frontiers in Psychology

METHODS
published: 10 January 2022
doi: 10.3389/fpsyg.2021.717568

Applying Evidence-Centered Design to Measure Psychological Resilience: The Development and Preliminary Validation of a Novel Simulation-Based Assessment Methodology

Sabina Kleitman[1], Simon A. Jackson[1], Lisa M. Zhang[1], Matthew D. Blanchard[1], Nikzad B. Rizvandi[2] and Eugene Aidman[3]*

[1] School of Psychology, The University of Sydney, Sydney, NSW, Australia, [2] Centre for Translational Data Science, The University of Sydney, Sydney, NSW, Australia, [3] Land Division, Defence Science and Technology Group, Edinburgh, SA, Australia

OPEN ACCESS

Edited by:
Kostas Karpouzis, Panteion University, Greece

Reviewed by:
Sofya K. Nartova-Bochaver, HSE University, Russia
Elina Roinioti, Panteion University, Greece

***Correspondence:**
Sabina Kleitman
sabina.kleitman@sydney.edu.au

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *01 June 2021*
Accepted: *10 December 2021*
Published: *10 January 2022*

Citation:
Kleitman S, Jackson SA, Zhang LM, Blanchard MD, Rizvandi NB and Aidman E (2022) Applying Evidence-Centered Design to Measure Psychological Resilience: The Development and Preliminary Validation of a Novel Simulation-Based Assessment Methodology. Front. Psychol. 12:717568. doi: 10.3389/fpsyg.2021.717568

Modern technologies have enabled the development of dynamic game- and simulation-based assessments to measure psychological constructs. This has highlighted their potential for supplementing other assessment modalities, such as self-report. This study describes the development, design, and preliminary validation of a simulation-based assessment methodology to measure psychological resilience—an important construct for multiple life domains. The design was guided by theories of resilience, and principles of evidence-centered design and stealth assessment. The system analyzed log files from a simulated task to derive individual trajectories in response to stressors. Using slope analyses, these trajectories were indicative of four types of responses to stressors: thriving, recovery, surviving, and succumbing. Using Machine Learning, the trajectories were predictive of self-reported resilience (Connor-Davidson Resilience Scale) with high accuracy, supporting construct validity of the simulation-based assessment. These findings add to the growing evidence supporting the utility of gamified assessment of psychological constructs. Importantly, these findings address theoretical debates about the construct of resilience, adding to its theory, supporting the combination of the "trait" and "process" approaches to its operationalization.

Keywords: assessment, simulation, resilience, evidence-centered design, machine learning

INTRODUCTION

Stress and adversity are an inevitable part of the human experience. However, not everyone is equally successful at overcoming potentially negative events. The construct of resilience offers one explanation for this. While there is no general agreement on what constitutes the psychological resilience, most researchers agree that it entails two core concepts: (1) the presence of a potential stressor, and (2) positive adaptation (see Fletcher and Sarkar, 2013 for a review). There is a growing body of research investigating different trajectories or divergent pathways of adjustment

in response to acute and chronic stressors (Bonanno and Diminich, 2013). These trajectories can include thriving, recovery, surviving, and succumbing (O'Leary and Ickovics, 1995; Carver, 1998). However, little empirical evidence exists for these trajectories in response to a clearly referenced, acute stressor. Such research, however, can provide evidence for the positive adaptation, the key process suggested to underlie mental resilience (see Fletcher and Sarkar, 2013 for a review).

Well-designed interactive technology (e.g., games, simulations), also referred to as Virtual Performance assessments (VPAs) show promise as a means to measure complex non-cognitive constructs including resilience, and have been referred to as the next-generation of assessment (de-Juan-Ripoll et al., 2018; Hao and Mislevy, 2018; Lee et al., 2019). A recent scoping review found interactive technology can deliver effective interventions to increase resilience (Pusey et al., 2020). The primary aim of this paper is to detail the design, development and validation of an immersive simulation-based assessment methodology to measure resilience, focusing on different trajectories within an acute stressful event.

The assessment framework draws from principles of evidence-centered design (Mislevy et al., 2003; Mislevy, 2013) and embedded stealth assessment (Shute, 2011), and is informed by well-established theories of resilience (Carver, 1998; Richardson, 2002). We created a dynamic simulation environment, where players had to adapt to and overcome various unexpected and challenging events to complete the task objective. Performance was reflective of different trajectories, and these pathways were validated against an existing gold-standard self-report measure of resilience, the Connor-Davidson Resilience Scale (CD-RISC; Connor and Davidson, 2003). The present research makes several novel theoretical and empirical contributions. First, we advance research on how game design elements, coupled with an evidence-based assessment framework, can make powerful assessment tools. Second, we give evidence of patterns of trajectories in response to adversity, emphasizing a holistic approach to resilience. Third, we contribute to recent advances in Machine Learning as a technique to analyze complex datasets and predict individual differences. Practically speaking, our work addresses the potential for gamified assessments as a supplementary method for measurement.

Assessment of Resilience

Given different conceptualizations of the resilience construct in the literature (as a *trait*, a *process*, and/or *outcome*; see Luthar et al., 2000; Fletcher and Sarkar, 2013; Fogarty and Perera, 2016 for reviews), varying methods have been used for its assessment. Common methods for the assessment of resilience include: (1) self-report scales and situational judgment tests (e.g., Windle et al., 2011; Pangallo et al., 2015; Teng et al., 2020), (2) indirect based on the presence of risk and positive adaptation (e.g., Luthar and Zelazo, 2003), and (3) measuring resistance or adaptation to negative life events, everyday stressors, or experimentally created stressors (e.g., Hildebrandt et al., 2016; Seery and Quinton, 2016). Despite the wide range of assessment formats, they are all evaluated against fundamental reliability and validity criteria. The psychometric properties of self-report methods of resilience are well-established, with some measures demonstrating excellent validity and reliability (e.g., CD-RISC; Connor and Davidson, 2003). Nevertheless, these scales are designed to capture resilience as a *trait*, although research has moved toward a broader *process* view, depicting the process through which internal and external factors interact to influence one's response to adversity (Pangallo et al., 2015). These scales also tend to approach resilience in basic terms of presence or absence of psychopathology (Bonanno and Diminich, 2013). However, this approach neglects the *distribution* of individual differences in resilience, that is, the different types or variations in responses to adversity. Researchers have also highlighted the need to develop multimodal methods of resilience, including moving toward objectively verified assessments, rather than a sole focus on resilience as a personality-like variable (Bonanno, 2012; Pangallo et al., 2015). Some self-report measures of resilience also have problems with discriminant validity as they fail to separate from related constructs (Fogarty et al., 2016; Fullerton et al., 2021), prompting some authors to advocate for an integrative process model of resilience (Fullerton et al., 2021). Game-based and simulation assessments offer a potential supplementary method for assessing the process of resilience, capturing responses to adverse or stressful events to produce an outcome. Integrating such an approach with the use of traditional self-report scales such as the CD-RISC to capture stable individual differences, would provide a more holistic depiction of the resilience process.

Self-report measures can also be affected by faking such as malingering, self-deception and impression management, especially in high-stakes contexts (Barrick and Mount, 1996; Bensch et al., 2019). This can impact both construct and criterion-related validity. Response distortion can negatively affect the factor structure of measures through differential item functioning (Zickar and Robie, 1999), and lead to misinformed decisions and predictions (Rosse et al., 1998). Different methods have been developed to detect response distortion. A common method is to correct scores using social desirability or lie scales. Yet, researchers have questioned the construct validity of these scales as they may confound response style with trait measurement (Ellingson et al., 2007). Other methods in detecting intentional response distortion include forced-choice response options (Cao and Drasgow, 2019), eye-tracking (van Hooft and Born, 2012), response latencies (Holden et al., 1992), and sophisticated mathematical algorithms to adjust scoring (Lee et al., 2014). However, these methods do not remedy the problem of reducing multiple sources or opportunities to distort responses. More recently, psychophysiological markers measuring responses to external stimuli (e.g., skin conductance) have been employed to indicate or predict resilience (Winslow et al., 2015). For instance, Walker et al. (2019) found individuals who rated themselves higher on trait resilience habituated quicker to acoustic startle stimuli (i.e., showed a reduction in the amplitude in skin conductance on repeated presentations). Although such approaches show promise for supplementing traditional methods, these studies tend to be resource-intensive, resulting in relatively small sample sizes typically below 40 participants (Berkovsky et al., 2019).

Using Games and Simulations to Assess Psychological Constructs

Modern technologies have allowed for the development of dynamic games and simulation environments (DiCerbo et al., 2017). These open-ended discovery spaces take many forms, including entertainment games, serious educational games, simulations, and virtual/augmented reality (see Marlow et al., 2016; Simmering et al., 2019; Taub et al., 2020 for reviews). Gamification is referred to as the use of game design elements in non-gaming contexts (Deterding et al., 2011). Games and simulations are defined as "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome" (Salen and Zimmerman, 2004, p. 80). In addition, Prensky (2001) includes goals, interaction, feedback, and representation as game design elements. In other words, a game necessarily involves creating an environment, objects, connections, rules and choices that allow the player to identify with the virtual character and immerse themselves into that game (Salen and Zimmerman, 2004).

Game-Based Assessment

The commercial use of game-based assessment has increased substantially in recent years. For example, personality (Barends et al., 2019; McCord et al., 2019) and non-cognitive constructs such as resilience, adaptability and flexibility (Georgiou et al., 2019; Nikolaou et al., 2019) have been assessed for personnel selection via games. However, the rapid expansion of commercial gamified assessments has attracted reservations about their psychometric properties. The lack of transparent and publicly available information about the design and data gathered from these instruments leave researchers with little evidence to evaluate their utility (Chamorro-Premuzic et al., 2016). Other criticisms include that they: lack a theory-driven design, tend to not be vetted in terms of scientific rigor, and have yet to demonstrate validity and reliability comparable with existing measures (Church and Silzer, 2016; Ihsan and Furnham, 2018).

Game-based assessments are also frequently used in educational settings, and their design is typically well-informed by evidence-based assessment frameworks. This type of assessment can support learning objectives and outcomes (Conrad et al., 2014; Hamari et al., 2014; Kapp et al., 2019). For instance, Shute et al. (2013, 2015) have demonstrated how games, coupled with evidence-based embedded assessment, can validly assess hard-to-measure constructs in educational contexts such as persistence (Ventura and Shute, 2013; DiCerbo, 2014), problem-solving, and creativity (Kim and Shute, 2015). Through log files, games allow us to record both the player's final choices and the decisions made before arriving at that choice (i.e., product and process data). This gives a rich bank of data that typically cannot be captured by closed-ended assessment tools (Ifenthaler et al., 2012).

Simulation-Based Assessment

Simulations are defined as "any artificial or synthetic environment that is created to manage the experiences of an individual with reality" (Marlow et al., 2016, p. 415). They share similar characteristics to games, however, can be distinguished in that they attempt to represent real-life situations and environments (Narayanasamy et al., 2006; Sauve et al., 2007; Marlow et al., 2016; Lamb et al., 2018). The level of realism or fidelity of simulated worlds can range from low to high, depending on how well the system represents reality (Beaubien and Baker, 2004). For instance, artificial fantasy in games vs. hyper-realistic 3D audio-visual rendering and motion platforms in large scale simulations. Similarly, user response options can range from purely symbolic such as keystrokes and mouse clicks, through to more realistic controls such as joysticks, steering wheels and pedals producing physically plausible changes, and all the way to locomotion in free-roam virtual reality applications. Representational and physical fidelity are important design considerations, and the cost-fidelity trade-off is a well-known problem. Whilst high fidelity may appear to be desirable, meaningful play actually comes from the interaction between players and the system of the game (Salen and Zimmerman, 2004). Thus, both low-and high-fidelity systems can foster meaningful interaction.

Using a combination of simulated environments and embedded performance assessment techniques, early systems measured behavioral task performance with minimal interference attributable to measurement itself. For instance, The Strategic and Tactical Assessment Record (STAR; Graham et al., 1985), produced a comprehensive array of perceptual and information processing parameters comparable with common laboratory measurements. One of its key advantages was that "all measurement procedures were embedded within the operations required to play a computer game" (Graham et al., 1985, p. 643). Shute (2011) describe the concept of *stealth assessment*, which involves embedding assessment unobtrusively and directly into the fabric of the gaming or simulated environment. Another strength of simulation technologies is the ability to continuously gather complex behavioral or performance data at a fine grain size, dynamically and in real-time (Aidman and Shmelyov, 2002; Johnson et al., 2016). For instance, a player can be immersed in a 3D augmented reality experience and presented with complex objects that move and rotate in space. In addition to recording their responses, it is also possible to capture their body and gaze movements, and how they rotate the objects around them. This richness of data collection makes simulated scenarios very powerful assessment tools. In this regard, simulated environments can exceed the usual physical and cost-prohibitive boundaries of space and time.

