. (1997). Temporal cognition. *Current Directions in Psychological Science, 6*, 12-16.

Appendix A

About how long do you think that the time estimation tasks took in total? Report in minutes.

How do you think you did on the tasks? Report what percentage of items you believe you answered correctly?

0 10 20 30 40 50 60 70 80 90 100

When estimating time during this study, what strategies did you use to improve your performance?

Prior to beginning today's study, did you remove all timing devices (watches, phones, etc.) and put them in the manila envelope provided to you until all tasks were completed?

○ Yes
○ No

Appendix B

What is your age in years and month? (Ex: 20-10)

What is your current grade point average (GPA)?

What gender do you identify with?

○ Male
○ Female
○ Neither
○ Other ____________________
○ Prefer not to answer

How would you describe yourself?

□ American Indian or Alaska Native
□ Asian
□ Black or African American
□ East Indian
□ Native Hawaiian or Other Pacific Islander
□ White or Caucasian
□ Other ____________________
□ Prefer not to answer

Are you of Spanish / Hispanic / Latino(a) origin? (Select the appropriate group(s)).

○ No, not of Hispanic, Latino(a), or Spanish origin
○ Yes, Mexican, Mexican American, Chicano
○ Yes, Puerto Rican
○ Yes, Cuban
○ Yes, other Hispanic, Latino(a), or Spanish origin, please specify: ____________________
○ Prefer not to answer

Have you been previously diagnosed with a psychological disorder?

○ Yes
○ Maybe
○ No
○ Prefer not to answer

If so, which one(s)?

□ Attention-Deficit Hyperactivity Disorder (ADHD)
□ Depression
□ Generalized Anxiety Disorder (GAD)
□ Obsessive Compulsive Disorder (OCD)
□ Schizophrenia
□ Other, please specify: ____________________
□ Other, please specify: ____________________
□ Other, please specify: ____________________
□ Prefer not to answer

Table 1

Total Scores of Time Estimation (TE) Task

Interval	AD	AE	RATIO	CV	Correct	CIT	b
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M*	%		
1s	0.50 (0.75)	0.50 (0.75)	1.48 (0.76)	0.51	61	.12	-1.77
2s	0.72 (1.08)	0.36 (0.54)	1.26 (0.59)	0.47	54	.44	-1.35
3s	1.01 (1.27)	0.34 (0.42)	1.22 (0.50)	0.41	42	.43	-0.72
5s	1.86 (2.35)	0.37 (0.47)	1.20 (0.57)	0.47	26	.56	0.24
7s	2.36 (2.86)	0.34 (0.41)	1.18 (0.50)	0.42	24	.44	0.37
11s	3.52 (3.78)	0.32 (0.34)	1.17 (0.44)	0.37	40	.58	-0.62
18s	5.79 (6.82)	0.32 (0.38)	1.16 (0.47)	0.41	18	.40	0.83
28s	8.21 (7.75)	0.29 (0.28)	1.09 (0.39)	0.36	22	.65	0.52
30s	9.37 (11.43)	0.31 (0.38)	1.15 (0.47)	0.41	30	.54	-0.02
43s	10.85 (9.35)	0.27 (0.27)	1.05 (0.37)	0.36	31	.53	-0.08
45s	11.42 (10.86)	0.25 (0.24)	1.06 (0.35)	0.33	26	.39	0.24
49s	13.26 (13.26)	0.27 (0.25)	1.07 (0.36)	0.34	20	.59	0.67
60s	14.48 (17.42)	0.24 (0.29)	1.07 (0.37)	0.35	36	.45	-0.38
Overall	-	0.32 (0.29)	1.17 (0.38)	0.40	33	-	-
α						.82	

Note. AD = absolute difference; AE = absolute difference; RATIO = estimated to target ratio; CV = coefficient of variance; CIT = corrected item-total correlation; b = item difficulty; α=coefficient alpha.

