Assessment
2022, Vol. 29(4) 630–650

DOI: 10.1177/1073191120985612
journals.sagepub.com/home/asm


Construct and Predictive Validity of an Assessment Game to Measure Honesty–Humility

Ard J. Barends[1], Reinout E. de Vries[1], and Mark van Vugt[1]

Abstract

Research on commercial computer games has demonstrated that in-game behavior is related to the players' personality profiles. However, this potential has not yet been fully utilized for personality assessments. Hence, we developed an applied (i.e., serious) assessment game to assess the Honesty–Humility personality trait. In two studies, we demonstrate that this game adequately assesses Honesty–Humility. In Study 1 ($N = 116$), we demonstrate convergent validity of the assessment game with self-reported Honesty–Humility and divergent validity with the other HEXACO traits and cognitive ability. In Study 2 ($N = 287$), we replicate the findings from Study 1, and also demonstrate that the assessment game shows incremental validity—beyond self-reported personality—in the prediction of cheating for financial gain, but not of counterproductive work and unethical behaviors. The findings demonstrate that assessment games are promising tools for personality measurement in applied contexts.

Keywords

Honesty–Humility, personality, serious game, game-based assessment, assessment game, in-game assessment, applied gaming

Computer gaming is one of the most popular forms of entertainment. Various researchers have attempted to utilize the popularity of games to achieve other goals beyond entertainment by developing applied games (Kato & De Klerk, 2017).[1] Applied games have mainly been developed to foster behavior change, training, and other educational goals (e.g., Bauer et al., 2017; Connolly et al., 2012). Typically, the motivational appeal of games has been utilized to increase people's engagement with the learning material (Prensky, 2001; Starks, 2014).

In recent years, another form of applied gaming has become more prevalent: using games for assessment purposes (Ifenthalter et al., 2012). Most of these game-based assessments have been conducted for educational assessments such as measuring the progress and outcomes of educational goals set forth in the game (e.g., Kiili et al., 2015; Shute et al., 2010; see Ventura & Shute, 2013, for an exception). However, game-based assessments can also be used for noneducational assessments. More specifically, games can be used to measure individual differences in personality because personality is linked to in-game behaviors in various commercial computer games (Tekofsky et al., 2013; Worth & Book, 2014, 2015). Such game-based assessments of personality may be useful for applied purposes in research, clinical assessments (Myers et al., 2016), and personnel selection and assessment (Fetzer et al., 2017). For instance, personality assessment games can be relevant for individuals who intellectually understand games but are unable to provide accurate self-reports, for instance, because they lack self-insight (e.g., individuals with borderline personality disorder; Morey, 2014) or because they self-enhance on self-reports (e.g., individuals with Autism Spectrum Disorder; Schriber et al., 2014).

In the current contribution, we describe the development and validation of an assessment game, called *Building Docks*, to measure the Honesty–Humility trait from the HEXACO model of personality (Ashton & Lee, 2007). In the present project, we will investigate the validity of our assessment game by measuring its convergent validity with self-reported Honesty–Humility and divergent validity with the other five self-reported personality HEXACO traits and cognitive ability. Furthermore, we will also assess its predictive validity by relating it to various outcomes that

[1]Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Corresponding Author:
Ard J. Barends, Department of Experimental and Applied Psychology, Institute for Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam 1081 HV, Netherlands.
Email: a.j.barends@vu.nl

have—in previous studies—been found to be associated with self-reports of Honesty–Humility. Finally, we will also investigate the incremental validity of the assessment game above and beyond self-reported Honesty–Humility in predicting these outcomes. However, we first briefly describe our new conceptual framework of applied gaming to clarify terminology and background.

Conceptual Framework

Applied gaming (Fleming et al., 2017) covers the application of computer games and gamification to achieve goals beyond entertainment. Gamification is the application of one or more game-design principles in a non-game context (Deterding et al., 2011) and the end result is a tool that itself cannot be considered a game (cf. Richter et al., 2015). Comparatively, applied games are full-fledged games that attempt to achieve goals beyond entertainment (Kato & De Klerk, 2017; Klabbers, 2009). Therefore, the concept of applied gaming presupposes a continuum of 'gamefulness' from gamification (low gamefulness) to applied games. In this continuum, we argue it is possible to distinguish between applied games that focus more on the "applied" aspect (intermediate gamefulness) and those that focus more on the "game" aspect (high gamefulness). The primary difference is whether the applied game feels more like a gamified application or an actual game. Furthermore, we argue that the two broad primary goals of applied games are "education" and "assessment."

Education broadly subsumes applied games that attempt to teach players' understanding of topics such as physics or math (e.g., Kiili et al., 2015), but also games that serve as a rehabilitation training after brain injury (Van der Kuil et al., 2018), or games that try to change attitudes (e.g., DeSmet et al., 2018). All these applications aim to create change in the player, and we argue that education is the most appropriate label.

Assessment broadly subsumes applied games that attempt to gain insight into a particular construct that is not developed or trained in the game itself. For instance, applied games that attempt to screen people at risk for developing diseases such as Alzheimer's (Coughlan et al., 2019) and games that are specifically developed to assess individual differences such as intelligence or personality are all covered under the goal of assessment. Such assessments can be used for personnel selection but also for clinical diagnosis.

Distinguishing between education and assessment goals of applied games has several advantages. First, this helps clearly determine the specific goals of applied games (e.g., Bellotti et al., 2013). Second, this distinction may be a helpful to map game attributes to the specific goals of applied games (e.g., Landers, 2014). Third, such a distinction can help select appropriate game genres for specific goals of applied games (e.g., Fetzer et al., 2017).

Combining the goals of education and assessment with the level of "gamefulness" results in a 2×3 taxonomy (see Table 1). This taxonomy distinguishes between the primary goal and the level of gamefulness. For instance, an assessment game is more like an assessment than an in-game assessment. Specifically, an assessment game is often designed in the form of a linear game in order to apply classical test theory to the assessment. In contrast, an in-game assessment more often has emergent properties and is usually designed in the form of a nonlinear game (e.g., adaptive toward the skill of the player) and, therefore, in-game assessments need to apply item response theory or Bayesian network analysis (Shute et al., 2010). Furthermore, we propose to use the more generic term game-based assessments to broadly refer to all these assessments applications regardless of the level of gamefulness. Similarly, we reserve the generic term game-based education to refer to all education applications regardless of the level of gamefulness (see Table 1 for all definitions, potential advantages, disadvantages, and examples for each goal and level of gamefulness). Overall, this taxonomy may add more precision and clarity to the terminology currently used in the field, which suffers from a lack of standardization.

Honesty–Humility and the HEXACO Model of Personality

According to lexical personality studies, the maximum cross-culturally replicable structure of personality is most optimally represented by six dimensions, referred to as the HEXACO dimensions of personality (Ashton et al., 2014; De Raad et al., 2014; Saucier, 2009). HEXACO is an acronym of the six traits that it encompasses: Honesty–Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to Experience. The HEXACO model fully encompasses the historically older Five-Factor/Big Five Model (FFM; Digman, 1990) and although there are also notable differences in the rotational positions of Agreeableness and (HEXACO) Emotionality/(FFM) Neuroticism in the two models, the main difference is that the HEXACO model consists of the addition of Honesty–Humility (see Ashton et al., 2014; Ashton & Lee, 2007).

Honesty–Humility encompasses the degree that people are honest, sincere, lack feelings of entitlement, and are uninterested in status and luxury (Ashton et al., 2004, 2014).

In terms of predictive validity, Honesty–Humility has been related to various outcomes such as counterproductive work behavior (CWB; Zettler & Hilbig, 2010), unethical business decisions (UBD; Ashton & Lee, 2008; Lee et al., 2013), and cheating for financial gain (Hilbig & Zettler, 2015; Thielmann et al., 2017). The inclusion of Honesty–Humility in the HEXACO model also has several practical advantages to the FFM. For instance, two

Table 1. Conceptual Taxonomy and Labeling of Applied Games in Relation to Their Primary Goal and Level of Gamefulness Including Definitions, Advantages, and Disadvantages.

	Game-based assessments		
	Gamified assessment	Assessment game	In-game assessment
Definition	*An approach in which game elements are added to an existing assessment instrument*	*An approach in which existing assessment items are incorporated in a full-fledged game*	*An approach in which in-game behaviors and metrics from a game with emergent properties are used for assessment*
Goal	Diagnosis, Selection, Assessment, giving diagnostic feedback	Diagnosis, Selection, Assessment, giving diagnostic feedback	(Diagnosis), Selection, Assessment, giving diagnostic feedback
Gamefulness	Low	Intermediate	High
Structure	Linear/Adaptive	Linear/Adaptive	Non linear
Statistical implications	Classical test theory/Item response theory	Classical test theory/Item response theory	Bayesian Network/Item response theory
+	Retains psychometric properties	Retains psychometric properties	High engagement
−	Relatively low engagement	Repetitive with repeated play Costly development	Difficult to incorporate feedback to players Cognitive overload Costly development
Example	Myers et al. (2016)	Current contribution	Ventura and Shute (2013)
	Game-based education		
	Gamified education	Educational game	In-game education
Definition	*An approach in which game elements are added to an existing educational approach*	*An approach in which existing education items are incorporated in a full-fledged game*	*An approach in which in-game behaviors and metrics from a game with emergent properties are used for education*
Goal	Training, Education, and giving developmental feedback	Training, Education, and giving developmental feedback	Training, Education, and giving developmental feedback
Gamefulness	Low	Intermediate	High
Structure	Linear/Adaptive	Linear/Adaptive	Nonlinear
+	Relatively inexpensive to develop Easy to incorporate feedback to players	Option to tailor content to current level of learning Easy to incorporate feedback to players	High engagement
−	Relatively low engagement	Repetitive with repeated play Costly development	Difficult to incorporate feedback to players Cognitive overload Costly development
Example	Barata et al. (2017)	Kiili et al. (2015) [Semideus game]	Liu et al. (2016)

Note. Diagnosis is added in brackets for in-game assessments as this seems difficult to create diagnostic cutoff scores on metrics derived from games with emergent properties.

recent meta-analyses found that the HEXACO model outperformed the FFM in the prediction of CWB (Pletzer et al., 2019; see also Pletzer et al., 2020) and prosocial behavior (Thielmann, Spadaro, et al., 2020). In both meta-analyses, the superior predictive value of the HEXACO model was due to its inclusion of Honesty–Humility.

Economic Games and Personality

Personality has been related to in-game behaviors of so-called economic games and social dilemmas (we will describe them jointly as economic games in the remainder of this manuscript). These economic games are abstract, text-based games that are used to investigate topics such as strategic and prosocial behaviors (see Fehr & Schmidt, 2006; Pruitt & Kimmel, 1977, for reviews). Although there are many different economic games, all of them involve decision-making situations in which someone can make a self-interested choice at the cost of the welfare of others or forgo the self-interested choice to foster the welfare of the collective (Thielmann et al., 2015).

In economic games, Honesty–Humility is mainly relevant for the degree that someone actively cooperates with others (Thielmann, Spadaro, et al., 2020).[2] More specifically,

Honesty–Humility has been related to cooperative behavior in various games such as the dictator game (Barends et al., 2019b; Hilbig & Zettler, 2009; Zhao et al., 2017), the public goods game (Hilbig et al., 2012), and the prisoner's dilemma (Zettler et al., 2013). Because of their relation with Honesty–Humility, adaptations of such economic may be useful additions for the current assessment game.