Nevertheless, well-designed simulations have many challenges and require substantial effort in the design and iteration phases. Some of these challenges include: (1) crafting appropriate and accurate digital environments to elicit the constructs of interest; (2) making valid inferences about the individual's behavior without disrupting the "free-flow" feel of the simulation; and (3) processing, synthesizing, extracting and interpreting the large quantities of data captured (de Klerk et al., 2015). The following sections detail how these challenges were addressed in designing the present simulation using an evidence-centered design framework.

Applying Evidence-Centered Design in Simulation-Based Assessment

For a simulation to be valid, we must consider psychometric principles from assessment design frameworks. Evidence-centered design (ECD; Mislevy et al., 2003, 2015; Mislevy, 2013) provides an excellent point of departure. ECD builds an evidentiary chain of reasoning (DiCerbo, 2017) and dates to Messick (1994, 1995), who lays out a series of questions for assessment design: (1) what knowledge, skills, or attributes should be assessed (are they valued by society)? (2) what behaviors reveal those constructs? and (3) what tasks or scenarios should elicit those behaviors? The ECD framework builds on the vision of Messick by formalizing these questions through three central models in the Conceptual Assessment Framework: Student (or Competency) Model, Task Model, and Evidence Model (Mislevy et al., 2003). The Student Model answers the question of *What are we measuring?* It defines the variables related to the knowledge, skills and abilities to be measured. The Task Model answers the question *Where do we measure it?* That is, what tasks elicit behaviors to produce the evidence? It describes the environment or scenarios in which individuals say or do something to produce evidence about the target construct. Finally, the Evidence Model bridges the Student and Tasks Models and answers the question *How do we measure it?* That is, what behaviors reveal different levels of the targeted construct? It analyses a player's interaction with, and responses to a given task. An evidence model consists of two parts: evidence rules and statistical model. Evidence rules take work product (e.g., a sequence of actions) that comes from the individual's interaction with the task as input, and produce observable variables (e.g., scores) as output, which are summaries of the work product. The statistical model expresses the relationship between the constructs of interest in the Student Model and the observable variables.

The Student Model: Defining Resilience

The rationale for investigating resilience as the focal construct is because stressful and aversive events continue to plague humans at each stage of the lifecycle (Bonanno and Diminich, 2013). How one responds to setbacks and uncertainty is critical for well-being and survival in a constantly evolving world. Resilience entails two core concepts: (1) the presence of a potential stressor, and (2) positive adaptation (Fullerton et al., 2021). The term "potential" is important as there are differences in how individuals react to a certain event; some people are overwhelmed by daily hassles whereas others thrive in testing experiences (Bonanno, 2012). This means that people do not uniformly perceive a potential stressor as stressful. Recent research also emphasizes the broad term *stressor*, which encompasses a range of demands that necessitate resilience (Fletcher and Sarkar, 2013). Earlier definitions used the narrow term *adversity*, which only captures significant negative life events. Davis et al. (2009, p. 1638) argued that "for most of us, the adversities we encounter do not constitute major disasters but rather are more modest disruptions that are embedded in our everyday lives." Thus, this study focuses on modest short-term stressors, rather than severe long-term hardship.

By integrating Carver's (1998) responses to adversity framework and Richardson's (2002) metatheory of resilience, four potential trajectories can result after experiencing a stressor: (1) resilient reintegration, where the individual returns to a higher level of homeostasis (thriving), (2) homeostatic reintegration, where the individual returns to their baseline level after decline (recovery), (3) reintegration with loss, leading to a lower level of functioning (surviving), and (4) dysfunctional reintegration, leading to maladaptive behaviors (succumbing). This is presented in **Figure 1**. In the present study, the model is applied to a discrete, time-bound adverse experience (although it can also be applied to a prolonged period of adversity; Carver, 1998; Bonanno and Diminich, 2013).

The Task Model: Design Elements

As mentioned earlier, the most common method to measure resilience is via self-report. A less commonly developed method is measuring resistance or adaptation to experimentally created stressors. Thus, we employed the latter method by aiming to capture responses to a clearly referenced adversity, where an individual is surrounded by high levels of emotional (and potentially physiological) stress. This "reactivity" approach involves measuring behavioral, subjective, or physiological responses to stimuli, whether naturalistic or experimental (Davydov et al., 2010). Adaptive reactivity would indicate resilience. The challenge was to foster high enough levels of stress and demand to elicit responses (not necessarily in magnitude, but in nature) to those experienced in real-world situations. To foster meaningful play, designing the formal system structures of the present simulation took into consideration core game elements, including objectives, procedures, rules, players and player interaction, conflict, and outcomes (Salen and Zimmerman, 2004).

The scenario was a driving simulation; however, it is important to note that the driving task is merely a medium or means to demonstrate the *method* of assessment (i.e., embedded evidence-based design framework), which is the focus of this research and

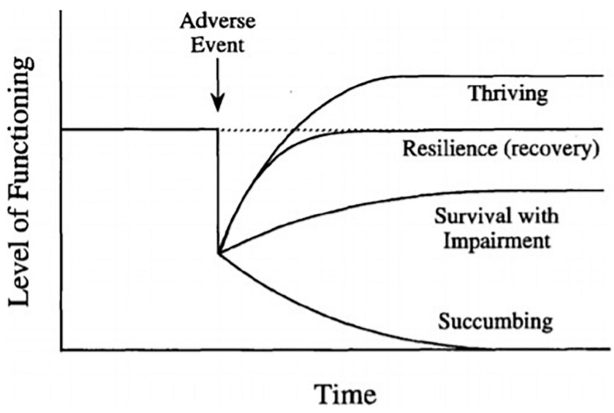

FIGURE 1 | Responses to adversity. From "Resilience and thriving: Issues, models, and linkages" by Carver (1998), *Journal of Social Issues*, p. 246. Copyright 1998 by Wiley.

described in detail below. That is, the assessment methodology is based on an ECD framework and intended to measure the target construct, thus should be independent of the medium. The driving scenario does not intend to measure driving or gaming experience *per se*, but rather, an individual's trajectory in response to stressors, which would demonstrate pathways of resilience as hypothesized in previous theories of resilience. Indeed, any type of simulated task could be used (e.g., flight, racing simulators), so long as it follows and is grounded in a strong theoretical assessment framework and methodology. A driving task was chosen out of practicality because they are middle ground in representational and physical fidelity, with physically consequential responses captured via steering action and pedals for acceleration and braking.

Players actively partook in a simulated training exercise for drivers of emergency vehicles (e.g., ambulance). Taking on the role of an emergency driver created a sense of urgency and time-pressure. Players drove around a metropolitan area designed as an urban grid, until they reached a final predetermined destination. This allowed players to be continuously directed along new routes in a relatively easily rendered space. Unbeknownst to them, they completed five different "laps" within the metropolitan area. Laps began and ended at the same location but took different paths around the city (see **Figure 2**).

Procedures, which are specific instructions of what actions to take, involved the player driving in a direction given by green arrows at every intersection. Directional arrows allowed the ability to create standardized and identical testing conditions. Hence, all players experienced the same events at approximately the same time. They could deviate from the arrows; however, they were instructed to stay on the desired path. Even if left unfollowed, participants could eventually return onto the directed path, although it would take longer for them to complete the task. If they did not follow the arrows, red no-go signs appeared, reminding them to make a U-turn which would lead back to the arrows. The *objective*, which aims to increase motivation and engagement, was to reach the destination as quickly and as safely as possible (i.e., maximize speed and minimize collisions). If present, another player controlled an Unmanned Aerial Vehicle (UAV) with a birds-eye view, in which they communicated information to the driver to help them drive faster and safer. For instance, instructing the driver when it was safe to use an incoming traffic lane. Depending on the helpfulness of the information, *player interaction* facilitated or inhibited the drivers' actions. *Rules*, which define conditions, restrict actions and determine effects on players, specified what players could and could not do. Players could break road rules (e.g., speeding, driving in the opposite lane), and could only drive on the road (e.g., could not drive on footpaths or through parks). Small road guards were added to prevent cars leaving the road. **Figure 3A** shows how this was managed via large invisible colliders.

Conflict, which emerges to prevent players from achieving their goal, was another core element. The type of conflict were physical obstacles encountered throughout the drive—what we term as "event probes." This event-based methodology is a systematic approach to designing simulated scenarios that are linked to the target constructs for assessment (Salas et al., 2009). Events must have clear start and end points to demarcate windows of meaningful data. Within these windows, the conditions players are exposed to are comparable. In line with stealth assessment, events were created and embedded with an emphasis on the scenario. Event probes appeared unexpectedly, disrupting players' actions such that a decrease in speed and increase in collisions was likely. Five events were designed: a falling lamp post, animals crossing the road, a stationary car blocking the driving lane, dense fog, and black ice. For the first three events, drivers needed to slow down and/or change lanes to avoid collisions. For the dense fog, drivers needed to reduce speed as long-distance vision was compromised. For the black ice, there was a loss of friction causing the car to slide, meaning drivers had to persist in accelerating and maneuvering the position of the car to avoid collisions (see **Figure 3B**).

Each event was presented once in each lap, meaning participants had five encounters with each event throughout the entire simulation. The insertion of event probes changed task demands in which participants were to adapt and overcome. The onset of each event was without warning, and embedded at a different location and order on each lap, in order to mitigate associating an event with a certain location. This is known as the "task-change paradigm" (Lang and Bliese, 2009). Event probes also provoked stress and frustration, requiring players to maintain composure. Emotional regulation is needed to redirect situational attention toward task demands (Niessen and Jimmieson, 2016). When no event probes took place, this was referred to as "probe-free periods." Nevertheless, participants still navigated heavy and changing traffic conditions. They also did not have an opportunity to practice as the aim was to capture dynamic responses under stressful and unpredictable conditions. In simulations aiming to capture other behaviors, it might be reasonable to inform players about the obstacles they are about to face and to provide them with the practice opportunities. This, however, would have compromised the purpose of the simulated environment.

The Evidence Model: Individual Performance Trajectories Using Slope Analysis

Log files record players' progress throughout the game or simulation. Analyzing log files involves parsing for relevant information and extracting performance indicators (Greiff et al., 2016; Hao et al., 2016; Hao and Mislevy, 2018). Identifying evidence that connects performance to the target construct requires well-structured log files and analysis methods (Hao et al., 2016). Three types data were collected:

1. A tab-separated values file for session-level information (e.g., start time of the simulation, session ID, player ID);
2. A tab-separated values file for time-stamped actions (e.g., collisions with other cars);
3. A directory of tab-separated values files, one for each value time-stamped and recorded frame-by-frame (e.g., angle of the steering wheel; position of the vehicle).

Evidence of the target constructs need to reflect how participants respond to the event probes, and how their responses change over the course of the simulation with each encounter.

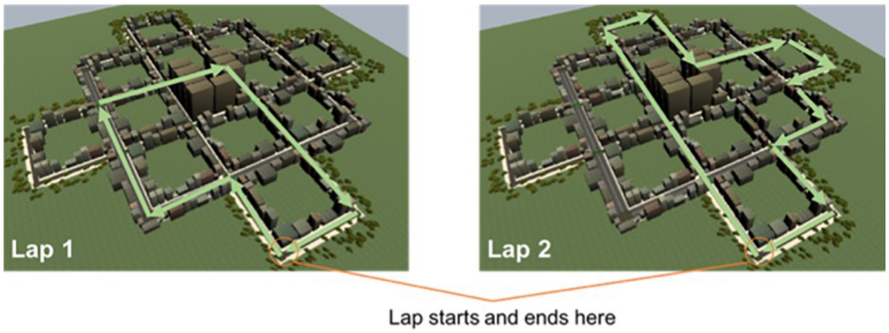

FIGURE 2 | Examples of paths taken by the driver on two consecutive laps.