Table 2
Total Scores of Time Production (TP) Task

Interval	AD	AE	RATIO	CV	Correct	CIT	b
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M*	%		
1s	0.43 (0.24)	0.43 (0.24)	0.65 (0.35)	0.53	6	.05	2.38
2s	0.66 (0.45)	0.33 (0.22)	0.74 (0.30)	0.41	21	.45	0.82
3s	0.94 (0.66)	0.31 (0.22)	0.76 (0.30)	0.40	18	.50	1.05
5s	1.39 (1.07)	0.28 (0.21)	0.83 (0.31)	0.37	28	.49	0.33
7s	1.92 (1.73)	0.27 (0.25)	0.85 (0.34)	0.40	24	.59	0.59
11s	2.71 (2.21)	0.25 (0.20)	0.89 (0.30)	0.34	22	.51	0.73
18s	4.11 (3.68)	0.23 (0.20)	0.88 (0.28)	0.32	31	.60	0.15
28s	6.98 (5.51)	0.25 (0.20)	0.89 (0.30)	0.34	24	.51	0.56
30s	7.35 (6.16)	0.24 (0.21)	0.87 (0.29)	0.34	28	.56	0.31
43s	11.10 (9.60)	0.26 (0.22)	0.89 (0.32)	0.36	27	.58	0.42
45s	11.35 (8.88)	0.25 (0.20)	0.91 (0.31)	0.34	23	.54	0.65
49s	13.14 (11.17)	0.27 (0.23)	0.88 (0.33)	0.38	24	.59	0.58
60s	15.62 (14.33)	0.26 (0.24)	0.88 (0.33)	0.38	28	.55	0.31
Overall	-	0.28 (0.18)	0.84 (0.27)	0.38	23	-	-
α						.86	

Note. AD = absolute difference; AE = absolute difference; RATIO = estimated to target ratio; CV = coefficient of variance; CIT = corrected item-total correlation; b = item difficulty; α=coefficient alpha.

Table 3

Total Scores of Time Reproduction (TR) Task

Interval	AD	AE	RATIO	CV	Correct	CIT	b
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M*	%		
1s	0.46 (0.60)	0.46 (0.60)	0.96 (0.75)	0.79	13	.14	1.52
2s	0.69 (1.27)	0.35 (0.64)	0.93 (0.72)	0.77	24	.22	0.66
3s	0.77 (0.54)	0.26 (0.18)	0.85 (0.28)	0.33	21	.25	0.89
5s	1.12 (2.73)	0.22 (0.55)	0.96 (0.59)	0.61	44	.29	-0.52
7s	1.19 (1.40)	0.17 (0.20)	0.92 (0.25)	0.27	49	.39	-0.71
11s	2.05 (2.65)	0.19 (0.24)	0.87 (0.28)	0.32	50	.47	-0.79
18s	3.15 (5.11)	0.18 (0.28)	0.95 (0.33)	0.35	61	.53	-1.38
28s	4.80 (6.07)	0.17 (0.22)	0.87 (0.25)	0.28	54	.42	-0.98
30s	4.86 (5.80)	0.16 (0.19)	0.93 (0.24)	0.26	62	.30	-1.38
43s	7.10 (9.19)	0.17 (0.21)	0.87 (0.24)	0.27	55	.46	-0.97
45s	6.47 (7.82)	0.14 (0.17)	0.94 (0.22)	0.23	58	.27	-1.15
49s	8.28 (9.46)	0.17 (0.19)	0.89 (0.23)	0.26	51	.33	-0.73
60s	7.81 (9.16)	0.13 (0.15)	0.96 (0.20)	0.21	62	.29	-1.25
Overall	-	0.22 (0.17)	0.91 (0.16)	0.38	45	-	-
α						.72	

Note. AD = absolute difference; AE = absolute difference; RATIO = estimated to target ratio; CV = coefficient of variance; CIT = corrected item-total correlation; b = item difficulty; α=coefficient alpha

Table 4
Total Scores for Tasks Across All Intervals

Total Score	Estimation	Production	Reproduction
AE	0.32 (0.29)	0.28 (0.18)	0.22 (0.17)
RATIO	1.17 (0.38)	0.84 (0.27)	0.91 (0.16)
CV	0.40	0.38	0.38
% Correct	33	23	45
α	.82	.86	.72

Note. AD = absolute difference; AE = absolute difference; RATIO = estimated to target ratio; CV = coefficient of variance; α=coefficient alpha.

