The potential of game- and video-based assessments for social attributes: examples from practice

Franziska Leutner
Experimental Psychology, University College London, London, UK
Sonia-Cristina
Clinical, Educational and Health Psychology, University College London, London, UK, and
Josh Liff and Nathan Mondragon
HireVue, Denver, Colorado, USA

Received 12 January 2020
Revised 9 July 2020
Accepted 18 August 2020

Abstract

Purpose – The purpose of this study is to describe the development and psychometric properties of a novel game- and video-based assessment of social attributes. Despite their increasing adaption, little research is available on the suitability of games and video analytics for measuring noncognitive attributes in the selection context.

Design/methodology/approach – The authors describe three novel assessments and their psychometric properties in a sample of 1,300 participants: a game-based adaptation of an Emotion Recognition Task, a chatbot-based situational judgment test for emotion management and a video-based conscientiousness assessment.

Findings – The novel assessments show good to moderate convergent validity for Emotional Recognition ($r = 0.42$), Emotion Management ($r = 0.39$) and Conscientiousness ($r = 0.21$). The video-based assessment demonstrates preliminary predictive validity for self-reported work performance. Novel game-based assessments (GBAs) are perceived as better designed and more immersive than traditional questionnaires. Adverse impact analysis indicates small group differences by age, gender and ethnicity.

Research limitations/implications – Predictive validity findings need to be replicated using objective measures of performance, such as performance ratings by supervisors and extended to the GBAs. Adverse impact should be evaluated using a real-life applicant pool and extended to additional groups.

Practical implications – Evidence for the psychometric validity of novel assessment formats supports their adoption in selection and recruitment. Improved user experience and shortened assessment times open up new areas of application.

Originality/value – This study gives first insights into psychometric properties of video- and game-based assessments of social attributes.

Keywords Emotional intelligence, Personality, Selection, Recruitment

Paper type Research paper

Introduction

The assessment of social attributes has a long history in psychological research due to their importance in predicting future job performance (Barrick and Mount, 1991). Consequently, the development of valid and reliable assessments measuring social attributes has received considerable attention. Most existing assessments follow traditional self-reported questionnaire formats, with little innovation in presentation format and assessment

Thank you to everyone involved in creating and supporting the game and video based assessments described in this article: Luca Boschetti, Maurizio Attisani, Clemens Aichholzer, Alixe Lay, Jasmin Roberts, Laryn Brown, Lindsey Zuloaga, Theodoros Bitsakis, and the HireVue Data Science team, Tom Cornell, Adam Bradshaw and the HireVue IO Psychology team.

Funding: Development of the game and video based assessments was funded by HireVue.

Journal of Managerial Psychology

0268-3946
DOI 10.1108/JMP-01-2020-0023

experience. However, recent technological developments, in particular video interviews and game-based assessments (GBAs), enable the use of alternative data sources for the assessment of psychological traits (Chamorro-Premuzic *et al.*, 2017).

Despite the increased number of studies describing the development and psychometric properties of GBAs, there is a lack of such assessments for social attributes. Most existing gamified assessments measure cognitive ability. Video interviews have mainly been studied outside of the work context, with few studies on video interviews for selection or psychometric assessment (Nguyen and Gatica-Perez, 2016; Rupasinghe *et al.*, 2016).

Social attributes and employability

Employers evaluate potential hires across multiple characteristics. Beyond cognitive ability, this includes social attributes and personality (Hogan *et al.*, 2013). Social attributes are important determinants of job success, predicting subjective career success, personal effectiveness and self-career evaluations (Heggestad and Morrison, 2008; Ng *et al.*, 2005), in particular where jobs include managerial tasks (Farh *et al.*, 2012). They exhibit low adverse impact, a significant advantage over cognitive ability tests (Newman and Lyon, 2009).

In the present study, we focus on Emotion Recognition, Emotion Management and Conscientiousness; Emotion Recognition and Management are widely researched and linked with job performance at both managerial and nonmanagerial levels (Joseph and Newman, 2010). Emotion Recognition describes the ability to identify emotion-based cues, read and recognize emotions in others (Bänziger *et al.*, 2009). Emotion Recognition is related to managerial job performance and outcomes such as negotiation performance (Elfenbein *et al.*, 2007). Emotion Management describes how an individual manages, enhances positive and regulates negative emotions of others (Allen *et al.*, 2015; MacCann and Roberts, 2008). Emotion Recognition and Emotion Management increase team effectiveness and job performance (Farh *et al.*, 2012), as well as transformational leadership behaviors (Rubin *et al.*, 2005).