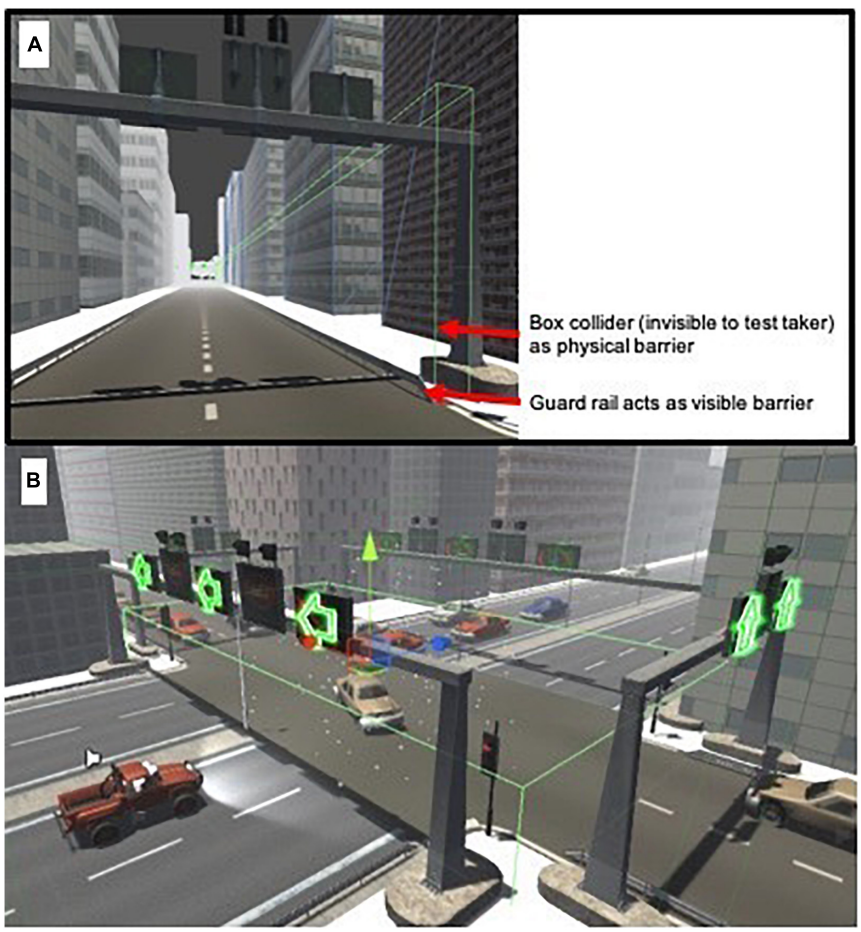

FIGURE 3 | (A) Guard rails and box colliders used to keep players on the road. **(B)** Birds-eye view of a player (red vehicle) entering the unpreventable (black ice) event. Spatially, the event is bound by a trigger collider appearing as fluorescent green lines (invisible to players). Upon entering, tire friction is reduced to near zero, causing the vehicle to slide. Upon exiting, tire friction returns to normal.

Speed and collisions were included as the key output indicators. They were chosen as variables of interest as they give measurable, unambiguous and objective outcomes of performance level. A response-time in complex simulation-based tasks is generally suggested to be in scoring (Lee et al., 2019). Speed was defined as an arbitrary digital miles per hour, and collisions were measured as the number of times a participant's vehicle collided with external objects. Below summarizes the process of how work product was taken as input (e.g., raw time-stamped data indicating a sequence of actions) and how observable variables (e.g., lap-level scores/measurements) were produced as output. This is known as the evidence rules.

Step 1: Lap-level measurements. The raw timestamped data included logs of collisions, speed, frame-by-frame recordings of the vehicle position, and steering wheel angle, accelerator and brake values. These data were used to create the following lap-level scores: collision frequency, average speed, time taken, distance traveled, and the mean and standard deviation of steering wheel, accelerator, and brake values. That is, each driver had five estimates for each of these variables corresponding to each of the five laps. Average speed and collisions were also estimated overall (i.e., across the entire drive), during event probes, and during probe-free periods. **Figure 4** shows the average speed and number of collisions of all drivers for each lap.

Step 2: Slope analysis. Derived variable analysis takes a collection of measurements (in this case, the five lap-level scores) and collapses them into a single meaningful summary feature (in this case, the slope) (Diggle et al., 2002). Slope analysis was used to determine individual performance trajectories. That is, we analyzed an individual's rate of change over time, with the slope as the key outcome. For each individual, the magnitude and direction of performance changes (Y) were estimated over baseline (intercept), linear, quadratic, and cubic slopes (see Equation 1). The model intercept (α) and slopes ($\beta_1 X_i$ *linear,* $\beta_2 X^2 i$ *quadratic* and $\beta_3 X^3 i$ *cubic*) were derived for each participant on each of the five estimates for both speed and collisions to represent a player's starting point and change over time. The statistical significance of slopes (beta weights) capture the strength of changes. The performance trajectories are indicative of individual differences in responses to stress (defined in the Student Model): thriving, recovery, surviving, or succumbing. This is known as the statistical model.

$$Y = \alpha + \beta_1 X_i(linear) + \beta_2 X^2{}_i(quadratic) + \beta_3 X^3{}_i(cubic) \pm \varepsilon \quad (1)$$

Slope Analysis: Linear, Quadratic, and Cubic Slopes

For speed, *thriving* is indicated by a strong positive linear slope (see **Figure 5A**). For collisions, *thriving* is indicated by a strong negative linear slope (see **Figure 5B**). These trends indicate the player continuously improved their performance (faster speeds and lower amount of collisions), despite obstacles encountered throughout the drive. These individuals adapted quickly to the changing situation; there was no or a relatively small loss of performance following obstacles or changes in the simulated environment in a lap 1, and they quickly relearned the changed situation. Thriving reflects decreased reactivity and faster recovery to subsequent stressful events, and a consistently high level of functioning (O'Leary and Ickovics, 1995; Carver, 1998). The strong the betas, the stronger the improvements across the five laps.

For speed, *recovery* is indicated by a strong positive quadratic beta, or a strong positive cubic beta (see **Figures 6A,B**). For collisions, *recovery*, is reflected by a strong negative quadratic beta, or a strong negative cubic beta (see **Figures 6C,D**). These trends demonstrate the ability to bounce back or return to former levels of functioning, after a decline in performance. After a downturn, they either return to baseline levels or continue toward an upward trend and function at a higher level than previously. After repeated experiences with the events, they are able to recover, should the stressor recur.

Weak or non-significant linear slope indicates merely *surviving* (i.e., strength of betas is relatively low). Whilst no significant improvements nor declines, these participants maintained homeostasis and were able to "just get past" the challenging events. Examples of surviving are displayed in **Figures 7A,B** for speed and collisions, respectively. These participants are able to withstand the challenging events, with no major deterioration nor improvement. The ability to withstand stressors is argued to be commonplace, emerging from the normative functions of human adaptation systems (Masten, 2001).

Finally, *succumbing* is indicated by various slopes. For speed, strong negative linear, quadratic, or cubic slopes indicate a poorer performance level relative to the initial baseline level (see **Figures 8A,B**). For collisions, strong positive linear, quadratic, or cubic slopes indicate succumbing (see **Figures 8C,D**). These participants show a continued downward slide (slower speeds and greater amount of collisions). They have a relatively large loss of performance following unexpected events, and they slowly or are unable to adapt to them. This leads to eventual succumbing after experience with repeated stressful events.

Validation With an Existing Resilience Measure

Following the approaches of Ventura and Shute (2013) and DiCerbo (2014), the individual performance trajectories were validated against an existing measure of resilience, the CD-RISC (Connor and Davidson, 2003). Connor and Davidson (2003) adopted Richardson's (2002) metatheory of resilience to develop their widely validated scale. It captures several trait-like aspects of resilience: the ability to adapt, to deal with stressors, to stay focus under stress, to handle unpleasant feelings, and the ability to stay on a task in the face of failure. The scale demonstrated strong psychometric properties. Reviews Windle et al. (2011) and Pangallo et al. (2015) have given the CD-RISC the highest rating in terms of quality criteria (validity and reliability), compared to other resilience scales. For validation, individual performance trajectories (linear, quadratic, and cubic slopes derived from lap-level measurements) are taken as inputs, and scores from the CD-RISC are taken as outcome criteria for predictive models. The experiment was conducted in a low-stakes environment to foster genuine responses on the CD-RISC, as the motivation to distort responses is more likely to occur in high-stakes situations such as job assessments (Donovan et al., 2014).

Research Questions and Aims

The overarching goal of the present research was to design a simulation-based assessment methodology to measure psychological resilience and provide data to evaluate its merits. Given the same experiences with the challenging simulated scenario and events, which people thrive and which are impaired? This study focuses on two key research questions to assess the validity of the simulation-based assessment:

1. Can slope analysis (i.e., rate of change over time in the simulated task) give empirical evidence of resilience

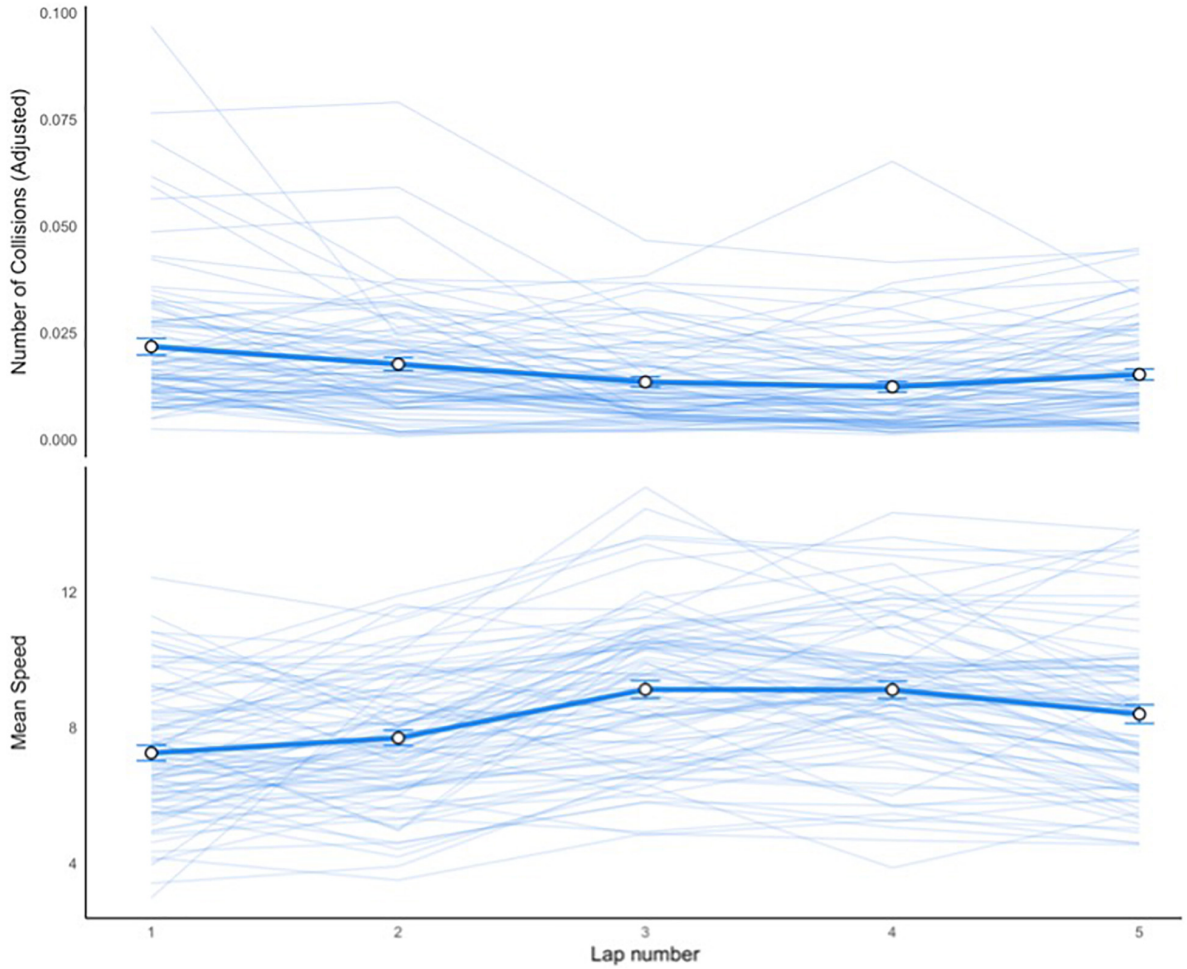

FIGURE 4 | Mean speed per lap. The thick lines represent the mean for all drivers with error bars representing 95% confidence intervals. Each thin line represents a single driver's number of collisions or average speed per lap. The number of collisions were adjusted to account for the different lap lengths. A ratio was calculated as the number of collisions per lap divided by the distance traveled per lap.

theories hypothesizing different individual trajectories of responses to stressors (thriving, recovery, surviving, succumbing)?

2. What is the relationship between individual trajectories and scores on an existing resilience measure (CD-RISC)?

METHODS

We report how we determined our sample size, all data exclusions, and all measures in the study.