Personality and Behavior in Commercial Games

Personality has also been related to behavior in various commercial computer games. For instance, Conscientiousness is negatively related to speed of play in a first-person shooter game (Tekofsky et al., 2013). Similarly, in the massively multiplayer online role-playing game World of Warcraft, all six HEXACO traits have been related to theoretically plausible in-game behaviors (Worth & Book, 2014). For instance, Conscientiousness was positively related to how frequent people engaged in in-game working (e.g., collecting resources and crafting items). Agreeableness was positively related to the frequency of helping behavior directed at other players (e.g., healing other players). Furthermore, other studies found meaningful relations between self-reported in-game preferences and behaviors on the one hand and personality traits on the other (Tabacchi et al., 2017; Worth & Book, 2015; Zeigler-Hill & Monica, 2015; cf. McCreery et al., 2012).

The Situation-Trait-Outcome Activation (De Vries et al., 2016) model may explain the relations between personality and different behaviors observed in computer games. The Situation-Trait-Outcome Activation model posits that personality is expressed in three different personality-situation interactions. We will illustrate each of these interactions with the findings of Worth and Book (2014) in their study on the HEXACO traits in World of Warcraft. First, situation activation means that people seek out specific situations that fit their personality (e.g., by selecting, perceiving, evoking, and/or manipulating situations to fit one's personality profile). For instance, people high in Openness to Experience more frequently explored the game world, similarly, individuals low in Honesty–Humility more frequently sought out player-versus-player activities. Second, specific situations activate the expression of specific personality traits (i.e., trait activation). For instance, people high in Openness to Experience were more likely to make unusual in-game items whereas people low in Honesty–Humility more frequently attempted to ruin other players' experience by stealing kills of other players (Worth & Book, 2014). Third, the expression of personality may be differentially related to outcomes such as rewards and punishments (i.e., outcome activation). This latter aspect was not studied by Worth and Book, but we would expect that people low in Honesty–Humility are more likely to gain high leaderboard scores and people high in Openness to Experience to gain exploration achievements.

However, commercial games are developed for the purpose of entertainment, making them less suitable for the in-game personality assessments. For instance, it is often impossible to access internal logging data of commercial games, many games require hours of play to master them, and a lot of in-game behavior is irrelevant for the inference of personality (see, e.g., Tekofsky et al., 2013). Consequently, utilizing such commercial games for personality assessment is likely to be a waste of assessment time. Furthermore, psychometric considerations are unlikely to have played a role in the development of commercial games. Therefore, assessment games are—ceteris paribus—much better equipped than commercial games for such practical purposes.

Prior Research on Assessment Games and Gamified Assessments of Personality

To date, to our knowledge, one in-game assessment and several gamified assessment applications have been developed to measure personality traits. Gamified assessment tools that assess personality usually do not incorporate actual game mechanics but use a storyline, avatars, and visual imagery to give a feeling of "gamefulness" to more traditional assessment tasks (e.g., questionnaires, situational judgment tests [SJT]; Georgiou et al., 2019; Levy et al., 2016; McCord et al., 2019; Myers et al., 2016; cf. Barends et al., 2019a).

Gamified assessment tools that utilized a so-called SJT format have generally found encouraging results in terms of construct validity. In a SJT, participants are confronted with a description of a particular situation, and participants choose one out of several response options. The contents of these SJTs can be presented in a text-based format (e.g., Oostrom et al., 2019), a video-based format (e.g., Dubbelt et al., 2015), or in a gamified format (e.g., Georgiou et al., 2019). To illustrate these studies' findings, McCord et al. (2019) found convergent validity between a gamified assessment of the FFM personality traits and several self-reported FFM personality traits. However, not all FFM personality traits assessed using the gamified assessment tool were significantly correlated with their corresponding trait. Furthermore, this study also found that some of these traits had considerable correlations with unintended personality traits (i.e., had low divergent validity). Similarly, in a gamified SJT, participants selected an avatar and completed a set of SJTs embedded within a virtual world with an overarching storyline (Georgiou et al., 2019). This gamified SJT had convergent validity with the four assessed skills (e.g., resilience) and divergent validity. Furthermore, this gamified SJT was also able to predict self-reported work and

academic performance (Nikolaou et al., 2019). Overall, these studies demonstrate that gamified assessments using SJTs can validly measure personality.

In addition to the gamification of SJT assessments, so-called virtual behavior cues (or virtual cues for short) can also be used as gamified assessment tools (Barends et al., 2019a). These virtual cues are visual customizations in a virtual environment. They can be made in the creation of avatars, the customization of a virtual car, or the decoration of a virtual office. Prior work has found that customization of avatars has been related to FFM personality (Bélise & Bodur, 2010; Fong & Mar, 2015). Barends et al. (2019a) developed a scale based on a variety of these virtual cues and showed that this scale had acceptable reliability, convergent validity with self-reported Honesty–Humility, and divergent validity with the other five HEXACO traits.

Finally, there is some evidence that in-game assessments can be used to measure particular personality traits (Ventura & Shute, 2013). Specifically, Ventura and Shute developed and validated an in-game assessment to measure the persistence facet of Conscientiousness. Their in-game assessment of persistence was significantly correlated with a behavioral assessment of persistence; however, their in-game assessment did not show any convergent validity with self-reported persistence.

The above studies suggest that it is possible to develop a game-based assessment to measure personality traits, especially if the assessment game is based on traditional assessment tasks (e.g., Georgiou et al., 2019; Myers et al., 2016). However, there are various potential challenges. The first challenge is that little is known about the psychometric properties of such game-based assessments (e.g., internal reliability). The second challenge is that it seems difficult to simultaneously achieve convergent and divergent validity. For instance, the gamified assessments described above often also measured unintended personality traits (Georgiou et al., 2019; McCord et al., 2019). Another threat to the construct validity game-based assessments of personality is that it may inadvertently (also) measure cognitive ability because games are immersive and engaging and can result in a high cognitive load for players (Gundry & Deterding, 2018). Consequently, managing such a cognitive load of game-based assessments may tap into cognitive ability more than in the targeted personality trait. The third challenge is that it is an outstanding question whether these game-based assessments of personality can predict relevant outcomes (i.e., have predictive validity) and if they do, whether they are able to predict these criteria above and beyond traditional self-report personality assessments (i.e., incremental validity).

Present Research

The overarching goal of the current set of studies was to investigate the potential utility of assessment games for personality assessment. For this purpose, a new personality assessment game was constructed, called *Building Docks*. With this assessment game, our first goal was to investigate the construct validity of *Building Docks*. We expected that the behaviors in *Building Docks* would be positively related to self-reported Honesty–Humility. We also expected that the game would show divergent validity in terms of the absence of correlations with the other five HEXACO traits and cognitive ability. Furthermore, our second goal was to investigate the predictive validity of *Building Docks* in predicting Honesty–Humility related outcomes (cheating, CWB, and UBD). Moreover, we also investigated whether *Building Docks* had incremental validity above and beyond self-reported Honesty–Humility in predicting these outcomes. The set of studies adds to the literature by demonstrating the utility of a personality assessment game and how this game can predict relevant outcomes.

Design of the Assessment Game "Building Docks"

The assessment game *Building Docks* was designed based on prior assessment methods that have shown trait activation of Honesty–Humility. Specifically, we based Building Docks on SJTs (De Meijer et al., 2010; Oostrom et al., 2019), virtual cues (Barends et al., 2019a), and economic games (e.g., Barends et al., 2019b; Hilbig et al., 2012; Zhao et al., 2017; see Zhao & Smillie, 2015, for a review).

In *Building Docks* these three types of tasks (SJTs, virtual cues, and economic games) are integrated into a full-fledged assessment game. In *Building Docks*, the player has to develop a harbor together with three computer-controlled characters. The economic games are used as the main game mechanic. Specifically, all player decisions in these economic games are consequential in the game and determine the way that the harbor develops. The virtual cues are used as rewards to further customize the harbor. Finally, the SJTs are used as an independent story line in the harbor (further details of the game can be found in the materials of Study 1). Compared with a gamified assessment tasks, *Building Docks* includes an in-game goal and consequential game mechanics for a player to achieve this in-game goal. *Building Docks* is a linear assessment game and is divided into four "quarters" of a year in which the player starts out as an employee in the first and second quarters, becomes a senior manager in the third quarter, and CEO in the fourth quarter.[3]

Study 1

Method

Below, we report how we determined our sample size, all data exclusions, all manipulations, and all measures in this study.

Participants and Procedure. Recent Dutch graduates who participated in a competition for match-making with various internationally operating companies were invited to complete *Building Docks* and the HEXACO-100 (Lee & Ashton, 2018) as additional assessments. The winner of the competition earned a cash prize of €10,000 (roughly $11,000). The graduates could also voluntarily complete a cognitive ability test as part of the competition. The invitation and reminders sent to the graduates emphasized that completing *Building Docks* and the HEXACO-100 was voluntary and had no effect on the competition. Furthermore, participants could receive a personal HEXACO-100 personality report in return for their participation. This personality report included the background of the instrument and definitions of the HEXACO traits, the raw scores and how the scores should be interpreted. We invited 1700 potential respondents, and in total, 116 people (M_{age} = 23.48 years; SD_{age} = 2.06 years; 56.9% men) completed the HEXACO-100 and *Building Docks*. Seventy-five of these participants also completed the cognitive ability test. Participants first completed the cognitive ability test, then the HEXACO-100, and finally, *Building Docks*. An apriori power analysis was conducted using G*Power 3.1.9.2 (Faul et al., 2007) to inform our minimally required sample size. We expected an effect size of r = .30 based on prior studies investigating convergent validity between self-reported Honesty–Humility and the other types of assessments of this trait (e.g., Barends et al., 2019a; Oostrom et al., 2019). This power analysis indicated that we required a sample of 82 participants to detect an effect of r = .30 (two-tailed test, 80% power, α = .05).

Materials

Cognitive Ability. The cognitive ability test was developed by LTP business psychologists and consisted of an abstract reasoning task, a verbal intelligence task, and a numerical intelligence task. This cognitive ability test has demonstrated convergent validity with another published cognitive ability test: the Multicultural Capacity Test (Bleichrodt & Van den Berg, 1999; Kappe & Van Der Flier, 2012). The data made available by LTP showed that in high-stakes assessment samples, these two cognitive ability tests were significantly and positively correlated (rs = .33 to .60 for the various subtests). In the original studies, each subtest had a Cronbach alpha reliability (α) of at least .73, and the overall cognitive ability test had an α of .91. However, we were not provided with the data at the item level and therefore could not calculate αs in the current study.[4]

HEXACO-100. The six HEXACO traits were measured with the Dutch HEXACO-100 (De Vries et al., 2009). The Dutch version is equivalent with the HEXACO-100 across other languages (Thielmann, Akrami, et al., 2020) in terms of factor structure and item loadings. Each trait was measured with 16 items each. This questionnaire also included four items to measure the interstitial Altruism facet. This Altruism theoretically covers the space between Honesty–Humility, Emotionality, and Agreeableness. Responses were self-reported on a 5-point Likert-type scale (1 = *strongly disagree* and 5 = *strongly agree*). All responses were checked for noncompliant responses using the procedure of Lee and Ashton (2018; see also Barends & De Vries, 2019), however, none of the responses were noncompliant. All six-factor scales in the current study had at least an α of .76.