Table 5

Average Correctness of the Discrimination Tasks Across Intervals by Difference

Difference (ms)	Comparative	Equality	Ternary
0	–	72	70
250	65	–	–
500	78	46	47
750	83	–	–
1000	–	68	72
1250	86	–	–
1500	91	78	79
1750	93	–	–
2000	–	80	85
2250	90	–	–
2500	88	83	85
Overall	81	66	68

Table 6

Average Correctness for Comparative Discrimination Task by Difference

Target (ms)	Comparison (ms)	Difference (ms)	Correctness (%)
1500	1250	250	63
1500	1750	250	73
2500	2250	250	47
2500	2750	250	83
3500	3750	250	84
3500	3250	250	40
1500	1000	500	83
1500	2000	500	95
2500	2000	500	68
2500	3000	500	86
3500	3000	500	56
1500	2250	750	90
2500	1750	750	80
2500	3250	750	90
3500	2750	750	70
1500	2750	1250	91
2500	1250	1250	86
2500	3750	1250	93
3500	2250	1250	73
1500	3000	1500	91
2500	1000	1500	91
3500	2000	1500	90
1500	3250	1750	96
3500	1750	1750	89
1500	3750	2250	92
3500	1250	2250	87
3500	1000	2500	88

Table 7

Average Correctness for Equality Discrimination Task by Difference

Target (ms)	Comparison (ms)	Difference (ms)	Correctness (%)
1000	1000	0	74
1500	1500	0	79
2000	2000	0	71
2500	2500	0	73
3000	3000	0	66
3500	3500	0	69
1000	1500	500	48
1500	1000	500	45
1500	2000	500	52
2000	1500	500	32
2000	2500	500	50
2500	2000	500	37
2500	3000	500	59
3000	2500	500	41
3000	3500	500	57
3500	3000	500	41
1000	2000	1000	73
1500	2500	1000	75
2000	1000	1000	59
2000	3000	1000	78
2500	1500	1000	64
2500	3500	1000	75
3000	2000	1000	67
3500	2500	1000	51
1000	2500	1500	81
1500	3000	1500	77
2000	3500	1500	80
2500	1000	1500	82
3000	1500	1500	71
3500	2000	1500	76
1000	3000	2000	86
1500	3500	2000	84
3000	1000	2000	72
3500	1500	2000	78
1000	3500	2500	85
3500	1000	2500	80

Table 8
Average Correctness for Ternary Discrimination Task by Difference

Target (ms)	Comparison (ms)	Difference (ms)	Correctness (%)
1000	1000	0	75
1500	1500	0	77
2000	2000	0	67
2500	2500	0	60
3000	3000	0	75
3500	3500	0	65
1000	1500	500	54
1500	1000	500	51
1500	2000	500	47
2000	1500	500	47
2000	2500	500	51
2500	2000	500	42
2500	3000	500	53
3000	2500	500	34
3000	3500	500	62
3500	3000	500	32
1000	2000	1000	82
1500	2500	1000	73
2000	1000	1000	71
2000	3000	1000	72
2500	1500	1000	63
2500	3500	1000	85
3000	2000	1000	68
3500	2500	1000	63
1000	2500	1500	85
1500	3000	1500	77
2000	3500	1500	89
2500	1000	1500	79
3000	1500	1500	67
3500	2000	1500	75
1000	3000	2000	87
1500	3500	2000	84
3000	1000	2000	88
3500	1500	2000	80
1000	3500	2500	85
3500	1000	2500	85

Table 9
Test-Retest Reliability

Task	Outcome	Time	*M* (*SD*)	Correlation	*t*	Cohen's *d*
Estimation	Absolute Error	1	0.27 (0.22)			
		2	0.26 (0.19)	.71	0.66	0.12
	RATIO	1	1.10 (0.32)			
		2	1.11 (0.28)	.85	-0.55	-0.09
	Correctness	1	0.27 (0.22)			
		2	0.36 (0.26)	.58	-2.22	-0.38
Production	Absolute Error	1	0.22 (0.12)			
		2	0.24 (0.14)	.85	-1.41	-0.25
	RATIO	1	0.90 (0.20)			
		2	0.84 (0.21)	.83	2.67	0.47
	Correctness	1	0.27 (0.25)			
		2	0.30 (0.26)	.62	-0.60	-0.11
Reproduction	Absolute Error	1	0.19 (0.12)			
		2	0.17 (0.13)	.64	0.85	0.14
	RATIO	1	0.90 (0.15)			
		2	0.89 (0.13)	.76	0.36	0.06
	Correctness	1	0.50 (0.21)			
		2	0.49 (0.23)	.64	0.21	0.04
Discrimination	Comparative	1	0.79 (0.17)			
		2	0.78 (0.19)	.39	0.17	0.03
	Equality	1	0.60 (0.25)			
		2	0.64 (0.20)	.54	-0.95	-0.17
	Ternary	1	0.65 (0.25)			
		2	0.61 (0.24)	.60	1.15	0.19

Note. All correlations were statistically significant ($p < .05$), whereas no *t* values were statistically significant.