Participants

109 undergraduate students participated and acted as drivers in the simulation, in return for course credit (59 females, 50 males; mean age = 20.10, *SD* = 4.72). Two participants were excluded because they did not complete the simulation. Four participants were excluded because they ignored instructions to follow the arrows directing their course, thus missing multiple events (more than 7). An additional 13 participants were excluded due to a hardware error which caused their teammate to observe different traffic conditions to what they experienced, making the teammate's instructions inaccurate. The final sample was composed of 90 participants who acted as drivers in the simulation (50 females, 40 males; mean age = 20.40, *SD* = 5.11). Machine Learning analyses (MLA) were utilized to quantify the accuracy of predictions. There are no agreed upon normative rules about how much data is needed to validate predictive models using MLA. Bzdok et al. (2018) recommends these analytics "when one is dealing with 'wide data,' where the number of input variables exceeds the number of subjects, in contrast to "long data" (2018, p. 233). The decision is typically based on the complexity of both the research question and the learning algorithm used in training and prediction, and the number of classes, input features, and model parameters used (see Raudys and Jain, 1991). Given the simplicity of the research questions and predictive model (based on eight main features used

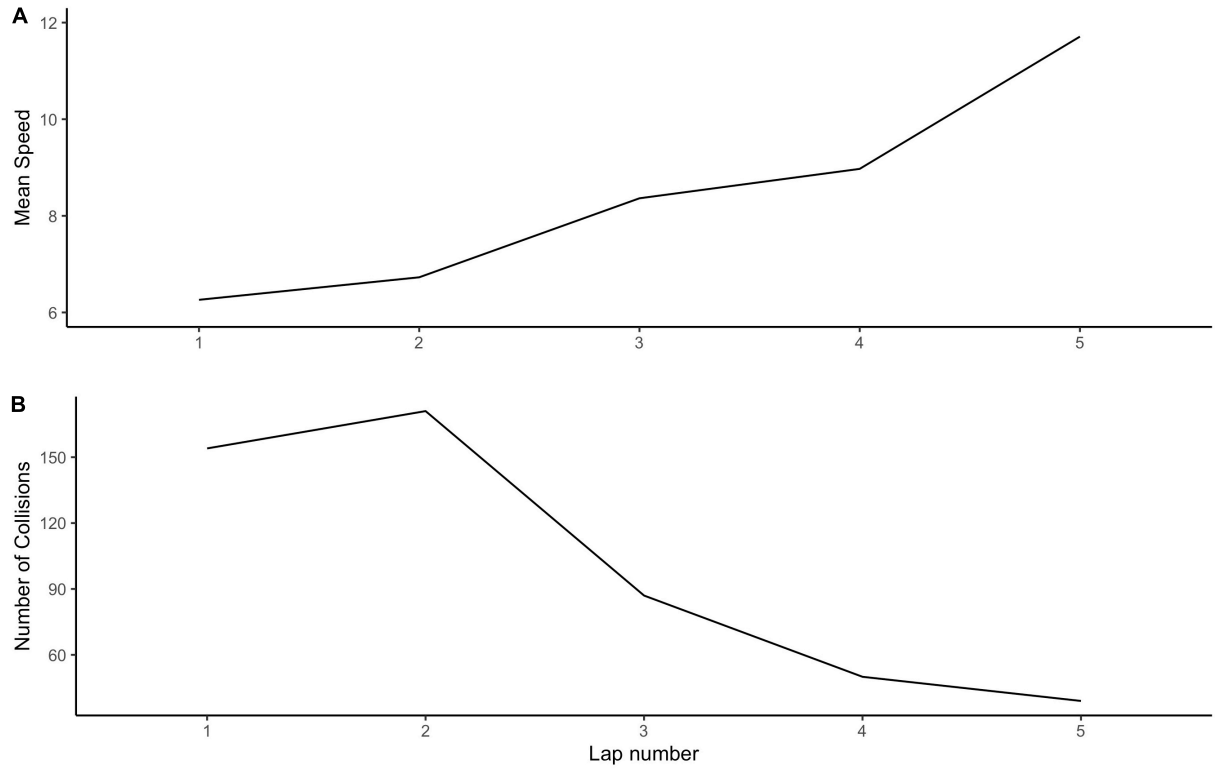

FIGURE 5 | Examples of different slopes indicating thriving.

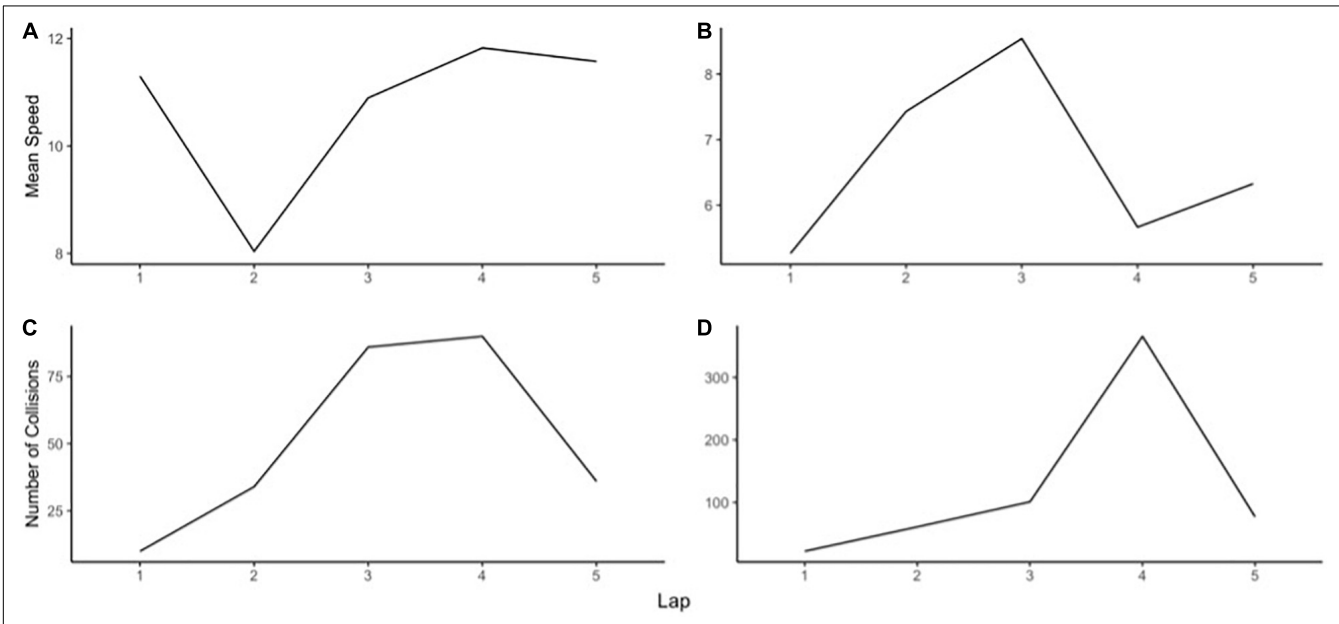

FIGURE 6 | Examples of different slopes indicating recovery.

separately for two metrics—speed and collision), and standard algorithms used, this sample size is more than satisfactory for the preliminary validation of the simulation before proceeding with a cross-validation of these results. This sample size was also adequate to examine a baseline linear regression model (with more than 10 people per feature/variable). More details are provided in the Machine Learning Analysis section of the results below.

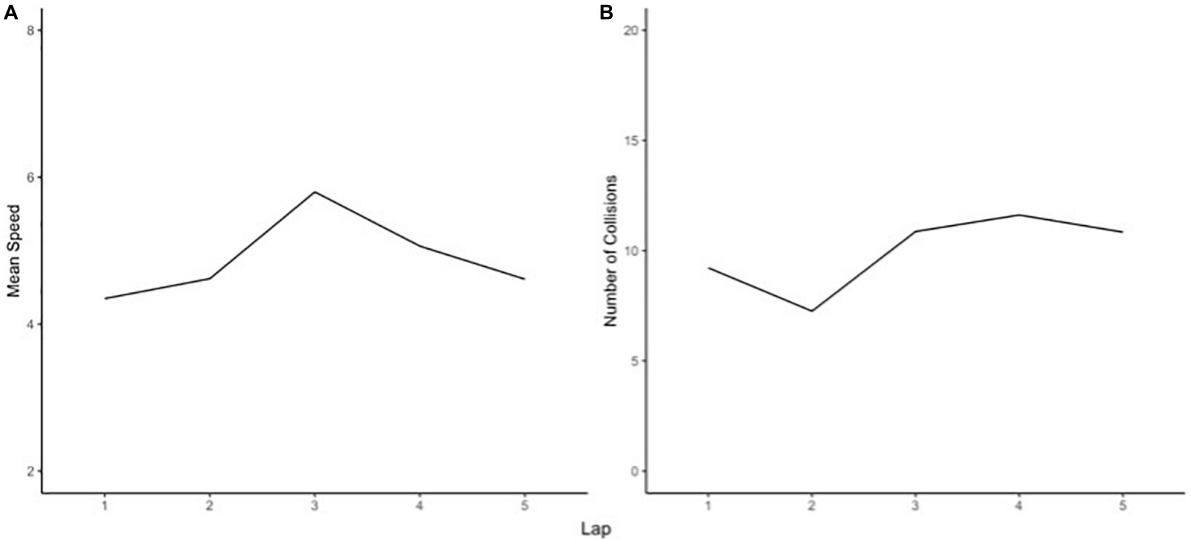

FIGURE 7 | Examples of different slopes indicating surviving.

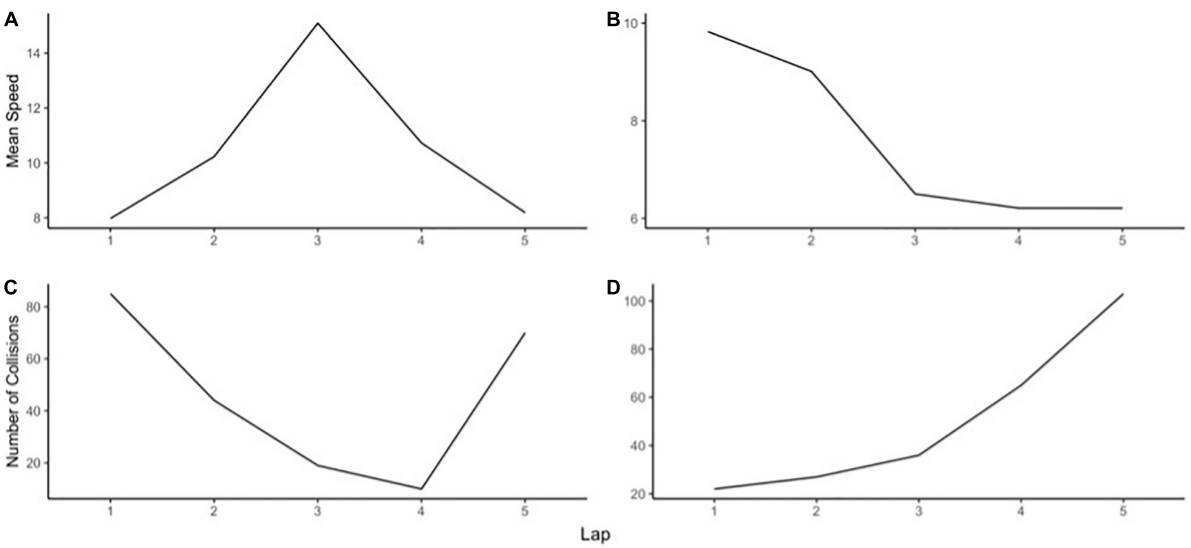

FIGURE 8 | Examples of different slopes indicating succumbing.

Measures

Simulation and Related Measures

Driving Simulation

The simulation is described above in the introduction.

NASA Task Load Index (NASA-TLX; Hart and Staveland, 1988). This is a 6-item measure of workload. It was administered immediately after, and in direct relation to the simulation. An example item is, "How mentally demanding was the task?" which was rated on a 7-point scale from (1) *very low* to (7) *very high*.

Simulator Sickness (Kennedy et al., 1993)

This questionnaire measures 16 symptoms of simulator sickness and cybersickness. There were three symptom clusters including Oculomotor (e.g., eyestrain, headache), Disorientation (e.g., dizziness, difficulty concentrating), and Nausea (e.g., stomach awareness, burping). Symptoms were scored on a 4-point scale from (0) *None* to (3) *Severe*. This questionnaire has demonstrated good internal consistency estimates (0.87; Bouchard et al., 2007).

Driving Experience and Gaming Intensity

Driving experience was recorded as the number of years driving a car, and gaming intensity was measured as the average number of hours spent playing video games in a single session.

Validation Measure

Connor-Davidson Resilience Scale (CD-RISC; Connor and Davidson, 2003). This measure consists of 25 items assessing the

ability to cope with stress and adversity. Items (e.g., "I am able to adapt to change") were rated on a 5-point Likert scale from (1) *not true at all* to (5) *nearly always true*. A higher total score indicates greater resilience. This scale has demonstrated good internal reliability (0.89; Connor and Davidson, 2003).