Figure 1. A screenshot from the assessment game "Building Docks."

Building Docks. Participants completed *Building Docks* in Dutch or in English.[5] *Building Docks* (see, e.g., Figure 1) was described as a task instead of a game to participants because prior work on economic games has suggested that this terminology may affect in-game decisions (Zhao et al., 2017). Participants were instructed to develop a harbor with three other (computer-controlled) dock-owners. These three dock-owners were introduced in a brief description and were represented by avatars (two were male, and one was female). The participants were instructed that every dock-owner had to reach a break-even point of $50,000, if they would not reach this point, they would have failed the task. In reality, all dock-owners always reached this break-even point. The break-even point was reached roughly in the middle of the game; the exact moment was based on the in-game decisions of the participant. After the break-even point was reached, the game continued, and participants were free to choose their own strategy and goal (as in sandbox games; e.g., Rollercoaster Tycoon). The initial collective goal was chosen because our prototype tests indicated that without receiving any instructions, people assumed their goal to be competitive.[6] The in-game money was earned by playing economic game scenarios (see, e.g., Figure 2). Participants could customize their harbor using

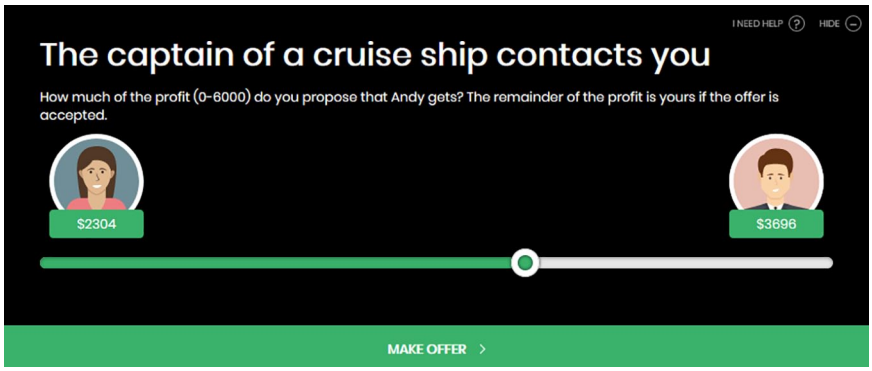

Figure 2. An example of an economic game scenario used in "Building Docks."

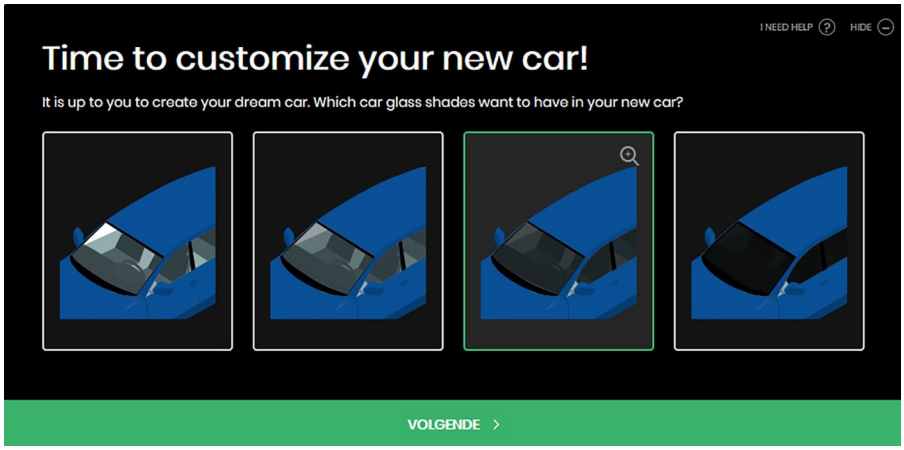

Figure 3. An example of virtual cue used in "Building Docks."

virtual cues (Barends et al., 2019a; see, e.g., Figure 3). The SJTs followed an overarching storyline in which the participant had to create a bid book (i.e., a proposal to realize a project such as the Olympic games) to get to host an event in the harbor. In the separate SJT items the participant had to decide how they approached this task and how they dealt with various events and setbacks (see, e.g., Figure 4). Although the *Building Docks* tasks were completed in a fixed order, to increase the feeling of playing a game, the three types of tasks were alternated at irregular intervals.

Building Docks economic games. *Building Docks* consisted of 12 economic game scenarios based on standard economic games and social dilemmas (e.g., dictator game; prisoner's dilemma; public goods game; Van Lange et al., 2014). However, the economic game scenarios were rewritten to fit into the narrative of *Building Docks*, to be more engaging, to decrease the level of abstraction of these tasks, and to increase their face validity. Thus, to increase the appeal of the items, in the scenarios, we refrained from using the matrix format to present information. Additionally, in two-person games (e.g., dictator game), the participant always played with one of the computer-controlled dock-owners. In the social dilemmas (e.g., public goods game), the player always played with all three computer-controlled dock-owners. The behaviors of the computer-controlled dock-owners were preprogrammed to standardize the game as much as possible and minimize dependence between scenarios. Eleven of the 12 economic game scenarios were developed to measure Honesty–Humility, and the other one served as a filler task. Additionally, two of the public goods games used a cheap talk paradigm (i.e., gave participants the opportunity to send a deceptive message). These tasks required players to indicate what they would do and what they actually did. We used both the difference between what they did and what they told others they would do and the actual behavior as outcomes. This resulted in a total of 13 economic game assessments of Honesty–Humility.

Building Docks situational judgment tests. A series of 12 SJTs followed a storyline that the participant had to create a bid book to get to host a Fleet parade in the harbor. These

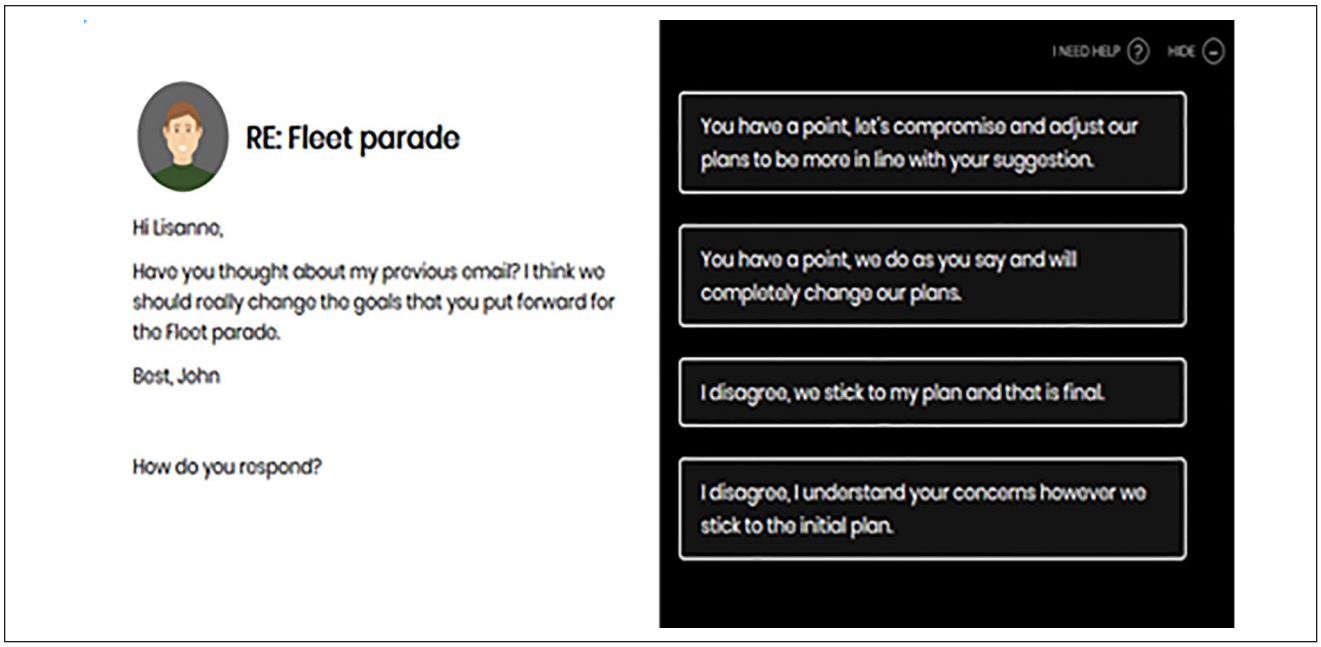

Figure 4. An example of a Situational Judgment Test item used in "Building Docks."

SJTs were developed using a construct-driven approach (Lievens, 2017; Oostrom et al., 2019). Eight of these SJTs were developed to measure Honesty–Humility, additionally, four SJTs were fillers (developed to measure some of the other HEXACO traits).[7] Respondents selected one of the four response options in each SJT. Each response option was developed to express a different level of the trait and were scored on the rank order of this expression (1 = *lowest level of Honesty–Humility* and 4 = *highest level of Honesty–Humility*). Per SJT, the order of the response options was randomized. Customized feedback was given based on the selected response (e.g., giving a different reason why someone disagreed with the decision), however, this feedback had no impact on the overall storyline (cf. Kanning et al., 2006). We decided not to branch the storyline to keep the assessment game as standardized as possible.

Building Docks virtual cues. Participants also completed 27 virtual cues (Barends et al., 2019a). New virtual cues were developed based on the original versions to fit into the visual style of *Building Docks*. Six virtual cues were used to create the player avatar, and five virtual cues were placed in the harbor environment throughout the game. The remaining 16 virtual cues did not have a visual reference point in the game after their selection (e.g., players selected and customized a car which was not placed in the harbor environment). During the selection, participants could zoom in to see all the pictures of the virtual cues, and participants had to choose one out of four alternatives per virtual cue. Each of the four alternatives was developed to be indicative of a different standing on the relevant personality trait. Answers were scored using the same rank ordering as used for the SJTs. Eighteen virtual cues were developed to measure Honesty–Humility, and the rest were fillers (all but two were designed to measure the other HEXACO traits).

All *Building Docks* responses were scored using the product file (i.e., the final in-game choices; De Klerk & Kato, 2017). We did not investigate any of the intermediate steps (e.g., process files, mouse clicks, and response times). To calculate *Building Docks* scores, all the scores per item were converted to z-scores. These z-scores were aggregated for every subtask and for the overall score. Note that some participants had missing data on some of their *Building Docks* items because responses were not saved if players lost their internet connection when an answer was transferred to the database. This happened for a total of eight items (i.e., 0.13% of the *Building Docks* data-points). The scores of participants who had missing Building Docks items were computed based on the available data.[8]

Results and Discussion

First, the *Building Docks* subtasks and the overall Honesty–Humility score had Cronbach αs between .45 and .78. Furthermore, Table 2 shows that all the *Building Docks* subtasks were all significantly and positively correlated (rs = .22 to .33, $p < .05$).