Table 10

Correlations Between Mean Total Scores Across Tasks, Age in Months, and Reported GPA

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
Estimation													
1. AE	--												
2. RATIO	.86**	--											
3. Correctness	-.66**	-.46**	--										
Production													
4. AE	.73**	.57**	-.64**	--									
5. RATIO	-.60**	-.77**	.36**	-.43**	--								
6. Correctness	-.51**	-.47**	.63**	-.70**	.42**	--							
Reproduction													
7. AE	.54**	.43**	-.37**	.45**	-.41**	-.28*	--						
8. RATIO	-.08	-.11	.19	-.21	.27*	.16	-.11	--					
9. Correctness	-.46**	-.38**	.34**	-.47**	.48**	.33**	-.74**	.32**	--				
Discrimination													
10. Comparative	-.19	-.11	.25*	-.37**	.20	.25*	-.44**	.23*	.46**	--			
11. Equality	-.07	.12	.24*	-.38**	.13	.13	-.30**	.01	.32**	.42**	--		
12. Ternary	-.24*	-.09	.15	-.47**	.27*	.21	-.42**	.20	.51**	.67**	.65**	--	
Covariates													
13. AGE	.11	.12	-.07	.05	-.03	-.18	.22*	.29**	-.14	-.01	-.16	-.15	--
14. GPA	-.15	-.18	.06	-.10	.14	.04	-.23*	-.03	.21*	.09	.16	.19	-.26**

Note. AE = absolute difference; RATIO = estimated to target RATIO.
*p < .05 (two-tailed).
**p < .001 (two-tailed).

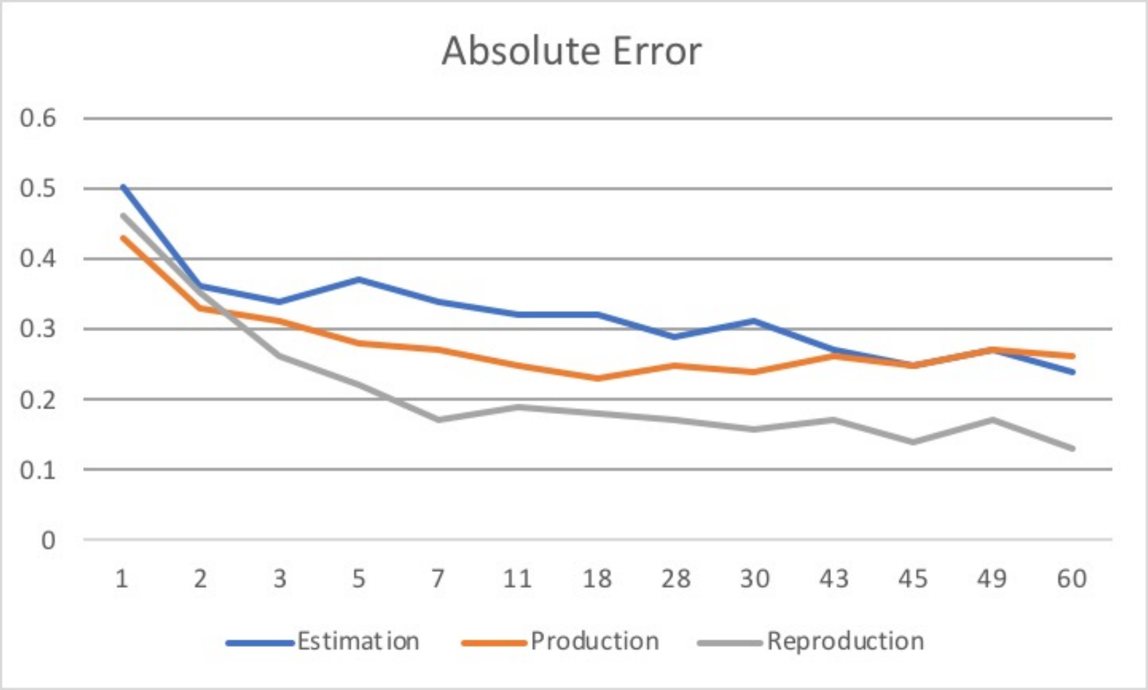

Figure 1. Average absolute error values for estimation, production, and reproduction tasks for all intervals. Smaller values indicate less error.