Procedure

The driving simulation was hosted on Unity game engine platform,[1] and presented on LG flat-screen monitors (43-inch screen size). The driving station consisted of Logitech G920 Driving Force Racing Wheel, Pedals and Playseat. All other measures were computerized and hosted on Qualtrics, a survey software platform. Sixty-seven (74.4%) drivers did the simulation with a teammate and twenty-three (25.6%) drivers completed it alone. Participants did not know each other before the study. They first completed demographic information (age, gender) and the simulation-related measures. They then completed the driving simulation (30 min), followed by the CD-RISC. Ethics approval was granted by Australian Defence Science and Technology Group Low-Risk Human Research Ethics Review (Protocol Number 07/415).

RESULTS

We begin by presenting descriptive statistics, providing a comprehensive examination of simulation-derived metrics, self-reported resilience, and evaluation of the simulation. Next, we present results of the slope analyses, including the proportion of individuals showing different performance trajectories. Given the complex distributional properties of the simulation-derived metrics and possible non-linear relationships between variables, we then used Machine Learning Analysis (MLA) to quantify the accuracy of predictions (Berkovsky et al., 2019; Jacobucci and Grimm, 2020). MLA was employed instead of traditional correlational analyses, because it can achieve relatively greater sensitivity compared to conventional techniques, which would likely deflate and/or fail to capture relationships between the variables (Koul et al., 2018). As MLA in psychological sciences is concerned with predictive accuracy, to optimize prediction we compared algorithms across degrees of complexity through the use of different linear and non-linear algorithms (including ridge regression, support vector machines, boosting, random forests) to examine relationships in the data (see Yarkoni and Westfall, 2017; Dwyer et al., 2018; Koul et al., 2018; Bleidorn and Hopwood, 2019; Jacobucci and Grimm, 2020; Orrù et al., 2020 for reviews). Accuracy of these algorithms were also compared to a baseline linear regression model to test whether they outperform the baseline.

[1]The simulation was originally designed to be Virtual Reality (VR) ready, and was piloted using the commercially released Oculus Rift on members of our research lab. However, we withdrew VR from testing of the main sample, due to virtual reality motion sickness. Whilst no participant in the pilot was physically ill, other symptoms (headaches, dizziness) from VR were adverse enough to interfere with the study.

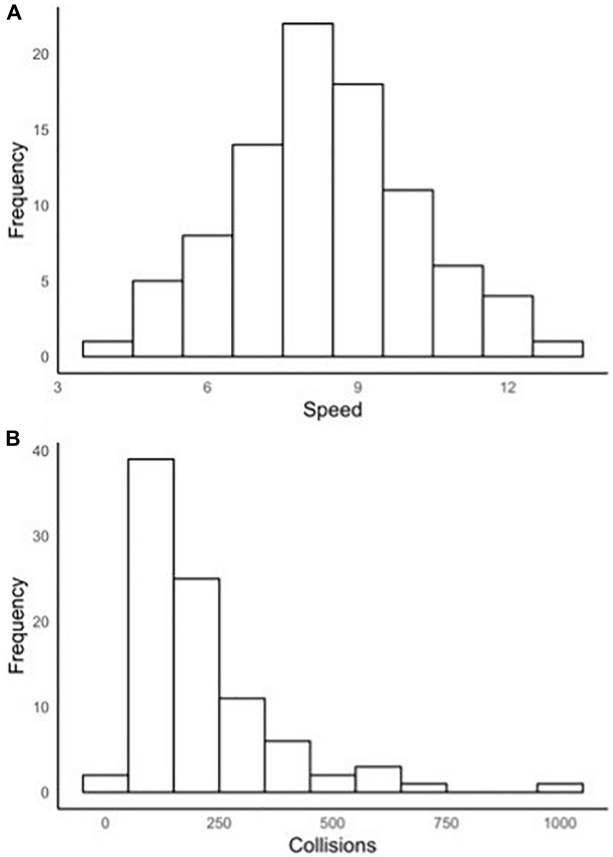

FIGURE 9 | (A,B) Histogram for A. speed & B. collisions.

Descriptive Statistics

Simulation-Derived Metrics

Figures 9A,B display the frequency distributions for overall speed and collisions in the simulation. Speed was normally distributed whilst collisions were positively skewed.

Table 1 presents descriptive statistics and Cronbach's alphas for average speed and number of collisions overall, during probe-free periods and during event probes. Internal consistency for speed and number of collisions overall across the five laps was high, and acceptable to good during probe-free periods and event probes. Paired sample *t*-tests were conducted to examine differences in speed and collisions during event probes vs. during probe-free periods. Participants had more collisions during event probes ($t = 4.05$, $p < 0.001$, Cohen's $d = 0.43$), however, there were no significant differences in average speed ($t = -1.42$, $p = 0.16$, Cohen's $d = -0.15$).

Simulation-Related Measures

Table 2 presents the descriptive statistics for NASA-TLX, driving experience, and gaming intensity. It appeared that participants were highly engaged and motivated to try their best in following the instructions to achieve the goal. Participants reported the

TABLE 1 | Descriptive statistics for simulation-derived metrics.

	Mean	*SD*	Range	Alpha	*t*-value	Cohen's *d*
Speed						
Overall	8.32	1.79	4.36–12.76	0.91		
Probe-free periods	9.29	2.12	4.98–14.93	0.87	–1.42	–0.15
Event probes	9.05	2.09	4.99–15.21	0.79		
Collisions						
Overall	211.77	162.49	31–966	0.80		
Probe-free periods	94.41	92.24	9–464	0.76	4.05***	0.43
Event probes	117.36	78.34	22–502	0.69		

****p < 0.001. Speed is reported as the average digital units per hour across the five laps. Collisions is reported as the total number of collisions across the five laps. For subsequent analyses using collisions, a ratio was calculated as the number of collisions per lap divided by the distance traveled per lap, to account for different lap lengths.*

TABLE 2 | Descriptive statistics for NASA-TLX, driving experience and gaming intensity.

Measure	Mean	*SD*	Range
NASA-TLX			
Effort	5.23	1.15	2–7
Frustration	3.83	1.78	1–7
Mental demand	4.84	1.39	1–7
Performance	3.20	1.10	1–5
Physical demand	3.80	1.60	1–7
Temporal demand	4.93	1.41	1–7
Driving and gaming			
Driving years	3.45	4.82	0–30
Gaming hours	1.24	1.09	0–5

highest score for effort, followed by mental and temporal (time-pressured) demands, then frustration, physical demand, and finally for performance.

Table 3 summarizes paired sample *t*-tests comparing pre- and post-simulation for 16 simulator sickness symptoms. Participants reported changes to half of the symptoms. Cohen's *d* were below medium effect sizes ($d < 0.50$) for 75% of affected symptoms. The most potent differences were sweating and fullness of the head with *d*-values over a medium effect size. The *Nausea* cluster of symptoms was most affected (in order of effect sizes): sweating, general discomfort, salivation, and nausea. The second most affected cluster was Disorientation (in order of effect sizes): fullness of head, and dizziness (eyes open). Two symptoms in the Oculomotor cluster (eyestrain, headache) were affected but their effect sizes were small. Participants reported no differences in fatigue, difficulty focusing and concentrating, blurred vision, dizziness (eyes closed) vertigo, stomach awareness, or burping.

An overall simulator sickness score was computed both pre- and post-simulation. The difference between the two scores was standardized. Majority of drivers (83.3%) were within 1 SD of the mean difference symptoms score. Four drivers (4.4%) reported an improvement in symptomatology in the size of 1 (3 drivers) and 3 (1 driver) SDs, respectively. Four drivers (4.4%) were 1 SD below the mean of the standardized difference score. Five and two drivers were 2 and 3 SDs below the mean, respectively. Thus, these seven people reported a notable detrimental change in symptomatology post-simulation.

Validation Measure

The mean score on the CD-RISC was 3.63 (*SD* = 0.40, Range = 2.20–4.84) and internal consistency was good ($\alpha = 0.86$). The frequency distributions represented a good spread of variance, rather than a skewed distribution (see **Figure 10**). In low-stakes situations such as that of the present research, people are less inclined to "fake good" or "fake bad," thus a normal distribution was observed, with an expected proportion of people reporting relatively low or high resilience levels.

Slope Analysis: Individual Performance Trajectories

Table 4 summarizes the percentage of people with different slopes for speed and collisions, during both the event probes and probe-free periods. Significant positive and negative slopes were investigated for the linear, quadratic, and cubic slopes. To allow for marginal error, the significance level was set at $p = 0.20$. Several findings are worth noting.

During the event probes, a small proportion of people had a positive linear slope for speed and a negative linear slope for collisions, which is indicative of *thriving*. With each encounter with an event, performance improved (speed increased, collisions decreased). These participants displayed a consistently high level of functioning despite constant challenges. During the event probes, a small proportion of people displayed positive quadratic and cubic slopes, respectively, for speed, and negative quadratic and cubic slopes, respectively, for collisions. These trends suggest the ability to *recover*—these people experienced a brief downward slide in performance but were able to bounce back from the embedded stressors and returned to their previous levels of functioning. A large proportion of participants (around 60–70%, with the percent varying for speed and collisions during events and probe-free periods) displayed non-significant trends (i.e., relatively weak betas or plateau-like slope). This suggests they were merely *surviving*—their performance level neither substantially increased nor decreased, but they maintained homeostasis. Finally, during the events, a minor proportion of people showed a negative linear slope for speed and a positive linear slope for collisions. This may suggest *succumbing*—these participants failed to adapt their behavior to each recurring event and could not recover after initial poor performance. With each encounter with events, there was a downslide in performance.

Machine Learning Analysis

One of the major applications of Machine Learning in psychological research is the development of models focused on predicting human behavior (Gonzalez, 2020). These analytics are recommended when *non-linear* relationships with well measured predictors are being modeled, which is a case for this research (Jacobucci and Grimm, 2020). Key machine learning techniques were considered in this study, including feature selection, cross-validation, and models/algorithms used. Feature selection is

TABLE 3 | Paired sample *t*-tests evaluating differences between simulator sickness symptoms pre- and post-simulation ($N = 90$, df_{89}).

	Mean pre	SD pre	Mean post	SD post	Mean diff.	SE mean diff.	*t*-value	Cohen's *d*
General discomfort	1.11	0.38	1.33	0.60	-0.22	0.07	–3.01**	0.40
Fatigue	1.24	0.53	1.34	0.62	–0.10	0.08	–1.32	–
Headache	1.07	0.25	1.18	0.41	–0.11	0.04	–2.58*	0.28
Eyestrain	1.16	0.42	1.29	0.57	–0.13	0.06	–2.16*	0.31
Difficulty focusing	1.11	0.35	1.17	0.46	–0.06	0.05	–1.15	–
Increased salivation	1.03	0.18	1.19	0.52	–0.16	0.05	–2.85**	0.38
Sweating	1.22	0.51	1.64	0.89	–0.42	0.09	–4.52**	0.60
Nausea	1.04	0.21	1.20	0.56	–0.16	0.06	–2.64*	0.35
Difficulty concentrating	1.10	0.40	1.17	0.43	–0.07	0.05	–1.23	–
Fullness of head	1.06	0.27	1.29	0.62	–0.23	0.07	–3.58**	0.53
Blurred vision	1.04	0.21	1.12	0.39	–0.08	0.04	–1.83	–
Dizziness (eyes open)	1.03	0.18	1.13	0.34	–0.10	0.04	–2.57*	0.36
Dizziness (eyes closed)	1.11	0.35	1.23	0.56	–0.12	0.07	–1.83	–
Vertigo	1.04	0.21	1.10	0.40	–0.06	0.04	–1.39	–
Stomach awareness	1.11	0.35	1.20	0.48	–0.09	0.05	–1.72	–
Burping	1.03	0.23	1.06	0.23	–0.02	0.03	–0.82	–

*$p < 0.05$; **$p < 0.01$.

selecting specific variables from a larger dataset to enhance accuracy and generalizability (Dwyer et al., 2018). Eight speed features and eight collisions features were selected to train models to predict scores on the CD-RISC for validation. The eight features were: first lap value, best lap value, worst lap value, overall average lap value, intercept term, linear term, quadratic term, and cubic term. For each of these features, there were estimates for during the event probes, during probe-free periods, and throughout the simulation overall. In order to mitigate the risk of overfitting, we used a train-test split to derive train and test subsets (7:3 ratio of train:test). The hyperparameters were only trained in the train subset, and the results were for the test subset only.