Subsequently, convergent validity was investigated with a correlational analysis. Table 2 shows that the overall *Building Docks* Honesty–Humility score was significantly and positively correlated to self-reported Honesty–Humility ($r = .33$, $p < .001$). At the subtask level, all correlations

Table 2. Descriptive Statistics of and Correlations Between the Demographic Variables, Personality, Intelligence, and Building Docks Variables in Study 1.

	M	*SD*	α	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Gender	0.57	0.50	—	—															
2 Age	23.48	2.06	—	−.06	—														
3 H	3.67	0.44	.76	−.17	.01	—													
4 E	2.94	0.54	.82	−.49**	−.15	.03	—												
5 X	3.63	0.61	.90	−.22*	.20*	−.06	−.04	—											
6 A	3.31	0.49	.80	−.08	.05	.08	.09	.01	—										
7 C	3.71	0.52	.85	−.16	−.03	.14	.07	.12	−.10	—									
8 O	3.43	0.64	.85	−.04	.31**	−.01	−.00	.30**	.07	.02	—								
9 Altruism	3.84	0.59	.57	−.28**	.11	.44**	.18	.20*	.24*	.10	.21*	—							
10 Abstract intelligence	4.96	1.04	—	.13	−.13	−.17	−.04	−.20	.17	.23	.02	−.06	—						
11 Numerical intelligence	4.97	1.10	—	.30**	−.21	−.19	−.26*	.13	−.13	.26*	−.18	−.14	.37**	—					
12 Verbal intelligence	3.98	1.13	—	.22	−.01	−.13	−.08	.05	.02	.08	.03	.07	.53**	.37**	—				
13 Overall intelligence	4.79	1.07	—	.25*	−.14	−.19	−.14	−.04	.05	.23	−.04	−.05	.85**	.67**	.82**	—			
14 BD: EG	−0.00	0.42	.62	.23*	.03	.10	−.10	−.19*	−.01	.02	.08	−.02	.10	.15	.19	.18	—		
15 BD: SJT	0.00	0.45	.45	.07	.05	.17	.02	.04	.07	−.02	−.03	.03	−.16	−.10	−.07	−.14	.22*	—	
16 BD: VC	0.00	0.45	.76	−.09	.18	.43**	−.04	−.02	.16	.21*	.14	.22*	.11	.04	.05	.09	.29**	.33**	—
17 BD: Total score	0.00	0.32	.78	.09	.12	.33**	−.06	−.08	.10	.10	.09	.11	.02	.04	.07	.06	.68**	.73**	.75**

Note. Gender is coded as F = 0, M = 1. BD = *Building Docks*; EG = economic games; SJT = situational judgment tests; VC = virtual cues.
*$p < .05$. **$p < .01$.

with self-reported Honesty–Humility were positive, however, only the virtual cues of Honesty–Humility score did reach conventional levels of significance ($r = .43$, $p < .001$), whereas the economic game Honesty–Humility score ($r = .10$, $p = .270$) and SJT Honesty–Humility score ($r = .17$, $p = .066$) did not reach this level of significance. The significant positive relation of the overall game score with self-reported Honesty–Humility supports the claim that *Building Docks* validly measures Honesty–Humility.

Second, to assess divergent validity, the *Building Docks* Honesty–Humility score was correlated with the other five HEXACO traits and the four intelligence scores. Table 2 shows that the overall *Building Docks* Honesty–Humility score did not significantly correlate with any of these variables (all $p > .25$). Only at the subtask level of *Building Docks*, two significant correlations were observed. Specifically, the *Building Docks* Honesty–Humility score obtained from economic games and self-reported Extraversion were significantly negatively correlated ($r = -.19$, $p = .037$), similarly, the *Building Docks* Honesty–Humility score obtained from virtual cues was significantly correlated to self-reported Conscientiousness ($r = .21$, $p = .027$). This proportion of significant correlations (2 out of 27 correlations or 7.5% of the correlations) is close to the rate of expected false positives based on the α level of .05 for significance testing.

Study 2

Study 1 provided initial evidence that we can use an assessment game to measure Honesty–Humility. In Study 2, we wanted to replicate whether *Building Docks* was able to assess this trait. Furthermore, an additional goal of Study 2 was to investigate the predictive and incremental validity of *Building Docks* in predicting three Honest–Humility-related outcomes, namely, the probability of cheating for financial gain, CBW, and UBD. We further improved on our Study 1 design by recruiting a larger sample, using a longer (and therefore more reliable) version of the HEXACO inventory (i.e., the HEXACO-208; De Vries et al., 2015) which allowed us to investigate the validity at both the factor and facet level. Furthermore, we improved the design by counterbalancing the order of administration of the assessment game and the HEXACO-208. To recruit a large sample, we conducted our study on Amazon Mechanical Turk (MTurk) as data on this platform tends of similar quality as high-quality commercial samples (McCredie & Morey, 2019; Thomas & Clifford, 2017). However, prior studies have indicated that a significant proportion of these MTurk respondents give noncompliant responses (e.g., Barends & De Vries, 2019). Therefore, we decided to also include several data quality checks in the HEXACO inventory.

Method

Below, we report how we determined our sample size, all data exclusions, all manipulations, and all measures in this study.

Participants and Procedure. Using MTurk, 500 American participants were recruited for a three-phase study. Only participants who had completed more than 5000 prior MTurk human intelligence tasks and were granted payment in at least 95% of them were eligible to participate. Each phase was completed one week after the other, and only participants who completed the previous phase could enter the subsequent phase (e.g., people who did not participate in Phase 2 were blocked from participating in Phase 3). Half of the participants completed the HEXACO inventory in Phase 1 and *Building Docks* in Phase 2, the other half of the participants completed *Building Docks* in Phase 1, and the HEXACO inventory in Phase 2. Individuals who gave noncompliant responses on the HEXACO inventory (either in Phases 1 or 2 depending on the condition) were blocked from further participation and their data was also not included in the analyses. In Phase 3, participants completed the three dependent variables (i.e., UBD, CWB, and a cheating task).

We recruited a larger sample compared with Study 1 because of potential dropouts between the different phases and the prevalence of noncompliant responses in such online samples (see Barends & De Vries, 2019). We expected an effect of $r = .20$ for the predictive validity of *Building Docks* as this is the typical effect in personality research (Gignac & Szodorai, 2016). Our a priori power analysis using G*Power 3.9.1.2 (Faul et al., 2007) indicated that we required a sample of at least 191 participants to detect this effect (two-tailed test, 80% power, $\alpha = .05$). Data of participants, who were not flagged as giving noncompliant responses and who had completed the first and second phases, were analyzed for the construct validity of *Building Docks* ($n = 287$; $M_{age} = 39.85$ years; $SD_{age} = 11.59$ years; 41.1% men)[9], similarly, data of participants who completed all three phases ($n = 241$; $M_{age} = 40.02$ years; $SD_{age} = 11.35$ years; 39.4% men; 91.2% of the respondents were currently employed) were analyzed for the predictive and incremental validity of *Building Docks*.

Materials

HEXACO-208. The six HEXACO traits were measured with the HEXACO-208 inventory (De Vries et al., 2015). The HEXACO-208 is an adapted version of the full-length HEXACO-PI-R (Lee & Ashton, 2006) and measures the six broad traits with 32 questions each and the interstitial Altruism and Proactivity facets with eight items each. Responses

were self-reported on a five-point Likert-type scale (1 = *strongly disagree* and 5 = *strongly agree*). In the current study, αs of all six-factor scales were at least .90 and of the 26 facets at least .75.

Data Quality Checks. Four instructed response items (e.g., *this is an attention check; please select "strongly agree"*) and four infrequency scale items (e.g., *I never bought anything in a store*; Fekken et al., 1987) were embedded within the HEXACO-208 to detect noncompliant responses. Additionally, the criteria of Lee and Ashton (2018) were used to filter out noncompliant responses based on a subset of the items (the HEXACO-100 items). See Barends and De Vries (2019) for further details.

Building Docks. We only used the English version of the assessment game reported in Study 1. The α of the overall game Honesty–Humility score was .73 and of the game's Honesty–Humility subtasks between .37 and .75.

Dependent Variables. The three dependent variables were collected in Phase 3 of the project. The three tasks were completed in a randomized order for every participant.

Counterproductive work behavior. CWB was measured with the 19-item self-report scale developed by Bennett and Robinson (2000). Participants indicated on a 7-point Likert-type scale how frequently they had engaged in various deviant behaviors at their work in the past year (1 = *never* and 7 = *daily*). Reliability in the current study was $\alpha = .92$.

Unethical business decisions. To measure UBD, six scenarios of Ashton and Lee (2008) were used. Respondents read a scenario about a potential UBD and indicated on a four-point Likert-type scale how likely they were to engage in the described activity (1 = *definitely not* and 4 = *definitely yes*). The responses of the six scenarios were aggregated, and the reliability was $\alpha = .78$.

Cheating. Participants had to flip a coin twice and received a $0.50 bonus if they reported two successive heads. If they reported any other results, they did not receive this bonus. Note that such a cheating task does not allow us to determine actual cheating but only the probability of cheating because a proportion of the participants will legitimately report two successive heads. The 25% probability of winning is considered an adequate tradeoff between observing a sufficient number of cheaters without inducing fear of incriminating oneself as a cheater (Moshagen & Hilbig, 2017). Responses were analyzed with the R script of Moshagen and Hilbig that corrects for the proportion of legitimate wins in the sample.

Results

Replicating Study 1. First, we investigated whether the convergent validity between self-reported Honesty–Humility and in-game behavior of Study 1 could be replicated. Table 3 shows that the different Honesty–Humility subtasks of *Building Docks* were all positively and significantly related (rs = .13 to .19, $p < .05$). Furthermore, the overall *Building Docks* Honesty–Humility score was significantly related to self-reported Honesty–Humility ($r = .28$, $p < .001$). Additionally, all *Building Docks* Honesty–Humility subtasks were significantly related to self-reported Honesty–Humility, (rs = .14 to .26, $p < .05$). Finally, all four Honesty–Humility facets were significantly related to the overall *Building Docks* Honesty–Humility score (Sincerity: $r = .21$, $p < .001$; Fairness: $r = .19$, $p = .001$; Greed-avoidance: $r = .26$, $p < .001$; Modesty: $r = .24$, $p < .001$; see Table S2, available in the online supplementary material, for further details).

Second, we also tested whether the divergent validity of *Building Docks* and the other self-reported HEXACO traits could be replicated. As Table 3 shows, the overall *Building Docks* Honesty–Humility score was significantly related to Agreeableness ($r = .12$, $p = .036$). None of the other HEXACO scales were significantly correlated to the overall *Building Docks* Honesty–Humility score. With respect to the *Building Docks* Honesty–Humility subtasks, the *Building Docks* SJT Honesty–Humility score was positively related to Emotionality ($r = .14$, $p = .014$), and also negatively related to Extraversion ($r = -.14$, $p = .022$). Moreover, the *Building Docks* virtual cues Honesty–Humility score was positively related to Agreeableness ($r = .13$, $p = .026$). This proportion is somewhat higher than the expected number of false positives based on the alpha level of .05 (about 20% of the correlations were significant at the scale level). Finally, at the HEXACO facet level, 15 significant divergent validity correlations between the facets of the non-Honesty–Humility HEXACO scales and the *Building Docks* overall Honesty–Humility score and Honesty–Humility subtasks were observed. This number was also somewhat higher than the expected number of false positives (e.g., 20% of the investigated correlations were significant; see the online supplementary Table S1). Generally, the patterns of divergent validity were replicated.