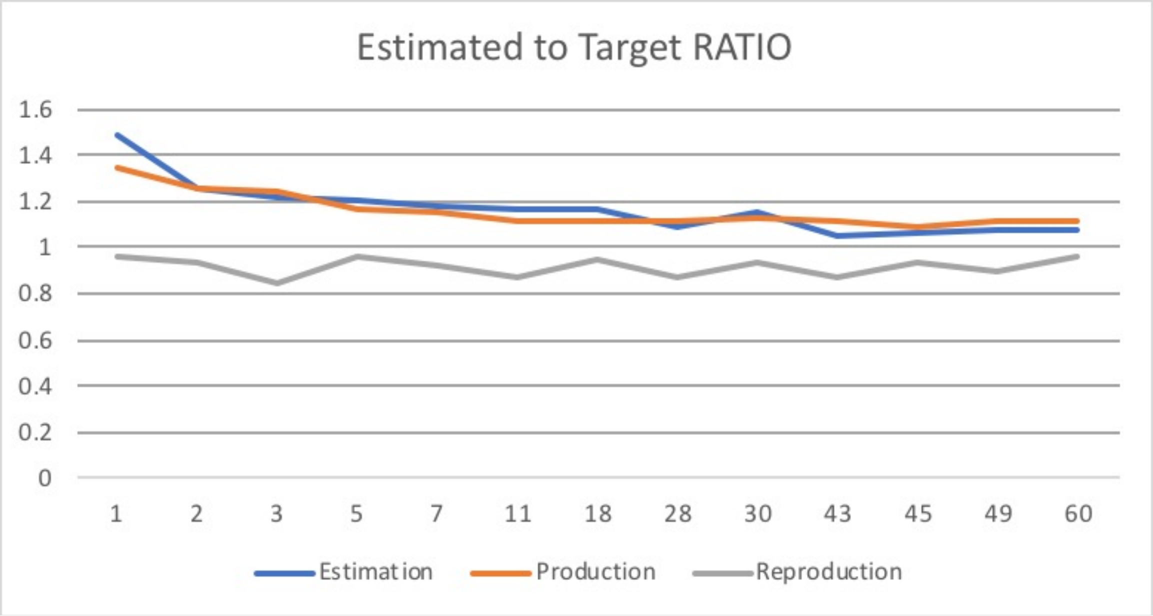

Figure 2. Average estimated to target RATIO values for estimation, production, and reproduction tasks for all intervals. Values above one represents overestimation, values less than one represent underestimation, and values closer to one represent higher accuracy. Production values have been reverse coded to accurate reflect the prior statement.