Conscientiousness is the personality trait most consistently linked with job performance outcomes across occupational groups (Barrick and Mount, 1991). Conscientiousness correlates with contextual performance, customer service orientation and dealing with others in an organizational context (Avis *et al.*, 2002), as well as predicting leadership effectiveness (e.g. Cavazotte *et al.*, 2012). In contrast, the remaining Big Five personality traits, agreeableness, neuroticism, openness and extraversion, are less consistently linked with job performance, with their predictive validity varying across job roles and industries (Barrick and Mount, 1991).

Personality assessment based on free text and video

Personality describes individual differences in behavioral tendencies and preferences. A rich body of research demonstrates the link between behavioral cues, in particular language use and the Big Five personality traits (Kern *et al.*, 2014; Pennebaker and King, 1999). Language use in written text is moderately correlated with the Big Five traits ($r = 0.37$ for Conscientiousness; Park *et al.*, 2015). Nonverbal cues are also linked with personality, and personality judgments made by others (Friedman and Miller-Herringer, 1991). Both language use and nonverbal cues are readily observed with videos; audio and visual features in YouTube clips predict observer ratings of the Big Five personality traits (Biel *et al.*, 2012). Prediction accuracy and relative importance of nonverbal cues and language use vary by Big Five traits; nonverbal content is particularly important for predicting extraversion, a trait linked to communication. For Conscientiousness, verbal content is more important than nonverbal content ($r = 0.19$ for language use only, compared with $r = 0.23$ for language use and nonverbal cues; Biel *et al.*, 2013).

Game-based assessments

There is increased interest in GBAs for recruitment practices (Chamorro-Premuzic *et al.*, 2017). GBAs offer short assessment times, high quality and quantity of data, and an engaging user experience (Lumsden *et al.*, 2016; Quiroga *et al.*, 2016). Preliminary studies indicate that game features such as real-time feedback, progression through levels and clear goals result in a more motivating testing environment and increase the flow experienced by users (Burgers *et al.*, 2015; Chen, 2007; Connolly *et al.*, 2012; Wood *et al.*, 2004). By framing assessments as challenges, lower levels of user anxiety are achieved, something that is particularly relevant in high-stakes assessments (Alter *et al.*, 2010; McPherson and Burns, 2008). By contrast, traditional assessments are typically longer and may induce stress in, or fail to motivate, test takers (Tlili *et al.*, 2016). Publications of GBAs predominantly describe GBAs of cognitive ability, showing promising validity in a range of settings (Atkins *et al.*, 2014; Gamberini *et al.*, 2010; Jimison *et al.*, 2004; Verhaegh *et al.*, 2013). In contrast, there is a dearth of GBAs for social abilities.

Aims and hypotheses

Given the lack of validated psychometric assessments using innovative formats, technological developments in video-based profiling of personality, and the success of GBAs in cognitive domains, we evaluate and test the psychometric properties and user experience of three novel assessments of social attributes; a video interview assessing Conscientiousness and two GBAs measuring Emotion Recognition and Emotion Management.

H1. The Emotion Management GBA correlates strongly with similar method, and moderately with different method measures of Emotion Management (convergent validity), and correlates weakly with Emotional Intelligence (discriminant validity).

H2. The Emotion Recognition GBA correlates strongly with a similar method Emotion Recognition Task (convergent validity), moderately with measures of emotionality and weakly with Emotional Intelligence (discriminant validity).

H3. The video-based Conscientiousness assessment correlates moderately with a different method measure of Conscientiousness (convergent validity), and weakly with agreeableness and neuroticism (negatively), and not with openness and extraversion (discriminant validity).

H4. Both the video- and questionnaire-based Conscientiousness assessments correlate significantly and positively with self-report work performance, and being in work (currently having a job) (predictive validity).

H5. User experience ratings for the novel GBAs will be better than for traditional self-report assessments.

According to the Uniform Guidelines on Employee Selection Procedures (1978), any measure used for selection must demonstrate a lack of adverse impact. We evaluate adverse impact for ethnicity, age and gender.

Method

Participants and measures

1,377 participants were recruited using the online panel service Prolific Academic, and compensated for their time (age mean = 33, SD = 10.45; 56% female). Data were collected through two separate panels: in sample 1 (H3), 718 participants completed the novel video personality assessment using the HireVue video interview platform and a Big Five questionnaire (Goldberg, 1999). In sample 2,765 participants (H1 and H2), including 70 also having completed the first panel (H4), completed the two novel GBAs and the remaining

assessment measures included in this study, using the MindX app. The user experience questionnaire was completed by a subset of 77 participants after the GBAs, and 75 after the questionnaire-based assessments (H5). We collected data through separate panels because of technical constraints related to administering the different assessments to participants.