A baseline regression model was first conducted using a linear regression; the outcome variable was separately regressed onto the eight features. A baseline model provides a reference point to which to compare different machine learning algorithms, including the extent to which these algorithms add improvement over and above the baseline (Brownlee, 2014). Then, three common linear and non-linear Machine Learning algorithms were applied: Random Forest, Bayesian Ridge regression and Support Vector Machine.

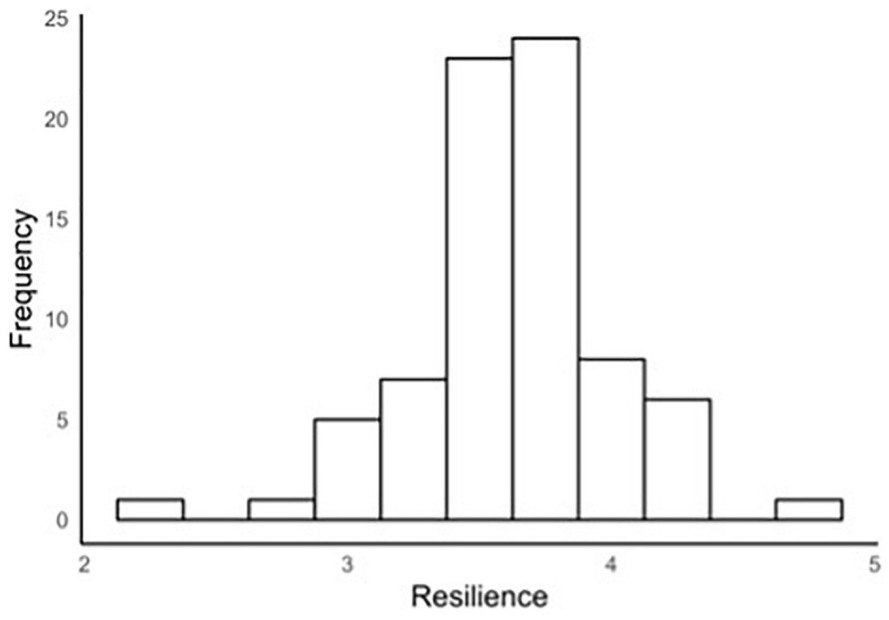

FIGURE 10 | Histogram of CD-RISC scores.

TABLE 4 | Percent (%) of participants with significant slopes (N = 90).

Slope		Speed		Collisions	
		Probe-free periods	Event probes	Probe-free periods	Event probes
Linear	Positive	14.44	11.11	11.11	7.78
	Negative	16.67	4.44	8.89	15.56
Quadratic	Positive	10.00	5.56	12.22	12.22
	Negative	11.11	12.22	6.67	8.89
Cubic	Positive	11.11	10.00	11.11	6.67
	Negative	11.11	5.56	2.22	8.89

Jacobucci and Grimm (2020) have recommended to compare predictive accuracy across algorithms with different degrees of complexity. The models were fitted and tested with predictions, and the Symmetric Mean Absolute Percentage Error (SMAPE) was used to quantify prediction accuracy. SMAPE is a widely used measure of accuracy, due to advantages of interpretability and scale-independency (Kim and Kim, 2016). These estimates are presented in **Table 5**. Each algorithm performed substantially better than the baseline linear regression model. While all methods produced good SMAPE estimates, attesting to the stability of the predictions, Bayesian Ridge regression algorithm produced the highest predictive accuracy (lowest error, with SMAPEs below 5%).

To demonstrate that the methodology was an assessment of resilience, rather than driving or gaming ability, both driving experience and gaming intensity were investigated as outcome variables in the machine learning models. The bottom of **Table 5** also presents the SMAPE estimates using the best performing algorithm, Bayesian Ridge regression (see **Supplementary Material** for results of other algorithms and baseline). Individual performance trajectories were weakly associated with driving experience, as it was predicted with relatively high error rates between 36.84 and 46.04%. Similarly, weak association was observed with gaming intensity: predictive error rates were high (between 36.57 and 50.15%). This suggests the simulation was assessing the target construct, resilience, and provides evidence of discriminant validity.

DISCUSSION

As modern technologies continue to progress, game and simulation design has correspondingly expanded for learning, assessment and training purposes. The present study aimed to detail the development, design and initial validation of a simulation-based assessment methodology to measure psychological resilience. The development of the simulation was guided by well-known resilience theories (Carver, 1998; Richardson, 2002), in addition to an evidence-centered design framework (Mislevy et al., 2003) and embedded stealth assessment (Shute, 2011). This study took a "reactivity" approach to measuring resilience, which involved systematic and deliberate exposure to stressful conditions (Davydov et al., 2010). These tests of reactivity to acute stressors have been proposed to assess different levels of adaptive or maladaptive responding. The findings demonstrate how game design elements, such as objectives, procedures, rules, and conflict, can be applied to make powerful assessment tools (Salas et al., 2009). Majority of players accomplished the objective of following arrows to reach the end destination. The success of this objective varied, with prominent individual differences in both absolute performance (speed and collisions) and trajectories (rate of change over the five laps). Conflict, crafted from the event probes, increased task difficulty by impeding players in reaching the objective of maximizing speed and minimizing collisions. Indeed, players collided more often during event probes, compared to probe-free periods. Players also reported the task being effortful, mentally demanding, and temporally demanding (fast-paced).

Performance in the simulation was recorded into log files unobtrusively—a key component of stealth assessment. Evidence extracted from these log files was used to identify different individual performance trajectories. That is, using derived variable analysis (Diggle et al., 2002), data from log files were transformed into lap-level measurements (i.e., average speed and collisions for each of the five laps), which were then collapsed into a meaningful summary feature—the slope. An individual's slope indicated their rate of change throughout the simulated task. Using slope analysis, different trends were observed, including linear, quadratic, and cubic slopes in both positive and negative directions. The strength of the slopes (i.e., betas) also held important information. One research question we aimed to investigate was whether these slopes indicated different responses to stressors, which represents a holistic approach to the resilience process (Carver, 1998; Richardson, 2002). We used MLA to build predictive models of resilience based on the simulation-embedded metrics. These analytics are especially recommended when non-linear relationships with well measured predictors are being modeled (Jacobucci and Grimm, 2020). Both are the case in this research.

For the simulation-embedded metrics, this indicated the stability of speed and collision tendencies of drivers within the simulation. Secondly, it appeared that a small proportion of participants demonstrated *thriving*, indicated by a positive linear slope for speed and a negative linear slope for collisions. Under stressful conditions, they were able to maintain emotional regulation and composure, and had no or minimal performance decrements. Another small percentage of people displayed *recovery*, shown by positive quadratic and cubic

TABLE 5 | Machine learning analysis: Predicting CD-RISC scores.

	Speed				Collisions			
	Probe-free periods		Event probes		Probe-free periods		Event probes	
Model	**SMAPE**	***SD***	**SMAPE**	***SD***	**SMAPE**	***SD***	**SMAPE**	***SD***
Baseline	7.90	7.63	7.91	7.08	7.78	7.11	7.92	7.12
Random forest	4.52	3.89	4.92	4.18	5.50	4.36	5.02	4.82
Bayesian Ridge	**4.04**	**3.91**	**3.80**	**2.80**	**3.85**	**5.15**	**3.21**	**2.74**
Support vector machine	4.22	4.16	4.42	5.32	4.13	5.68	3.63	3.60
Bayesian Ridge for								
Driving years	46.04	32.52	36.84	27.86	40.46	30.83	40.34	29.32
Gaming time	37.47	23.57	41.46	32.68	36.57	30.65	50.15	35.97

SMAPE, symmetric mean absolute percentage error. The best performing algorithm is bolded.

slope for speed and negative quadratic and cubic slopes for collisions. These people recovered from an initial setback in performance and returned to their previous level of functioning. The majority of people showed *survival*, as they had no significant slope (e.g., plateau-like slope). They were able to withstand the stressors and showed adaptive behavior to maintain their performance level. This process of survival and minimal impact has been argued to be commonplace and arises from the basic, normative functions of human adaptation systems (Masten, 2001; Bonanno and Diminich, 2013). Hence, it is not surprising that a large proportion of people showed the ability to maintain homeostasis. Finally, a small proportion of people showed trends indicative of *succumbing*. They were unable to adapt to the changing conditions or bounce back after initial poor performance, possibly indicating maladaptive reactivity.

The findings have practical implications for simulation-based training to support resilience, particularly for those showing a trajectory of succumbing, or those who seek to improve their resilience in the face of adversity (see Pusey et al., 2020, for a review). Existing resilience training programs have shown promise in contexts such as defense, workplace, and medical (see Leppin et al., 2014; Joyce et al., 2018 for reviews). There is potential to employ the present simulation-based assessment methodology in conjunction with resiliency programs—a randomized controlled design with follow-up measurements can ascertain the effectiveness of such training programs on raising performance levels to a point of thriving. Few studies have investigated this approach of building adversity into resilience training programs by systematic exposure to realistic simulations. To illustrate, in studies by Arnetz et al. (2009) and McCraty and Atkinson (2012), first-responders participated in realistic simulation scenarios (e.g., high-speed car pursuits). Compared to a control condition, those in resilience training programs reported less psychophysiological stress and better performance in the simulation. Indeed, moderate exposure to adversity with appropriate challenges may help individuals develop resilience, particularly for future stressful situations (Robertson et al., 2015). Thus, combining simulation-based assessment and training may be a promising paradigm to building resilience.

Another research question we aimed to answer was how these individual slopes would relate to an existing self-report measure of resilience (CD-RISC), for purposes of validation. Using machine learning techniques, the individual response trajectories were predictive of CD-RISC scores with high accuracy, provides evidence of construct validity. Error rates were below 5% for the best performing algorithm, Bayesian Ridge regression; and, importantly, were similar for other algorithms used, attesting to the stability of the predictions. Each of the machine learning models used outperformed a baseline linear regression model also tested in terms of predictive validity. Hence, as recommended by Jacobucci and Grimm (2020), machine learning approaches are more sensitive to modeling non-linear relationships, which can complement traditional statistical analysis techniques. To demonstrate discriminant validity, driving experience and gaming intensity were also placed in the models as outcome variables. These variables were predicted with relatively high error rates (above 35%), implying that behaviors in the simulation was not necessarily sensitive to the reports of driving or gaming ability. Thus, we reiterate that it is the design and validation of the *assessment methodology*, and *not* the driving task, which was the focus of this research. No predictions about the actual driving skills and abilities can be made based on the simulation. The driving scenario was merely a convenient medium to demonstrate how ECD could be applied to assess different resilience trajectories in response to stressors. It is of course possible and needed for future research to develop and test other simulated tasks using a variety of mediums (e.g., flight simulators), embedding the methodology presented to demonstrate that the method of assessment is independent of the medium.

These findings advance the way we construe resilience by demonstrating the dynamic process through which individuals adapt to stressors. The prediction of CD-RISC scores of trait resilience from the trajectories supports the role of stable individual differences in shaping one's adaptation to adverse events. However, the simulation goes beyond capturing these stable traits to assess real-time responses to stressors, i.e., the in-lab window into a *process* of resilience. We propose that capturing both is necessary to deepen our understanding of the psychological resilience construct.

Challenges, Limitations, and Future Directions

Despite holding promise as an alternative or supplementary assessment method, gamified assessments are still in its early stages. The present study is a step in the direction toward a "next-generation" of assessment. However, there is a need for further validation with other well-validated measures of resilience, as well as with measures of similar constructs in the nomological network (e.g., adaptability), and related real-world outcomes (Aidman, 2020). This newly developed simulation methodology must also be tested and replicated across multiple samples and contexts to determine the generalizability of the findings.

A significant challenge encountered in this study was the need for multidisciplinary expertise. For example, development of the simulation required software developers, game developers and 3D modelers; data management and analysis (particularly predictive modeling) required data scientists and programmers; and understanding and implementing the theory and constructs required psychologists and cognitive scientists. Not only must these personnel have expertise in their respective areas, but they must also develop their work output with consideration for other experts. For example, the software developer must program the simulation to output data logs which capture the target constructs as defined by the research psychologist. These data logs must also be suitably formatted for use by the data scientist.

Another challenge relates to issues of accessibility and feasibility, including the need for specialized equipment (both software and hardware). The present simulation used Unity development platform, however there are many other game engines such as Bohemia Interactive Simulations, Unreal Engine, and Godot. Specialized hardware (e.g., driving equipment, virtual reality headsets and equipment) can also increase costs and the need for dedicated space. Since game- and simulation-based assessments are more expensive and can take greater time to develop and validate, researchers must weigh the costs and benefits about their needs and goals. Thus, it remains a future research direction to explore ways to create gamified assessment protocols that are accessible, accurate, and cost-effective for both researchers and end-users, so their full benefit can be realized.