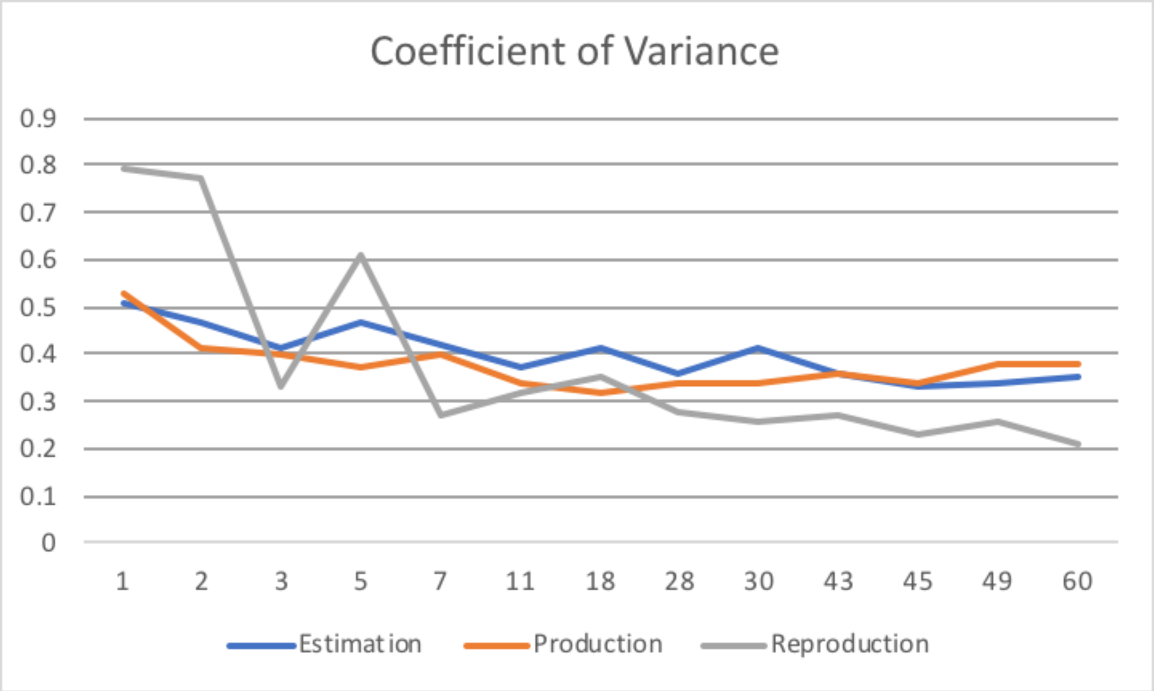

Figure 3. Average coefficient of variance values for estimation, production, and reproduction tasks for all intervals. Smaller values indicate less variability.

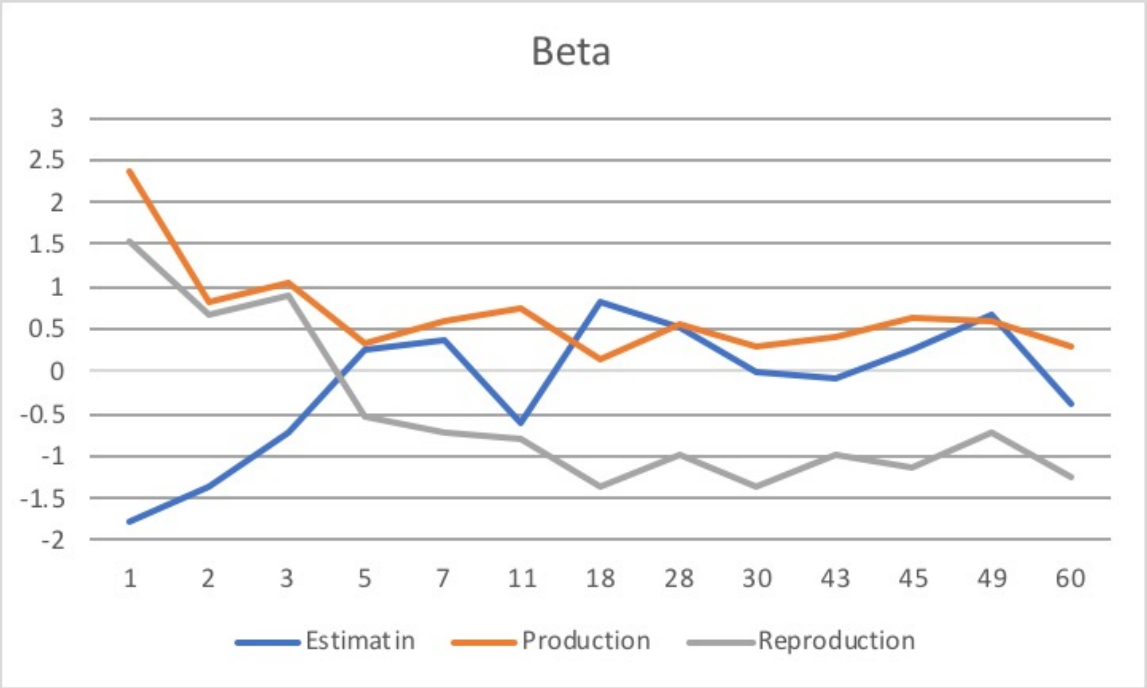

Figure 4. Beta (b) values for intervals across estimation, production, and reproduction tasks. Larger positive values indicate more difficult intervals and larger negative values indicate easier intervals.