Traditional assessment measures

Sample 1

The Big Five Inventory (BFI; John and Srivastava, 1999) is a 44-item measure of Big Five personality traits conscientiousness ($\alpha = 0.82$), extraversion ($\alpha = 0.86$), openness ($\alpha = 0.84$), neuroticism ($\alpha = 0.87$) and agreeableness ($\alpha = 0.76$) measured on a 5-point Likert scale ("Strongly disagree" to "Strongly agree").

Sample 2

Geneva Emotion Recognition Test (GERT-S; Schlegel and Scherer, 2016) measures the ability to recognize emotions in others. Fourteen emotions (six positive, eight negative) are presented in 42 short audio–video clips with different intensities and verbal content.

Situational Judgement Test of Emotion Management (STEM; MacCann and Roberts, 2008) measures how people manage others' emotions. An adapted 16-scenario version of STEM is used, where only items describing workplace situations were selected.

The COPE inventory – Disengagement scale (Carver *et al.*, 1989) measures how individuals engage in Emotion Management when faced with stressful situations. Participants indicated how frequently they used each coping strategy over 15 items on a 4-point Likert scale ("I usually do not do this at all" to "I usually do this a lot"). The disengagement scale showed good internal consistency (Cronbach's alpha $\alpha = 0.78$).

The Trait Emotional Intelligence Questionnaire – Short Form (TEIQue-S; Petrides and Furnham, 2006) measures behavioral tendencies and preferences related to Emotional Intelligence, grouped into four facets of self-control, emotionality, sociability and well-being. TEIQue has 30 items answered on a 7-point Likert scale ("Disagree completely" to "Agree completely"), and showed excellent internal consistency (Cronbach's alpha $\alpha = 0.91$).

Self-Reported In-Role Job Performance questionnaire (Podsakoff and MacKenzie, 1989) assesses the quality and quantity of in-role work tasks across six items on a 7-point Likert scale ("Strongly disagree" to "Strongly agree"), and showed good internal consistency (Cronbach's alpha $\alpha = 0.72$).

Employed measured by asking: "Are you currently employed?", answered as "yes" or "no".

User experience

User experience survey (O'Brien and Toms, 2010). Adapted 22 item measure of user reactions to interactive systems. Two factors were extracted: Design task, e.g. "I liked the graphics used on the questionnaire/game" (Cronbach's alpha = 0.93) and Involvement, e.g. "This questionnaire/game experience was fun" (Cronbach's alpha = 0.98). After deleting five items that lowered the factor alpha, each factor included 11 items.

Novel assessment measures

Each novel assessment was developed to measure a distinct social attribute, and modeled after tasks used in existing psychometric assessments of the given attribute (see Figure 1):

Sample 1

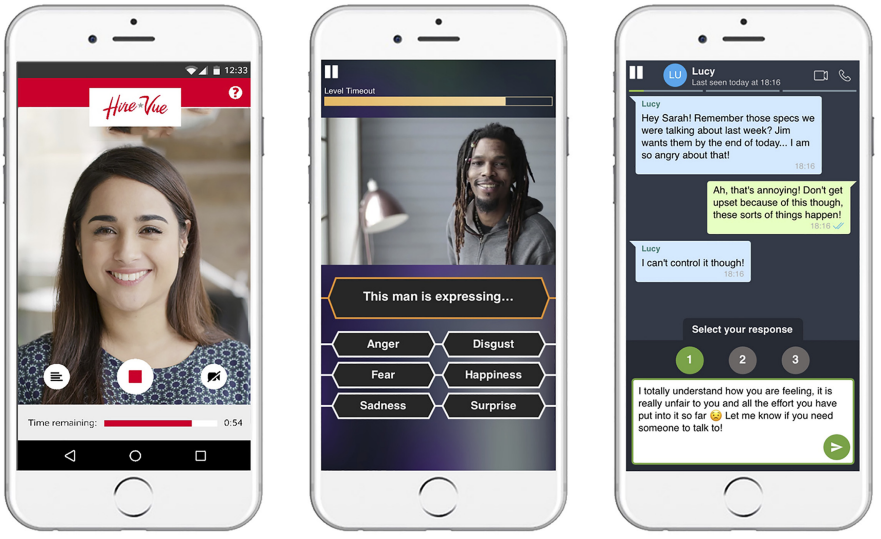
Conscientiousness recorded video interview (Hirevue Inc, 2018) with six questions designed to elicit responses that vary across personality traits. For example, the question "Tell me about a

time you had a month or more to prepare for an important presentation [. . .]" was asked to elicit responses related to Conscientiousness.