Specific to driving and other motion-based simulations, the potential impact of simulation sickness on performance must be fully investigated to limit its severity. High levels of simulator sickness can affect performance by confounding data and influence participant dropout rates (Brooks et al., 2010). Indicated by the simulator sickness questionnaire measured pre- and post-simulation, majority of the sample (83.3%) were not affected by the simulation. However, 7.8% of the sample reported a notable increase in symptoms. About 50% of the 16 symptoms showed significant changes, but only 25% of them were of notable effect size. Moreover, while five laps were sufficient to examine rate of change over time, the stability of the slope analysis could be strengthened by increasing the number of laps. However, longer exposures can produce more symptoms (Kennedy et al., 2000; Brooks et al., 2010). Perhaps in future iterations, the lap length could be reduced and consequently, the number of laps could be increased, without increasing the duration of the task. Additional research is also needed to determine the optimal length of a single exposure.

While we focused on speed and collision metrics, other indicators could be used to measure resilient responses. This could include, for instance, intentional lane-changing strategies (data points that collect the location of the vehicle), and maintaining emotional composure (i.e., modulating the level of control over one's responses to match environmental demands). Recording psychophysiological (e.g., skin conductance, heart rate variability) whilst completing the simulation may also give valuable information. Several studies have investigated indicators of physiological arousal whilst participants completed a stressful laboratory task (see Walker et al., 2017, for a review). For instance, Hildebrandt et al. (2016) placed participants in a threatening and changing immersive virtual environment while recording skin conductance, and found self-rated resilience predicted arousal during the sustained experience of threat. Other studies have investigated regulation and recovery from stressors via startle responses (Walker et al., 2019) or matching emotional responses to changing stimuli (Waugh et al., 2011). Employing biomarkers in conjunction with game- and simulation-based assessment protocols can act as an additional source of validation. It would be interesting to determine whether those with performance levels indicating thriving or recovery show better regulation of psychophysiological arousal (e.g., lower skin conductance)—this may provide evidence that resilient people can regulate and change their affective and physiological responses to match the demands of changing environmental circumstances.

This study recruited university students in a low-stakes context. While the sample size was appropriate, a larger sample is recommended to replicate these results. The promising aspect of this research is the stability of predictions across different ML algorithms; thus we anticipate that these results will replicate on a larger sample. Also, whilst there are limits of generalizability due to the sample characteristics, the results still show promising utility of the simulation-based assessment. Future studies should examine specific samples where resilience is critical for success (e.g., elite athletes, defense personnel, business managers). On a related note, the incremental and criterion-related validity of this methodology is yet to be established, above and beyond existing measures of resilience. Iterative validation of game- and simulation-based assessments includes determining their utility in predicting real-world outcomes (and being implemented in high-stakes environments; Georgiou et al., 2019; Nikolaou et al., 2019). These outcomes could be subjective or objective, for instance, attrition rates and posttraumatic stress trajectories in military personnel (Bonanno, 2012); game performance consistency and injury rehabilitation in competitive athletes (Sarkar and Fletcher, 2013); or job performance and burnout in employees such as healthcare professionals (Robertson et al., 2015). Finally, Machine Learning approaches typically require large sample sizes to train the data.

Future research is recommended to obtain larger samples to strengthen model predictions.

CONCLUSION

This work is contributing to the growing literature on gamified psychometrics, and to the theory of mental resilience, integrating the process model of resilience to its measurement. Game- and simulation-based assessment is a nascent research area, with promising progress being made toward their theory, design, validation, and implementation for end-users in various contexts (e.g., education, defense, organizational). Well-designed games and simulations provide opportunities to assess "hard-to-measure" constructs, particularly those regarded as twenty-first century skills (OECD, 2018); not to replace, but to supplement traditional measures and methods. Data can be collected continuously and unobtrusively (stealth assessment), providing a rich bank of information about individuals' skills, abilities, and attributes. However, iterative, and rigorous validation is necessary for the utility of gamified assessments to be fully achieved. We look forward to the continued investigation of gamified methods that may change how we think about assessment.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the release will need to be approved via DSTG internal review process. Requests to access the datasets should be directed to corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Defence Science and Technology Low-Risk Human Research Ethics Panel (Protocol Number 07/415). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SK, SJ, and EA designed the study. SJ built the simulation. SJ, MB, and LZ collected the data. SK, SJ, MB, and NR analyzed the data. SJ, LZ, and SK wrote the manuscript. SK and EA revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Defence Science and Technology Group, Australia, Research Agreements 5861, 8050, and 6082.

ACKNOWLEDGMENTS

We acknowledge Dr. Luke Thiele's assistance with software/hardware integration, and members of the CODES lab at the University of Sydney for their valuable comments on earlier drafts of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.717568/full#supplementary-material

REFERENCES

Aidman, E. (2020). Cognitive fitness framework: towards assessing, training and augmenting individual-difference factors underpinning high-performance cognition. *Front. Hum. Neurosci.* 13:466. doi: 10.3389/fnhum.2019.00466

Aidman, E., and Shmelyov, A. G. (2002). Mimics: a symbolic conflict/cooperation simulation program, with embedded protocol recording and automatic psychometric assessment. *Behav. Res. Methods Instr. Comput.* 34, 83–89. doi: 10.3758/BF03195426

Arnetz, B. B., Nevedal, D. C., Lumley, M. A., Backman, L., and Lublin, A. (2009). Trauma resilience training for police: psychophysiological and performance effects. *J. Police Crim. Psychol.* 24, 1–9. doi: 10.1007/s11896-008-9030-y

Barends, A. J., de Vries, R. E., and van Vugt, M. (2019). Gamified personality assessment: virtual behavior cues of honesty-humility. *Z. Psychol.* 227, 207–217. doi: 10.1027/2151-2604/a000379

Barrick, M. R., and Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *J. Appl. Psychol.* 81, 261–272. doi: 10.1037/0021-9010.81.3.261

Beaubien, J. M., and Baker, D. P. (2004). The use of simulation for training teamwork skills in health care: how low can you go? *Qual. Saf. Health Care* 13, i51–i56. doi: 10.1136/qhc.13.suppl_1.i51

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., and Ziegler, M. (2019). The nature of faking: a homogeneous and predictable construct? *Psychol. Assess.* 31, 532–544. doi: 10.1037/pas0000619

Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J., et al. (2019). "Detecting personality traits using eye-tracking data," in *Proceedings of CHI Conference on Human Factors in Computing Systems Proceedings*. (Minato City: Yokohama), 1–12.

Bleidorn, W., and Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Pers. Soc. Psychol. Rev.* 23, 190–203. doi: 10.1177/1088868318772990

Bonanno, G. A. (2012). Uses and abuses of the resilience construct: loss, trauma, and health-related adversities. *Soc. Sci. Med.* 74, 753–756. doi: 10.1016/j.socscimed.2011.11.022

Bonanno, G. A., and Diminich, E. D. (2013). Annual research review: positive adjustment to adversity-trajectories of minimal-impact resilience and emergent resilience. *J. Child Psychol. Psychiatry* 54, 378–401. doi: 10.1111/jcpp.12021

Bouchard, S., Robillard, G., and Renaud, P. (2007). Revising the factor structure of the simulator sickness questionnaire. *Annu. Rev. Cyber Ther. Telemed.* 5, 117–122.

Brooks, J. O., Goodenough, R. R., Crisler, M. C., Klein, N. D., Alley, R. L., and Wills, R. F. (2010). Simulator sickness during driving simulation studies. *Accid. Anal. Prevent.* 42, 788–796. doi: 10.1016/j.aap.2009.04.013

Brownlee, J. (2014). *How To Get Baseline Results And Why They Matter*. Available online at: https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/ (accessed June 27, 2017).

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nat. Methods* 15, 233–234.

Cao, M., and Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *J. Appl. Psychol.* 104, 1347–1368. doi: 10.1037/apl0000414

Carver, C. S. (1998). Resilience and thriving: issues, models, and linkages. *J. Social Issues* 54, 245–266. doi: 10.1111/0022-4537.641998064

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., and Hogan, R. (2016). New talent signals: shiny new objects or a brave new world? *Ind. Organ. Psychol.* 9, 621–640. doi: 10.1017/iop.2016.6

Church, A. H., and Silzer, E. (2016). Are we on the same wavelength? Four steps for moving from talent signals to valid talent management applications. *Ind. Organ. Psychol.* 9, 645–654.

Connor, K. M., and Davidson, J. R. T. (2003). Development of a new resilience scale: the connor-davidson resilience scale (CD-RISC). *Depress. Anxiety* 18, 76–82. doi: 10.1002/da.10113

Conrad, S., Clarke-Midura, J., and Klopfer, E. (2014). A framework for structuring learning assessment in a massively multiplayer online educational game: experiment centered design. *Int. J. Game Based Learn.* 4, 37–59. doi: 10.4018/IJGBL.2014010103

Davis, M. C., Luecken, L., and Lemery-Chalfant, K. (2009). Resilience in common life: introduction to the special issue. *J. Pers.* 77, 1637–1644. doi: 10.1111/j.1467-6494.2009.00595.x

Davydov, D. M., Stewart, R., Ritchie, K., and Chaudieu, I. (2010). Resilience and mental health. *Clin. Psychol. Rev.* 30, 479–495. doi: 10.1016/j.cpr.2010.03.003

de Klerk, S., Veldkamp, B. P., and Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: a systematic review and a Bayesian network example. *Comput. Educ.* 85, 23–34. doi: 10.1016/j.compedu.2014.12.020

de-Juan-Ripoll, C., Soler-Domínguez, J. L., Guixeres, J., Contero, M., Álvarez Gutiérrez, N., and Alcañiz, M. (2018). Virtual reality as a new approach for risk taking assessment. *Front. Psychol.* 9:2532. doi: 10.3389/fpsyg.2018.02532

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. E. (2011). *Gamification: Towards a Definition*. Vancouver, BC: CHI.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educ. Technol. Soc.* 17, 17–28. doi: 10.1037/a0023786

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *J. Appl. Test. Technol.* 18, 7–18.

DiCerbo, K. E., Shute, V., and Kim, Y. J. (2017). "The future of assessment in technology rich environments: psychometric considerations of ongoing assessment," in *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy*, eds J. M. Spector, B. Lockee, and M. Childress (Berlin: Springer).

Diggle, P. J., Heagerty, P. J., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Donovan, J. J., Dwight, S. A., and Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *J. Bus. Psychol.* 29, 479–493. doi: 10.1007/s10869-013-9318-5

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annu. Rev. Clin. Psychol.* 7, 91–118. doi: 10.1146/annurev-clinpsy-032816-045037

Ellingson, J. E., Sackett, P. R., and Connelly, B. S. (2007). Personality assessment across selection and development contexts: insights into response distortion. *J. Appl. Psychol.* 92, 386–395. doi: 10.1037/0021-9010.92.2.386

Fletcher, D., and Sarkar, M. (2013). Psychological resilience: a review and critique of definitions, concepts, and theory. *Eur. Psychol.* 18, 12–23. doi: 10.1027/1016-9040/a000124

Fogarty, G. J., Furst, A., Thomas, P., and Perera, N. (2016). Evaluating measures of sport confidence and optimism. *Meas. Phys. Educ. Exerc. Sci.* 20, 81–92. doi: 10.1080/1091367X.2015.1111220

Fogarty, G. J., and Perera, N. (2016). Resilience: distinct construct or conglomerate of existing traits? *Ind. Organ. Psychol.* 9, 422–429.

Fullerton, D., Zhang, L., and Kleitman, S. (2021). An integrative process model of resilience in an academic context: resilience resources, coping strategies, and positive adaptation. *PLos One* 2:e0246000. doi: 10.1371/journal.pone.0246000

Georgiou, K., Gouras, A., and Nikolaou, I. (2019). Gamification in employee selection: the development of a gamified assessment. *Int. J. Select. Assess.* 27, 91–103. doi: 10.1111/ijsa.12240

Gonzalez, O. (2020). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychol. Methods* [Epub ahead of print]. doi: 10.1037/met0000317

Graham, G., Cook, M., Cohen, M., Phelps, M., and Gerkovich, M. (1985). STAR: a unique embedded performance assessment technique. *Behav. Res. Methods Instr. Comput.* 17, 642–651.

Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Hum. Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095

Hamari, J., Koivisto, J., and Sarsa, H. (2014). "Does gamification work? – A literature review of empirical studies on gamification," in *Procedings of the 47th Hawaii International Conference on System Science*, Waikoloa.