Predictive and Incremental Validity of Building Docks. The negative correlations reported in Table 3 also demonstrate that the *Building Docks* Honesty–Humility score had predictive validity for two outcomes. Specifically, people who had a higher score on *Building Docks* Honesty–Humility had a lower probability of cheating in the cheating task than individuals with a lower score ($r = -.19$, $p = .004$).

Table 3. Descriptive Statistics of and Correlations Between the Demographic Variables, Personality, and the Outcome Variables, and Building Docks Variables in Study 2.

	M	*SD*	α	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Gender	.41	.49	—	—															
2 Age	39.85	11.59	—	−.17**	—														
3 H	3.70	.68	.94	−.23**	.27**	—													
4 E	3.18	.60	.91	−.39**	.09	.12	—												
5 X	3.23	.74	.95	.07	.12*	−.09	−.24**	—											
6 A	3.16	.59	.92	.05	.13*	.37**	−.02	.33**	—										
7 C	3.75	.55	.91	−.04	.06	.32**	−.06	.30**	.17**	—									
8 O	3.41	.58	.90	.07	.08	.13*	−.12*	.12*	.17**	.16**	—								
9 BD: EG	.00	.46	.68	.05	.09	.15**	−.04	.02	.09	−.06	−.01	—							
10 BD: SJT	−.00	.43	.37	−.20**	.13*	.26**	.14*	−.14*	.03	.08	−.08	.13*	—						
11 BD: VC	.00	.45	.75	−.02	.09	.14*	.00	−.09	.13*	−.03	.01	.14*	.19**	—					
12 BD: total	.00	.29	.73	−.08	.16**	.28**	.05	.11	.12*	−.01	−.04	.65**	.65**	.68**	—				
13 UBD	1.96	.64	.78	.22**	−.31**	−.46**	−.26**	.10	−.26**	−.10	−.26**	−.09	−.19**	−.13	−.21**	—			
14 CWB: interp.	1.47	.85	.90	.15*	−.13*	−.34**	−.06	−.07	−.24**	−.34**	−.17*	−.06	−.16*	−.08	−.15*	.31**	—		
15 CWB org.	1.76	.85	.87	.04	−.18**	−.35**	.05	−.31**	−.27**	−.47**	−.13*	−.05	−.09	−.03	−.08	.21**	.69**	—	
16 CWB total	1.65	.79	.92	.09	−.17**	−.38**	.01	−.24**	−.28**	−.46*	−.15*	−.06	−.13	−.05	−.12	.27**	.87**	.96**	—
17 Cheating (nr of reported wins)	.66	.47	—	.03	−.12	−.12	−.04	.10	−.03	.08	.09	−.18**	−.12	−.08	−.19**	.18**	.05	.00	.02

Note. Gender is coded as F = 0, M = 1. CWB = counterproductive work behavior; UBD = unethical business decisions; SJT = situational judgment tests; BD = *Building Docks*, EG = economic games, VC = virtual cues.
*$p < .05$. **$p < .01$.

Additionally, people with a higher *Building Docks* Honesty–Humility score were less likely to engage in UBD than individuals with a lower score ($r = -.21$, $p = .001$). The *Building Docks* Honesty–Humility score did not show predictive validity for overall CWB ($r = -.12$, $p = .070$); it did, however, have predictive validity for the interpersonal deviance subscale ($r = -.15$, $p = .020$) but not for the organizational deviance subscale ($r = -.08$, $p = .193$).[10]

To assess the incremental validity of the *Building Docks* Honesty–Humility score above and beyond self-reported personality (including Honesty–Humility), several multiple hierarchical regressions were conducted. In Step 1, the demographic variables and the five self-reported HEXACO traits not of focal interest were entered. In Step 2a, the *Building Docks* Honesty–Humility assessment was entered in the model. This model allowed us to investigate the incremental validity of the *Building Docks* Honesty–Humility assessment above and beyond these control variables. In Step 2b, the model from Step 1 was used, and the self-reported Honesty–Humility score was entered into the model. Subsequently, in Step 3, the *Building Docks* Honesty–Humility assessment was included in the model. The results of Step 3 allowed us to investigate the incremental validity of the *Building Docks* Honesty–Humility score above and beyond the demographics and the HEXACO-208 inventory. Specifically, the *Building Docks* Honesty–Humility score had incremental validity if it was able to predict the outcome when all these other variables were also included in the regression.

As Table 4 shows, the *Building Docks* Honesty–Humility assessment had incremental validity in predicting the cheating task above and beyond all variables (both in the model of Step 2a and in the model of Step 3), more specifically, it explained additional variance above and beyond all other predictors (including self-reported Honesty–Humility) in both models. However, in predicting CWB and the UBD the self-reported personality traits had the strongest predictive validity, and the *Building Docks* Honesty–Humility score was not a significant predictor in any of these models. Therefore, the *Building Docks* Honesty–Humility score did not have any incremental validity for these two outcomes.

General Discussion

In two studies, we investigated the construct validity and predictive validity of the assessment game, *Building Docks*, which was designed to assess the Honesty–Humility personality trait of the HEXACO framework. Both studies found support for the construct validity of the assessment game by demonstrating significant correlations between the Building Docks Honesty–Humility scores and self-reported Honesty–Humility. Additionally, Study 1 demonstrated that the game scores were not substantially related to intelligence, and both studies demonstrated that the game scores were not significantly correlated to the other five self-reported HEXACO traits. Furthermore, Study 2 demonstrated that *Building Docks* showed predictive validity for outcomes relevant to Honesty–Humility and had incremental validity beyond self-reported Honesty–Humility in predicting cheating, the most objective of the three outcomes. Therefore, these studies provide initial evidence of the utility of *Building Docks*.

Theoretical and Practical Implications

Our studies add to the literature in several ways. First, as Kato and De Klerk (2017) noted, the construct validity of assessment games is an understudied topic. Importantly, although constructing an assessment game to measure one or more personality traits clearly is a labor-intensive and psychometrically challenging assignment, we demonstrate that it is possible to develop a reliable and valid assessment game. We validated this game by showing its convergent validity with a well-established indicator of the construct of interest and its divergent validity with variables theoretically unrelated to the construct. Furthermore, compared with many of the prior studies using gamified assessments (e.g., McCord et al., 2019), we were able to develop an assessment game that has convergent, divergent, and predictive validity.

Second, our studies also demonstrated that purposefully designing an assessment game to assess personality has an advantage compared with assessments based on metrics from commercial games. Note that such games and their metrics were not purposefully designed to assess personality, however, they provide a clear benchmark for comparison. For instance, Tekofsky et al. (2013) investigated how self-reported FFM personality was correlated with 175 game metrics from the first-person shooter game Battlefield 3. They reported only four correlations out of their 249 significant correlations that were slightly greater than an absolute $r = .10$, specifically, the strongest correlation between a personality trait and a game metric was $r = .12$. Comparatively, the correlation coefficients obtained in the current studies were clearly higher and showcase the applicability of personality assessment games.

In terms of convergent validity, the correlation between the assessment game score with self-reported Honesty–Humility was around $r = .30$. These effects can be considered modest in terms of classical interpretations (Cohen, 1988), however, in terms of more modern standards these effects can actually be considered "large" (Funder & Ozer, 2019; Gignac & Szodorai, 2016). Furthermore, this degree of convergent validity is comparable to other personality assessment methods other than self-reports (e.g., Barends et al., 2019a; Oostrom et al., 2019). We argue that based on these modern effect size standards, there is initial evidence

Table 4. Investigating the Incremental Validity of Building Docks in the Prediction of the Outcome Variables.

	UBD	CWB	Cheating task
	β	β	*OR*
Step 1	$R^2 = .33$	$R^2 = .27$	Nagelkerke $R^2 = .07$
Gender	.13*	.07	.94
Age	−.26**	−.12*	.97*
E	−.18*	−.02	1.01
X	.23**	−.06	1.70**
A	−.30**	−.17*	.60
C	−.06	−.40**	1.33**
O	−.24**	−.03	1.42**
Step 2a	$R^2 = .34$ ($\Delta = .01$)	$R^2 = .28$ ($\Delta = .01$)	Nagelkerke $R^2 = .10$ ($\Delta = .03$*)
Gender	.12*	.06	.90
Age	−.25**	−.11	.97
E	−.18**	−.03	.98
X	.21*	−.08	1.65**
A	−.28**	−.15*	.62
C	.06	−.40**	1.34**
O	−.24**	−.03	1.39**
Building Docks H	−.10	−.09	.21*
Step 2b	$R^2 = .39$ ($\Delta = .06$**)	$R^2 = .31$ ($\Delta = .04$*)	Nagelkerke $R^2 = .08$ ($\Delta = .01$)
Gender	.07	.02	.81
Age	−.22**	−.09	.97
E	−.18*	−.02	1.01
X	−.12	−.14*	1.49**
A	−.18*	−.07	.75
C	.04	−.32**	1.61**
O	−.23**	−.02	1.42**
H	−.30**	−.24*	.59
Step 3	$R^2 = .39$ ($\Delta 2a = .05$**/$\Delta 2b = .00$)	$R^2 = .31$ ($\Delta 2a = .03$*/$\Delta 2b = .00$)	Nagelkerke $R^2 = .11$ ($\Delta 2a = .01$/$\Delta 2b = .02$*)
Gender	.07	.02	.79
Age	−.21**	−.08	.97
E	−.18*	−.03	.98
X	.12	−.15*	1.50**
A	−.17*	−.07	.74
C	.04	−.33**	1.56**
O	−.24**	−.03	1.40**
H	−.29**	−.23*	.67
Building Docks H	−.05	−.06	.24*

Note. The cheating task was analyzed using the script of Moshagen and Hilbig (2017). Gender F = 0; M = 1. UBD = unethical business decision making; CWB = Counterproductive work behavior; *OR* = odds ratio.
*$p < .05$. **$p < .01$.

of the construct validity of the assessment game. However, future research might also want to further consider the construct validity of *Building Docks* by also investigating convergent validity with observer reports of Honesty–Humility using close acquaintances.

Practically, personality assessment games may be useful to complement the traditional self-report assessment of personality. Importantly, self-reports of Honesty–Humility generally had the best predictive validity of our studied outcomes, however, the assessment game was of added value in predicting the cheating task. Notably, it seems that both self-reports and assessment games of Honesty–Humility each had unique predictive validity for the selected outcomes. This finding aligns with prior research that found that a combination of self-reports and behavioral assessments of self-control leads to high predictive validity as both methods can complement for each other's inherent weaknesses (Sharma et al., 2014).

However, it is not yet clear whether such assessment games may be especially viable in applied settings such as high stakes assessments (e.g., personnel selection; see directions for future research for further details). For scientific research, personality assessment games may have several advantages. Specifically, assessment games may improve research participants' engagement, which may lower dropout and improve data-quality in scientific research. However, *Building Docks* took more time for participants to complete than it took them to complete the long HEXACO-208 self-report instrument, which also measures five additional personality traits. For instance, *Building Docks* took participants about 45 minutes to complete, whereas the HEXACO-208, which measures the complete gamut of personality, took 30 minutes to complete. Furthermore, we want to stress that assessment games are costly to create in terms of time and human and financial resources.

Finally, we would like to address an important point: there are currently no agreed on standards for the psychometric properties of game-based assessments for both assessment games and in-game assessments (see our theoretical framework for the distinction). It seems questionable whether the psychometric standards that were developed in light of traditional assessments apply to game-based assessments (DiCerbo et al., 2017). For instance, *Building Docks* is a 38-item game with an alpha reliability of around .70. Given that a higher alpha reliability may be required for diagnostic purposes (Nunnaly & Bernstein, 1994), it seems informative to demonstrate the number of items that need to be developed and their associated costs to reach a higher alpha level (e.g., .90). Before demonstrating this, it is important to note *Building Docks* was a linear game (i.e., participants encountered all items in a fixed order), which arguably may have resulted in a higher alpha level than in-game assessments with a nonlinear structure (e.g., branching structure; a player-selected structure). For instance, prior work found that a linear version of a game-based assessment had a higher Cronbach alpha reliability of eight achievements ($\alpha = .63$) than a nonlinear variant of the same game using the same achievements ($\alpha = .50$; Kim & Shute, 2015).[11] Based on the Spearman-Brown prophecy formula (Brown, 1910), we calculated that the nonlinear variant of Kim and Shute required an additional 61 items and the linear variant an additional 33 items to reach an α of .90. Similarly, our (linear) assessment game required an additional 82 Honesty–Humility items to reach an α of .90 (based on the lowest overall reliability obtained in our studies [$\alpha = .73$]). To make a calculation, the development of our 38 items [and 17 filler and tutorial items] was about €100.000 (roughly $110.000), so about €1.800 per item.[12] Extending the linear game by another 82 items would cost an extra €147.600. Similarly, taking the rough estimate based on the findings of Kim and Shute that a nonlinear version requires twice as many items, then the total development cost of a nonlinear version of *Building Docks* (of 257 items) would have been roughly €462.600. Of course, these are rather rough estimates and future research may want to give more precise estimations of development costs and the difference in reliabilities of linear assessment games and nonlinear in-game assessments.

Therefore, for such assessment games to be used in applied settings, it is important to have a broader discussion about the psychometric standards that need to be adhered to. Nonetheless, there is a clear primacy of predictive validity in tools of personnel selection (Morgeson et al., 2007), and our Study 2 suggests that assessment games do have predictive and incremental validity.

Limitations and Directions for Future Research

The current studies are not without limitations. First, the studies were conducted in low stakes testing situations instead of the high stakes situations often used in personnel selection. Therefore, we do not yet know how viable assessment games are in such a context. However, we do want to point out that our Study 1 sample may be considered to be in a 'medium stakes' situation as they were invited to this study as part of a competition in which the best candidate could win €10.000. We clearly notified the respondents that their participation would not have any impact on their chance to win the prize but it was likely that many of the Study 1 respondents thought that they would make a good impression if they participated. The fact that this sample took this research seriously is also evidenced by the fact that none of the respondents were flagged for giving noncompliant responses on the HEXACO inventory (see Barends & De Vries, 2019).

Second, the number of participants that completed the cognitive ability test in Study 1 was somewhat limited and therefore the divergent validity with this cognitive ability test may reflect more a lack of statistical power than an absence of a true effect smaller than we were able to detect (i.e., our sensitivity analysis using G*Power 3.1.9.2 [Faul et al., 2007] indicated that an effect of $r = .31$ was the smallest effect we could detect with 80% power for the cognitive ability tests).

Third, some of the findings may be influenced by common method variance. Specially, the assessment game had no incremental validity beyond self-reported Honesty–Humility in predicting the two self-reported outcomes (i.e., CWB and UBD). This finding may be explained—at least partly—by the common method effects (Podsakoff et al., 2003). Specifically, all self-report instruments may have shared variance because they were all assessed using a self-report instrument and because they were completed by the same source. Although, as suggested by Podsakoff et al., we separated the measurements of the predictors and outcomes

by 1 week to decrease potential carry-over effects, such common method effects may have still played a role. Importantly, the findings did show that the assessment game had incremental validity in the prediction of the behavioral outcome (cheating task) beyond self-reported Honesty–Humility. A game-based assessment of persistence has found similar results, specifically, it was related to a behavioral indicator of the trait and not to a self-rated indicator (Ventura & Shute, 2013).

More broadly, our results align with prior findings that rating-scale measures of personality tend to be more highly correlated with each other than behavioral assessments of personality (Duckworth & Kern, 2011; Sharma et al., 2014). Additionally, these studies also showed that rating-scales are also somewhat more strongly related to outcomes than behavioral assessments. Furthermore, personality self-reports are much more strongly related to self-rated outcomes than outcomes gathered using other methods (Zettler et al., 2020). Similarly, although behavioral assessments are usually less strongly related (Sharma et al.), the incremental validity of Building Docks in the prediction of the cheating task could reflect a common method effect because both are behavioral assessments. Therefore, future research may want to compare the incremental validities of self-reported Honesty–Humility and *Building Docks* to objective outcomes with which neither one shares method effects (e.g., employer records of theft).

Finally, future research may want to further investigate the viability of game-based assessments of personality for personnel selection. Theoretically, an advantage of computer games, and thus our assessment game, is that they may get players in a state of flow (Sweetser & Wyeth, 2005; see Boyle et al., 2012, for a review). Flow leads people to forget their surroundings when they are playing a game. Arguably, a person in a state of flow may forget that a game is being played as a part of a personnel selection assessment and may, therefore, decrease socially desirable responding (i.e., faking). Specifically, job applicants tend to score about half a standard deviation higher on socially desirable traits (such as Honesty–Humility) than job incumbents (Anglim et al., 2017; Birkeland et al., 2006; cf. Grieve & de Groot, 2011). Potentially, *Building Docks* and other personality assessment games or in-game assessments may be able to decrease faking. However, some caution is warranted as prior research has demonstrated that a standalone gamified personality assessment was just as fakeable as a self-report personality inventory (Barends et al., 2019a). Therefore, it is an open question whether at a higher level of gamefulness it may be possible to counteract faking with a game-based personality assessment.

Similarly, an important avenue for future research is to investigate if game-based personality assessments have adverse impact. Adverse impact means that particular groups (e.g., women; ethnic minorities) receive substantially different scores than other groups (e.g., men; ethnic majority) on a selection instrument (Bartram, 1995). Traditional personality assessments do not tend to result in adverse impact against particular groups in society (Berry et al., 2012) but it is not yet known whether game-based assessments result in adverse impact. Specifically, there are some stereotypes about gaming in terms of age (McLaughlin et al., 2012) and gender (Wasserman & Rittenour, 2019). Therefore, it is important to investigate whether groups with little gaming experience are not disadvantaged by assessment games. However, the findings of the current studies did not indicate that men or women received different *Building Docks* Honesty–Humility scores, such as have been consistently found for self-reported Honesty–Humility (favoring women, who have higher scores on average; De Vries et al., 2009; Lee & Ashton, 2018). Neither did older people receive lower scores than younger people on *Building Docks* Honesty–Humility in Study 1, and only in Study 2 a significant positive correlation was obtained mirroring age differences in self-reports of Honesty–Humility (Ashton & Lee, 2016; De Vries et al., 2009). Nonetheless, if game-based personality assessments are to be used in personnel selection it is important to demonstrate that they are free from such adverse impact.

Conclusion

Our research show that assessment games, such as *Building Docks*, may be considered valid assessment tools that deserve further study. Specifically, the Honesty–Humility scores obtained in our assessment game *Building Docks* had convergent validity with self-reported Honesty–Humility and divergent validity with self-reports of the other five HEXACO personality traits and a cognitive ability test. Furthermore, we showed that assessment games can have unique predictive validity above and beyond traditional self-reported personality measures in the prediction of outcomes. Specifically, *Building Docks* was a better predictor than self-reported HEXACO personality of the probability that someone cheated for financial gain. Arguably, assessment games may also be able to assess constructs and predict outcomes that are more difficult to ascertain using self-reports. To find out more about the theoretical implications and practical utility of assessment games, we call on scholars to further investigate the use of assessment games in personnel assessment procedures and clinical practice.

Acknowledgments

We would like to thank Marian de Joode, Rob Fraats, Dean Meis and IJsfontein for their help with the design of the assessment game. We would also like to thank Labrooms for the development

of the assessment game. Additionally, we also want to thank Joost Jongeneel for his help with data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was financially supported by a grant from LTP business psychologists to support a PhD position for the first author at the VU University, The Netherlands.

ORCID iD

Ard J. Barends https://orcid.org/0000-0001-7067-4463

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Note that the term "serious games" is also frequently used (e.g., Bellotti et al., 2013; Kato & De Klerk, 2017), however, this term is controversial as it is ambiguous and as it suggests that entertainment games are not taken seriously by gamers (Klabbers, 2009). Therefore, we believe the term applied games is to be preferred.
2. This active cooperation is opposed to reactive cooperation, which is captured by HEXACO Agreeableness (Hilbig et al., 2013; Hilbig et al., 2016).
3. This progression may be considered a role manipulation of power (Galinsky et al., 2015), which has been argued to magnify the expression of personality (Galinsky et al., 2008). In the supplemental files, we show that our exploratory analysis on whether in-game power manipulation increases convergent validity between self-reported Honesty–Humility and the *Building Docks* assessment of this trait did not yield any significant results.
4. Also, note that three different versions of this cognitive ability test were used. These versions only had some overlapping items but generally have comparable correlations between the subtasks as the maximal [r] difference = .10.
5. The game was first developed in English and then translated into Dutch. The translation followed a back-translation method.
6. We wanted to avoid a competitive orientation to avoid that players followed a self-maximizing dominant strategy because this might potentially reduce the impact of personality on in-game behavior. Of course, the collective orientation instruction may also result in dominant strategies. Therefore, this collective break-even point was achieved already quite early in the game to limit the development of dominant strategies.
7. We only report the results of the *Building Docks* fillers for the other five HEXACO traits in the supplemental materials because of the limited number of items per trait (1-3).
8. We reported the alpha reliabilities using the listwise deletion procedure from SPSS as we find highly comparable findings using the pairwise deletion procedure of the *psych* package in R (Revelle, 2018). The differences are at most [α] = .01.
9. A majority of our sample (Phases 1 and 2 n = 206; all phases n = 176) was from the group that first completed the HEXACO because we informed these participants via a message when a new phase was available. However, in the group that first completed *Building Docks* (Phases 1 & 2 n = 81; all phases n = 65) we forgot to remind these participants when Phase 2 was available. Therefore, this sample contained fewer participants that completed more than one phase.
10. The results for CWB were virtually unchanged if we excluded the 26 participants who indicated that they were currently unemployed (see the supplementary materials for further details).
11. Note that we use these alpha reliability findings as an illustrative case and that Kim and Shute (2015) also found much higher reliabilities on another indicator of internal reliability (ω). However, the linear variant of the game also had higher ω than the nonlinear version.
12. We divided the estimated development cost of €100.000 by the total of 55 items (€1,818.18 per item). For ease of communication we rounded this cost per item to €1,800 and used this in the subsequent calculations. In all calculations we used the Spearman–Brown formula to determine the number of Honesty–Humility items required to reach the intended alpha level of .90. Finally, note that in the final calculation we estimated that 120 Honesty–Humility items were required for the linear version and 240 for the nonlinear version. The total number of items in this hypothetical nonlinear version of the game is 257 due to the 17 filler and tutorial items. Note that this is a rough estimation as it does not account for the economy of scale meaning that subsequent items are cheaper to implement.