Hao, J., and Mislevy, R. J. (2018). The evidence trace file: a data structure for virtual performance assessments informed by data analytics and evidence-centered design. *ETS Res. Rep. Ser.* 2018, 1–16. doi: 10.1002/ets2.12215

Hao, J., Smith, L., Mislevy, R. J., von Davier, A., and Bauer, M. (2016). Taming log files from game/simulation-based assessments: data models and data analysis tools. *ETS Res. Rep. Ser.* 2016, 10–16. doi: 10.1002/ets2.12096

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," in *Advances in Psychology, 52. Human Mental Workload*, eds P. A. Hancock and N. Meshkati (Amsterdam: Elsevier), 139–183.

Hildebrandt, L. K., McCall, C., Engen, H. G., and Singer, T. (2016). Cognitive flexibility, heart rate variability, and resilience predict fine-grained regulation of arousal during prolonged threat. *Psychophysiology* 53, 880–890. doi: 10.1111/psyp.12632

Holden, R. R., Kroner, D. G., Fekken, G. C., and Popham, S. M. (1992). A model of personality test item response dissimulation. *J. Pers. Soc. Psychol.* 63, 272–279. doi: 10.1037/0022-3514.63.2.272

Ifenthaler, D., Eseryel, D., and Ge, X. (2012). "Assessment in game-based learning," in *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, eds D. Ifenthaler, D. Eseryel, and X. Ge (Berlin: Springer), 1–8.

Ihsan, Z., and Furnham, A. (2018). The new technologies in personality assessment: a review. *Consult. Psychol. J.* 70, 147–166. doi: 10.1037/cpb0000106

Jacobucci, R., and Grimm, K. J. (2020). Machine learning and psychological research: the unexplored effect of measurement. *Perspect. Psychol. Sci.* 15, 809–816. doi: 10.1177/1745691620902467

Johnson, K., Aidman, E., Paech, G. M., Pajcin, M., Grant, C., LaValle, C., et al. (2016). Early morning repeat-dose caffeine mitigates driving performance impairments during 50 hours of sleep deprivation. *Road Transp. Res. J. Austral. N. Zeal. Res. Pract.* 25:3.

Joyce, S., Shand, F., Tighe, J., Laurent, S. J., Bryant, R., and Harvey, S. B. (2018). Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. *BMJ Open* 8:e017858. doi: 10.1136/bmjopen-2017-017858

Kapp, F., Spangenberger, P., Kruse, L., and Narciss, S. (2019). Investigating changes in self-evaluation of technical competences in the serious game serena supergreen: findings, challenges and lessons learned. *Metacogn. Learn.* 14, 387–411. doi: 10.1007/s11409-019-09209-4

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3, 203–220. doi: 10.1207/s15327108ijap0303_3

Kennedy, R. S., Stanney, K. M., and Dunlap, W. P. (2000). Duration and exposure to virtual environments: sickness curves during and across sessions. *Presence Teleoperat. Virt. Environ.* 9, 463–472.

Kim, S., and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *Int. J. Forecast.* 32, 669–679. doi: 10.1016/j.ijforecast.2015.12.003

Kim, Y. J., and Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Comput. Educ.* 87, 340–356. doi: 10.1016/j.compedu.2015.07.009

Koul, A., Becchio, C., and Cavallo, A. (2018). PredPsych: a toolbox for predictive machine learning-based approach in experimental psychology research. *Behav. Res. Methods* 50, 1657–1672. doi: 10.3758/s13428-017-0987-2

Lamb, R. L., Annetta, L., Firestone, J., and Etopio, E. (2018). A meta-analysis with examination of moderators of student cognition, affect, and learning outcomes

while using serious educational games, serious games, and simulations. *Comput. Hum. Behav.* 80, 158–167. doi: 10.1016/j.chb.2017.10.040

Lang, J. W. B., and Bliese, P. D. (2009). General mental ability and two types of adaptation to unforeseen change: applying discontinuous growth models to the task-change paradigm. *J. Appl. Psychol.* 94, 411–428. doi: 10.1037/a0013803

Lee, P., Stark, S., and Chernyshenko, O. S. (2014). Detecting aberrant responding on unidimensional pairwise preference tests: an application of lz based on the Zinnes–Griggs ideal point IRT model. *Appl. Psychol. Meas.* 38, 391–403. doi: 10.1177/0146621614526636

Lee, Y.-H., Hao, J., Man, K., and Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. *Front. Psychol.* 10:906. doi: 10.3389/fpsyg.2019.00906

Leppin, A. L., Bora, P. R., Tilburt, J. C., Gionfriddo, M. R., Zeballos-Palacios, C., Dulohery, M. M., et al. (2014). The efficacy of resiliency training programs: a systematic review and meta-analysis of randomized trials. *PLoS One* 9:e111420. doi: 10.1371/journal.pone.0111420

Luthar, S. S., Cicchetti, D., and Becker, B. (2000). The construct of resilience: a critical evaluation and guidelines for future work. *Child Dev.* 71, 543–562. doi: 10.1111/1467-8624.00164

Luthar, S. S., and Zelazo, L. B. (2003). "Research on resilience: an integrative review," in *Resilience and Vulnerability: Adaptation in the Context of Childhood Adversities*, ed. S. S. Luthar (Cambridge, MA: Cambridge University Press), 510–549.

Marlow, S. L., Salas, E., Landon, L. B., and Presnell, B. (2016). Eliciting teamwork with game attributes: a systematic review and research agenda. *Comput. Hum. Behav.* 55, 413–423. doi: 10.1016/j.chb.2015.09.028

Masten, A. S. (2001). Ordinary magic: resilience processes in development. *Am. Psychol.* 56, 227–238. doi: 10.1037/0003-066X.56.3.227

McCord, J., Harman, J. L., and Purl, J. (2019). Game-like personality testing: an emerging mode of personality assessment. *Pers. Individ. Diff.* 143, 95–102. doi: 10.1016/j.paid.2019.02.017

McCraty, R., and Atkinson, M. (2012). Resilience training program reduces physiological and psychological stress in police officers. *Glob. Adv. Health Med.* 1, 44–66. doi: 10.7453/gahmj.2012.1.5.013

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.* 23, 13–23. doi: 10.3102/0013189X023002013

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Mil. Med.* 178, 107–114. doi: 10.7205/MILMED-D-13-00213

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). *A Brief Introduction To Evidence-Centered Design*. Princeton, NJ: ETS.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., et al. (2015). "Psychometrics and game-based assessment," in *Technology and Testing: Improving Educational and Psychological Measurement*, ed. F. Drasgow (Milton Park: Routledge), 23–48.

Narayanasamy, V., Wong, K. W., Fung, C. C., and Rai, S. (2006). Distinguishing games and simulation games from simulators. *ACM Comput. Entertain.* 4, 1–18.

Niessen, C., and Jimmieson, N. L. (2016). Threat of resource loss: the role of self-regulation in adaptive task performance. *J. Appl. Psychol.* 101, 450–462. doi: 10.1037/apl0000049

Nikolaou, I., Georgiou, K., and Kotsasarlidou, V. (2019). Exploring the relationship of a gamified assessment with performance. *Span. J. Psychol.* 1:e6. doi: 10.1017/sjp.2019.5

OECD (2018). *The Future of Education and Skills: Education 2030*. Paris: OECD.

O'Leary, V. E., and Ickovics, J. R. (1995). Resilience and thriving in response to challenge: an opportunity for a paradigm shift in women's health. *Womens Health* 1, 121–142.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Front. Psychol.* 10:2970. doi: 10.3389/fpsyg.2019.02970

Pangallo, A., Zibarras, L., Lewis, R., and Flaxman, P. (2015). Resilience through the lens of interactionism: a systematic review. *Psychol. Assess.* 27, 1–20. doi: 10.1037/pas0000024

Prensky, M. (2001). Digital natives, digital immigrants. *Horizon* 9, 1–6. doi: 10.1108/10748120110424816

Pusey, M., Wong, K. W., and Rappa, N. A. (2020). Resilience interventions using interactive technology: a scoping review. *Interact. Learn. Environ.* [Epub ahead of print]. doi: 10.1080/10494820.2020.1772837

Raudys, S. J., and Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intellig.* 13, 252–264.

Richardson, G. E. (2002). The metatheory of resilience and resiliency. *J. Clin. Psychol.* 58, 307–321. doi: 10.1002/jclp.10020

Robertson, I. T., Cooper, C. L., Sarkar, M., and Curran, T. (2015). Resilience training in the workplace from 2003 to 2014: a systematic review. *J. Occup. Organ. Psychol.* 88, 533–562. doi: 10.1111/joop.12120

Rosse, J. G., Stecher, M. D., Miller, J. L., and Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *J. Appl. Psychol.* 83, 634–644. doi: 10.1037/0021-9010.83.4.634

Salas, E., Rosen, M. A., Held, J. D., and Weissmuller, J. J. (2009). Performance measurement in simulation-based training: a review and best practices. *Simul. Gam.* 40, 328–376. doi: 10.1177/1046878108326734

Salen, K., and Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals*. Cambridge, MA: MIT Press.

Sarkar, M., and Fletcher, D. (2013). How should we measure psychological resilience in sport performers? *Meas. Phys. Educ. Exerc. Sci.* 17, 264–280. doi: 10.1080/1091367X.2013.805141

Sauve, L., Renaud, L., Kaufman, D., and Marquis, J. S. (2007). Distinguishing between games and simulations: a systematic review. *Educ. Technol. Soc.* 10, 247–256.

Seery, M. D., and Quinton, W. J. (2016). "Understanding resilience: from negative life events to everyday stressors," in *Advances in Experimental Social Psychology: Vol. 54*, eds J. M. Olson and M. P. Zanna (Amsterdam: Elsevier), 181–245.

Shute, V. J. (2011). "Stealth assessment in computer-based games to support learning," in *Computer Games and Instruction*, eds S. Tobias and J. D. Fletcher (Charlotte: Information Age Publishers), 503–524.

Shute, V. J., Ventura, M., and Ke, F. (2015). The power of play: the effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Comput. Educ.* 80, 58–67. doi: 10.1016/j.compedu.2014.08.013

Shute, V. J., Wang, L., Greiff, S., Zhao, W., and Moore, G. (2013). Measuring problem solving skills via stealth assessment in an engaging video game. *Comput. Hum. Behav.* 63, 106–117. doi: 10.1016/j.chb.2016.05.047

Simmering, V. R., Ou, L., and Bolsinova, M. (2019). What technology can and cannot do to support assessment of non-cognitive skills. *Front. Psychol.* 10:2168. doi: 10.3389/fpsyg.2019.02168

Taub, M., Azevedo, R., Bradbury, A. E., and Mudrick, N. V. (2020). "9 Self-regulation and reflection during game-based learning," in *Handbook of Game-Based Learning*, eds J. L. Plass, R. E. Mayer, and B. D. Homer (Cambridge, MA: MIT Press).

Teng, Y., Brannick, M. T., and Borman, W. C. (2020). Capturing resilience in context: development and validation of a situational judgment test of resilience. *Hum. Perform.* [Epub ahead of print]. doi: 10.1080/08959285.2019.1709069

van Hooft, E. A. J., and Born, M. P. (2012). Intentional response distortion on personality tests: using eye-tracking to understand response processes when faking. *J. Appl. Psychol.* 97, 301–316. doi: 10.1037/a0025711

Ventura, M., and Shute, V. J. (2013). The validity of a game-based assessment of persistence. *Comput. Hum. Behav.* 29, 2568–2572. doi: 10.1016/j.chb.2013.06.033

Walker, F. R., Pfingst, K., Carnevali, L., Sgoifo, A., and Nalivaiko, E. (2017). In the search for integrative biomarker of resilience to psychological stress. *Neurosci. Biobehav. Rev.* 74(Pt B), 310–320.

Walker, F. R., Thomson, A., Pfingst, K., Vlemincx, E., Aidman, E., and Nalivaiko, E. (2019). Habituation of the electrodermal response – A biological correlate of resilience? *PLoS One* 14:e0210078. doi: 10.1371/journal.pone.0210078

Waugh, C. E., Thompson, R. J., and Gotlib, I. H. (2011). Flexible emotional responsiveness in trait resilience. *Emotion* 11, 1059–1067. doi: 10.1037/a0021786

Windle, G., Bennett, K. M., and Noyes, J. (2011). A methodological review of resilience measurement scales. *Health Qual. Life Outcomes* 9:8. doi: 10.1186/1477-7525-9-8

Winslow, B. D., Carroll, M. B., Martin, J. W., Surpris, G., and Chadderdon, G. L. (2015). Identification of resilient individuals and those at risk for performance deficits under stress. *Front. Neurosci.* 9:328. doi: 10.3389/fnins.2015.00328

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zickar, M. J., and Robie, C. (1999). Modeling faking good on personality items: an item-level analysis. *J. Appl. Psychol.* 84, 551–563. doi: 10.1037/0021-9010.84.4.551

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.