References

Anglim, J., Morse, G., De Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing job applicants to non-applicants using an item-level bifactor model on the HEXACO personality inventory. *European Journal of Personality*, *31*(6), 669-684. https://doi.org/10.1002/per.2120

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150-166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2008). The prediction of Honesty-Humility-related criteria by the HEXACO and five-factor models of personality. *Journal of Research in Personality*, *42*(5), 1216-1228. https://doi.org/10.1016/j.jrp.2008.03.006

Ashton, M. C., & Lee, K. (2016). Age trends in the HEXACO-PI-R self-reports. *Journal of Research in Personality*, *64*(October), 102-111. https://doi.org/10.1016/j.jrp.2016.08.008

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139-152. https://doi.org/10.1177/1088868314523838

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psychological studies in seven languages. *Journal of Personality and Social Psychology*, *86*(2), 356-366. https://doi.org/10.1037/0022-3514.86.2.356

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2017). Studying student differentiation in gamified education: A long-term study. *Computers in Human Behavior*, *71*(June), 550-585. https://doi.org/10.1016/j.chb.2016.08.049

Barends, A. J., & De Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, *143*(June), 84-89. https://doi.org/10.1016/j.paid.2019.02.015

Barends, A. J., De Vries, R. E., & Van Vugt, M. (2019a). Gamified personality assessment: Virtual behavior cues of Honesty-Humility. *Zeitschrift für Psychologie*, *227*(3), 207-217. https://doi.org/10.1027/2151-2604/a000379

Barends, A. J., De Vries, R. E., & Van Vugt, M. (2019b). Power influences the expression of Honesty-Humility: The power-exploitation affordances hypothesis. *Journal of Research in Personality*, *82*(October), Article 103856. https://doi.org/10.1016/j.jrp.2019.103856

Bartram, D. (1995). Predicting adverse impact in selection testing. *International Journal of Selection and Assessment*, *3*(1), 52-61. https://doi.org/10.1111/j.1468-2389.1995.tb00007.x

Bauer, M., Wylie, C., Jackson, T., Mislevy, B., Hoffman-John, E., John, M., & Corrigan, S. (2017). Why video games can be a good fit for formative assessment. *Journal of Applied Testing Technology*, *18*(S1), 19-31. http://www.jattjournal.com/index.php/atp/article/view/118673

Bélise, J.-F., & Bodur, H. O. (2010). Avatars as information: Perception of consumers based on their avatars in virtual worlds. *Psychology & Marketing*, *27*(8), 741-765. https://doi.org/10.1002/mar.20354

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013*, Article 136864. https://doi.org/10.1155/2013/136864

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, *85*(3), 349-360. https://doi.org/10.1037/0021-9010.85.3.349

Berry, C. M., Kim, A., Wang, Y., Thompson, R., & Mobley, W. H. (2012). Five-factor model personality measures and sex-based differential prediction of performance. *Applied Psychology*, *62*(1), 13-43. https://doi.org/10.1111/j.1464-0597.2012.00493.x

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Branninck, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317-335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bleichrodt, N., & Van den Berg, R. H. (1999). *Handleiding MCT-H* [Manual MCT-H]. NOA/VU.

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771-780. https://doi.org/10.1016/j.chb.2011.11.020

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296-322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature of empirical evidence on computer games and serious games. *Computers & Education*, *59*(2), 661-686. https://doi.org/10.1016/j.compedu.2012.03.004

Coughlan, G., Coutrot, A., Khondoker, M., Minihane, A.-M., Spiers, H., & Hornberger, M. (2019). Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proceedings of the National Academy of Sciences*, *116*(19), 9285-9292. https://doi.org/10.1073/pnas.1901600116

De Klerk, S., & Kato, P. M. (2017). The future value of serious games for assessment: Where do we go now? *Journal of Applied Testing Technology*, *18*(S1), 32-37.

De Meijer, L. A., Born, M. Ph., Van Zielst, J., & Van Der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity. *European Psychologist*, *15*(3), 229-236. https://doi.org/10.1027/1016-9040/a000027

De Raad, B., Barelds, D. P. H., Timmerman, M. E., De Rover, K., Mlačić, B., & Church, A. T. (2014). Towards a pan-cultural personality structure: Input from 11 psycholexical studies. *European Journal of Personality*, *28*(5), 497-510. https://doi.org/10.1002/per.1953

De Vries, R. E., Ashton, M. C., & Lee, K. (2009). De zes belangrijkste persoonlijkheidsdimensies en de HEXACO persoonlijkheidsvragenlijst [The six most important personality dimensions and the HEXACO personality inventory]. *Gedrag & Organisatie*, *22*(3), 232-274.

De Vries, R. E., Tybur, J. M., Pollet, T. V., & Van Vugt, M. (2016). Evolution, situational affordances, and the HEXACO model of personality. *Evolution and Human Behavior*, *37*(5), 407-421. https://doi.org/10.1016/j.evolhumbehav.2016.04.001

De Vries, R. E., Wawoe, K. W., & Holtrop, D. (2015). What is engagement? Proactivity as the missing link in the HEXACO model of personality. *Journal of Personality*, *84*(2), 178-193. https://doi.org/10.1111/jopy.12150

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Deboutte, G., Herrewijn, L., Malliet, S., Pabian, S., Van Broeckhoven, F., De Troyere, O., Deglorie, G., Van Hoecke, S., Samyn, K., & De Bourdeaudhuij, I. (2018). The efficacy of the Friendly Attac serious digital game to promote prosocial bystander behavior in cyberbullying among young adolescents: A cluster-randomnized controlled trial. *Computers in Human Behavior*, *78*(January), 336-347. https://doi.org/10.1016/j.chb.2017.10.011

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness defining "gamification." In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9-15). Association for Computing Machinery. https://www.researchgate.net/publication/230854710_From_Game_Design_Elements_to_Gamefulness_Defining_Gamification

DiCerbo, K. E., Shute, V., & Kim, Y. J. (2017). The future of assessments in technology-rich environments: Psychometric considerations. In M. Spector, B. B. Lockee, & M. D.

Childress (Eds.) *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1-21). Springer. https://doi.org/10.1007/978-3-319-17727-4_66-1

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417-440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Dubbelt, L., Oostrom, J. K., Hiemstra, A. M. F., & Modderman, J. P. L. (2015). Validation of a digital work simulation to assess machiavellianism and compliant behavior. *Journal of Business Ethics*, *130*(3), 619-637. https://doi.org/10.1007/s10551-014-2249-x

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259-268. http://www.peggykern.org/uploads/5/6/6/7/56678211/duckworth___kern_2011_-_meta-analysis_of_self-control_measures.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism: Experimental evidence and new theories. In S.-C. Kolm & J. M. Ythier (Eds.), *Handbook of the economics of giving, altruism and reciprocity: Foundations* (Vol. 1, pp. 615-691). Elsevier Science. https://doi.org/10.1016/S1574-0714(06)01008-6

Fekken, G. C., Holden, R. R., Jackson, D. N., & Guthrie, G. M. (1987). An evaluation of the personality research form with Filipino university students. *International Journal of Psychology*, *22*(4), 399-407. https://doi.org/10.1080/00207598708246781

Fetzer, M., McNamara, J., & Geimer, J. L. (2017). Gamification, serious games and personnel selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 293-309). Wiley & Sons. https://doi.org/10.1002/9781118972472.ch14

Fleming, T, M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S. N., Cheek, C., Lucassen, M., Ming Lau, H., Pollmuller, B., & Hetrick, S. (2017). Serious games and gamification for mental health: Current status and promising directions. *Frontiers in Psychiatry*, *7*, Article 215. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5222787/

Fong, K., & Mar, R. A. (2015). What does my avatar say about me? Inferring personality from avatars. *Personality and Social Psychology Bulletin*, *41*(2), 237-249. https://doi.org/10.1177/0146167214562761

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Galinsky, A. D., Magee, J. C., Gruenfeld, D. H., Whitson, J. A., & Liljenquist, K. A. (2008). Power reduces the press of situations: Implications for creativity, conformity, and dissonance. *Journal of Personality and Social Psychology*, *95*(6), 1450-1466. https://doi.org/10.1037/a0012633

Galinsky, A. D., Rucker, D. D., & Magee, J. C. (2015). Power: Past findings, present considerations, and future directions. In M. Mikulincer & P. R. Shavers (Eds.), *APA handbook of personality and social psychology: Vol. 3: Interpersonal relations* (pp. 421-460). American Psychological Association. https://doi.org/10.1037/14344-016

Georgiou, K., Gouras, A., & Nikolauo, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International Journal of Selection and Assessment*, *27*(2), 91-103. https://doi.org/10.1111/ijsa.12240

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Grieve, R., & de Groot, H. T. (2011). Does online psychological test administration facilitate faking? *Computers in Human Behavior*, *27*(6), 2386-2391. https://doi.org/10.1016/j.chb.2011.08.001

Gundry, D., & Deterding, S. (2018). Validity threats in quantitative data collection with games: A narrative survey. *Simulation & Gaming*, *50*(3), 302-328. https://doi.org/10.1177/1046878118805515

Hilbig, B. E., Thielmann, I., Klein, S. A., & Henniger, F. (2016). Two faces of cooperation: On the unique role of HEXACO agreeableness for forgiveness versus retaliation. *Journal of Research in Personality*, *64*(October), 69-78. https://doi.org/10.1016/j.jrp.2016.08.004

Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty-Humility, social value orientation and economic behavior. *Journal of Research in Personality*, *43*(3), 516-519. https://doi.org/10.1016/j.jrp.2009.01.003

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality*, *57*(August), 72-88. https://doi.org/10.1016/j.jrp.2015.04.003

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts toward cooperation as a matter of dispositional Honesty-Humility. *European Journal of Personality*, *26*(3), 245-254. https://doi.org/10.1002/per.830

Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–humility and agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, *54*(5), 598-603. https://doi.org/10.1016/j.paid.2012.11.008

Ifenthalter, D., Eseryel, D., & Ge, X. (Eds.). (2012). *Assessment in game-based learning: Foundations, innovations, and perspectives*. Springer. https://doi.org/10.1007/978-1-4614-3546-4

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment*, *22*(3), 168-176. https://doi.org/10.1027/1015-5759.22.3.168

Kappe, R., & Van Der Flier, H. (2012). Predicting academic success in higher education: What's more important than being smart? *European Journal of Psychology of Education*, *27*(4), 605-619. https://doi.org/10.1007/s10212-011-0099-9

Kato, P. M., & De Klerk, S. (2017). Serious games for assessment: Welcome to the jungle. *Journal of Applied Testing Technology*, *18*(S1), 1-6. https://www.researchgate.net/publication/321492367_Serious_Games_for_Assessment_Welcome_to_the_Jungle

Kiili, K., Devlin, K., Perttula, A., Tuomi, P., & Lindstedt, A. (2015). Using video games to combine learning and assessment in mathematics education. *International Journal of Serious Games*, *2*(4), 37-55. https://doi.org/10.17083/ijsg.v2i4.98

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessments. *Computers & Education*, *87*(September), 340-356. https://doi.org/10.1016/j.compedu.2015.07.009

Klabbers, J. H. G. (2009). Terminological ambiguity: Game and simulation. *Simulation & Gaming*, *40*(4), 446-463. https://doi.org/10.1177/1046878108325500

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming*, *45*(6), 752-768. https://doi.org/10.1177/1046878114563660

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO personality inventory: Two new facet scales and an observer report form. *Psychological Assessment*, *18*(2), 182-191. https://doi.org/10.1037/1040-3590.18.2.182

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, *25*(5), 543-556. https://doi.org/10.1177/1073191116659134

Lee, K., Ashton, M. C., Wiltshire, J., Bourdage, J. S., Visser, B. A., & Gallucci, A. (2013). Sex, power, money: Prediction from the dark triad and Honesty-Humility. *European Journal of Personality*, *27*(2), 169-184. https://doi.org/10.1002/per.1860

Levy, L., Solomon, R., Johnson, J., Wilson, J., Lambeth, A., Gandy, M., Joann, M., Way, J., & Liu, R. (2016). Grouches, extraverts, and jellyfish: Assessment validity and game mechanics in a gamified assessment. In *Proceedings of the 1st International Joint Conference of DiGra and FDG*. Digital Games Research Association and Society for the Advancement of the Science of Digital Games. http://www.digra.org/wp-content/uploads/digital-library/paper_327.pdf

Lievens, F. (2017). Construct-driven SJTs: Towards an agenda for future research. *International Journal of Testing*, *17*(3), 1-7. https://ink.library.smu.edu.sg/lkcsb_research/5771/

Liu, Y., Holden, D., & Zheng, D. (2016). Analyzing students' language learning experience in an augmented mobile game: An exploration of an emergent learning environment. *Procedia: Social and Behavioral Sciences*, *228*(July), 369-374. https://doi.org/10.1016/j.sbspro.2016.07.055

McCord, J.-L., Harman, J. L., & Purl, J. (2019). Game-like personality testing: An emerging mode of personality assessment. *Personality and Individual Differences*, *143*(June), 95-102. https://doi.org/10.1016/j.paid.2019.02.017

McCredie, M., & Morey, L. C. (2019). Who are the turkers? A characterization of MTurk workers using the personality assessment inventory. *Assessment*, *26*(5), 759-766. https://doi.org/10.1177/1073191118760709

McCreery, M. P., Krach, S. K., Schrader, P. G., & Boone, R. (2012). Defining the virtual self: Personality, behavior, and the psychology of embodiment. *Computers in Human Behavior*, *28*(3), 976-983. https://doi.org/10.1016/j.chb.2011.12.019

McLaughlin, A., Gandy, M., Allaire, J., & Whitlock, L. (2012). Putting fun into video games for older adults. *Ergonomics in Design*, *20*(2), 13-22. https://doi.org/10.1177/1064804611435654

Morey, L. C. (2014). Borderline features are associated with inaccurate trait self-estimations. *Borderline Personality Disorder and Emotion Dysregulation*, *1*(4), 1-6. https://doi.org/10.1186/2051-6673-1-4

Morgeson, F. P., Campion, M., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683-729. https://msu.edu/~morgeson/morgeson_campion_dipboye_hollenbeck_murphy_schmitt_2007a.pdf

Moshagen, M., & Hilbig, B. E. (2017). The statistical analysis of cheating paradigms. *Behavioral Research Methods*, *49*(2), 724-732. https://doi.org/10.3758/s13428-016-0729-x

Myers, C. E., Kostek, J. A., Ekah, B., Sanchez, R., Ebanks-Williams, Y., Krusznis, A. L., Weinflash, N., & Servatius, R. J. (2016). Watch what I do, not what I say I do: Computer-based avatars to assess behavioral inhibition, a vulnerability factor for anxiety disorders. *Computers in Human Behavior*, *55*(February), 804-816. https://doi.org/10.1016/j.chb.2015.07.067

Nikolaou, I., Georgiou, K., & Kotsasarlidou, V. (2019). Exploring the relationship of gamified assessment with performance. *Spanish Journal of Psychology*, *22*, Article e6. https://doi.org/10.1017/sjp.2019.5

Nunnaly, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). McGraw-Hill.

Oostrom, J. K., De Vries, R. E., & De Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1-29. https://doi.org/10.1080/08959285.2018.1539856

Pletzer, J.-L., Bentvelzen, M., Oostrom, J. K., & De Vries, R. E. (2019). A meta-analysis of the relations between personality and workplace deviance: Big Five versus HEXACO. *Journal of Vocational Behavior*, *112*(June), 369-383. https://doi.org/10.1016/j.jvb.2019.04.004

Pletzer, J.-L., Oostrom, J. K., Bentvelzen, M., & De Vries, R. E. (2020). Comparing domain-and facet-level relations of the HEXACO personality model with workplace deviance: A meta-analysis. *Personality and Individual Differences*, *152*(January), Article 109539. https://doi.org/10.1016/j.paid.2019.109539

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Prensky, M. (2001). Fun, play, and games: What makes games engaging. In M. Prensky (Ed.), *Digital game-based learning* (pp. 106-145). McGraw-Hill.

Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Reviews of Psychology*, *28*, 363-392. https://doi.org/10.1146/annurev.ps.28.020177.002051

Revelle, W. (2018). *psych: Procedures for personality and psychological research* (Version = 1.8.12) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Richter, G., Raban, D. R., & Rafaeli, S. (2015). Studying gamification: The effects of rewards and incentives on motivation In T. Reiners & L. C. Woods (Eds.), *Gamification in education and business* (pp. 21-46). Springer International. https://doi.org/10.1007/978-3-319-10208-5_2

Saucier, G. (2009). Recurring personality dimensions in inclusive lexical studies: Indications for a big six structure. *Journal of Personality*, *77*(5), 1577-1614. https://doi.org/10.1111/j.1467-6494.2009.00593.x

Schriber, R. A., Robins, R. W., & Solomon, M. (2014). Personality and self-insight in individuals with Autism Spectrum Disorder. *Journal of Personality and Social Psychology*, *106*(1), 112-130. https://doi.org/10.1037/a0034950

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*(2), 374-408. https://doi.org/10.1037/a0034418

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Technology, Instruction, Cognition and Learning*, *8*, 137-161. https://myweb.fsu.edu/vshute/pdf/TICL2010.pdf

Starks, K. (2014). Cognitive behavioral game design: A unified model for designing serious games. *Frontiers in Psychology*, *5*, Article 28. https://doi.org/10.3389/fpsyg.2014.00028

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, *3*(3), 1-24. https://doi.org/10.1145/1077246.1077253

Tabacchi, M. E., Caci, B., Cardaci, M., & Perticone, V. (2017). Early usage of Pokémon Go and its personality correlates. *Computers in Human Behavior*, *72*(July), 163-169. https://doi.org/10.1016/j.chb.2017.02.047

Tekofsky, S., Spronck, P., Plaat, A., Van Den Herik, J., & Broersen, J. (2013). PsyOps: Personality assessment through gaming behavior. In *Proceedings of the International Conference on the Foundations of Digital Games*. https://liacs.leidenuniv.nl/~plaata1/papers/fdg2013.pdf

Thielmann, I., Akrami, N., Baborović, T., Belloch, A., Bergh, R., Chirumbolo, A., Čolović, P., De Vries, R. E., Dostál, D., Egorova, M., Gnisci, A., Heydasch, T., Hilbig, B. E., Hsu, K-Y., Izdebski, P., Leone, L., Marcus, B., Međedović, J., Nagy, J., . . . Lee, K. (2020). The HEXACO-100 across 16 languages: A large-scale test of measurement invariance. *Journal of Personality Assessment*, *102*(5), 714-726. https://doi.org/10.1080/00223891.2019.1614011

Thielmann, I., Böhm, R., & Hilbig, B. E. (2015). Different games for different motives: Comment on Haesevoets, Folmer, and Van Hiel (2015). *European Journal of Personality*, *29*(4), 506-508. https://doi.org/10.1002/per.2007

Thielmann, I., Hilbig, B. E., Zettler, I., & Moshagen, M. (2017). On measuring the sixth basic personality dimension: A comparison between HEXACO Honesty-Humility and big six honesty-propriety. *Assessment*, *24*(8), 1024-1036. https://doi.org/10.1177/1073191116638411

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30-90. https://doi.org/10.1037/bul0000217

Thomas, K. A., & Clifford, S. (2017). Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*(December), 184-197. https://doi.org/10.1016/j.chb.2017.08.038

Van der Kuil, M. N. A., Visser-Meily, J. M. A., Evers, A. W. M., & Van der Ham, I. J. M. (2018). A usability study of a serious game in cognitive rehabilitation: A compensatory navigation training in acquired brain injury patients. *Frontiers in Psychology*, *9*, Article 846. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5996119/

Van Lange, P. A. M., Balliet, D., Parks, C. D., & Van Vugt, M. (2014). *Social dilemmas: The psychology of human cooperation*. Oxford University Press.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568-2572. https://doi.org/10.1016/j.chb.2013.06.033

Wasserman, J. A., & Rittenour, C. E. (2019). Who wants to play? Cueing perceived sex-based stereotypes of games. *Computers in Human Behavior*, *91*(February), 252-262. https://doi.org/10.1016/j.chb.2018.09.003

Worth, N. C., & Book, A. S. (2014). Personality and behavior in a massively multiplayer online role-playing game. *Computers in Human Behavior*, *38*(September), 322-330. https://doi.org/10.1016/j.chb.2014.06.009

Worth, N. C., & Book, A. S. (2015). Dimensions of video game behavior and their relationships with personality. *Computers in Human Behavior*, *50*(September), 132-140. https://doi.org/10.1016/j.chb.2015.03.056

Zeigler-Hill, V., & Monica, S. (2015). The HEXACO model of personality and video game preferences. *Entertainment Computing*, *11*(November), 21-26. https://doi.org/10.1016/j.entcom.2015.08.001

Zettler, I., & Hilbig, B. E. (2010). Honesty-Humility and a person-situation interaction at work. *European Journal of Personality*, *24*(7), 569-582. https://doi.org/10.1002/per.757

Zettler, I., Hilbig, B. E., & Heydasch, T. (2013). Two sides of one coin: Honesty-Humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality*, *47*(4), 286-295. https://doi.org/10.1016/j.jrp.2013.01.012

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomilogical net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, *15*(3), 723-760. https://doi.org/10.1177/1745691619895036

Zhao, K., Ferguson, E., & Smillie, L. D. (2017). Individual differences in good manners rather than compassion predict fair allocations of wealth in the dictator game. *Journal of Personality*, *85*(2), 244-256. https://doi.org/10.1111/jopy.12237

Zhao, K., & Smillie, L. D. (2015). The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*, *19*(3), 277-302. https://doi.org/10.1177/1088868314553709