To score video interviews, responses were transcribed to text using HireVue's proprietary algorithms. Over 400 variables relating to word use, meaning and sentiment were extracted using the Natural Language Toolkit and the General Inquirer (Loper and Bird, 2002; Stone *et al.*, 1966). Spectral audio characteristics were extracted using a feature learning algorithm for the Big Five personality traits (Carbonneau *et al.*, 2016). Facial expressions were extracted using automatic expression analysis based on the Facial Action Coding System (Ekman and Rosenberg, 1997; Friesen and Ekman, 1978). All variables were entered into a machine learning prediction model, with self-reported Conscientiousness scores as the outcome variable. The model produces a single score for Conscientiousness.

Sample 2

Emotion Recognition GBA. Gamified Emotion Recognition Task that creates a motivating and enjoyable testing environment. Game elements include player advancement through levels of increasing difficulty, level timers, perceptual aspects (sounds, videos, countdowns) and feedback on player's performance at the end of the game (Robison and Bellotti, 2013). The game was designed using the evidence-centered design framework for assessments (Mislevy *et al.*, 2003), measuring the ability to identify expressions associated with Ekman's (1992) six basic emotions anger, fear, disgust, happiness, surprise and sadness in videos (the competency model). The game environment elicits the competency of interest (the task model): participants identify emotions displayed in video clips showing either a full person, a face or a close up. People shown are diverse in terms of gender, ethnicity and age to account for cultural familiarity effects (Elfenbein and Ambady, 2003). Evidence of performance on the game is operationalized as several variables including win ratio of levels won versus levels played, and the highest level reached (the evidence model).



Note(s): Example GBA Emotion Recognition task. Example GBA Emotion Management task. GBAs were hosted on the MindX mobile app

Figure 1.
From left to right: Interviews for the video-based Conscientiousness assessment are conducted using the HireVue video interviewing platform (Hirevue Inc, 2018)

Emotion Management GBA. A chatbot-based Situational Judgement Tests (SJTs) where participants chat with a fictitious colleague. The messenger increases assessment fidelity, by simulating everyday interaction, which is particularly realistic for the measurement of interpersonal skills and enhances validity estimates (Christian *et al.*, 2010; Weekley *et al.*, 2015). Message content was based on critical incidents eliciting Emotion Management strategies, collected from employees and managers in various roles and in a variety of contexts (Bledow and Frese, 2009; Motowidlo *et al.*, 1990). Response options were designed to reflect varying levels of Emotion Management, based on theoretical models (Carver *et al.*, 1989; Kaplan *et al.*, 2014), and scored based on effectiveness ratings provided by subject matter experts (Weekley *et al.*, 2006).

Results

Descriptive statistics show a good distribution of scores on the novel assessments (see Table 1).

Validity

H1 was supported. The Emotion Management GBA correlated strongly with the similar method SJT, STEM ($r = 0.39$), and less strongly with the COPE Likert scale assessment ($r = 0.31$; convergent validity) (see Table 2). The GBA also correlated weakly with Emotional Intelligence, with stronger correlations with related TEIQue subfacet Emotionality than with remaining subfacets ($r = 0.22$) (TEIQue, $r = 0.16$; discriminant validity) (see Table 3).

H2 was also supported. The Emotion Recognition GBA showed stronger correlations with the similar method Emotion Recognition Task (GERT-S, $r = 0.42$; convergent validity), and weaker correlations with Emotional Intelligence ($r = 0.15$; discriminant validity).

Table 1.
Descriptive statistics for scores on the new and standard assessments used in the study

		Mean	SD	Min	Max	*N*	Mins
Novel assessments (standardised scores)	Emotion Management	0.00	1	−3.42	2.06	729	6
	Emotion Recognition	0.00	1	−3.64	2.45	729	6
	Conscientiousness-video	0.00	1	−3.46	3.17	718	3
Traditional Emotion Management assessments	STEM	7.35	1.57	0	10.5	729	10
	COPE-disengagement	1.62	0.62	1	4	729	5
Traditional Emotion Recognition Task	GERT-S	23.47	6.00	5	36	729	10
Traditional Emotional Intelligence assessment	TEIQue	4.73	0.93	1.4	7	634	10
	Emotionality	4.45	1.08	1	7	634	
	Well-being	5.09	1.22	1	7	634	
	Sociability	4.62	1.09	1	7	634	
	Self-control	4.87	1.03	1.75	7	634	
Traditional Big Five	Conscientiousness	35.30	5.51	15	45	718	3.5
	Agreeableness	34.58	4.22	16	44	718	3.5
	Extraversion	25.57	6.33	9	40	718	3.5
	Openness	37.87	5.79	17	50	718	3.5
	Neuroticism	22.11	6.61	8	40	718	3.5
User experience: design	Games	50.58	8.52	27	65	77	4
	Questionnaire	42.48	8.43	0	55	75	4
User experience: task involvement	Games	41.35	8.15	16	55	77	4
	Questionnaire	39.01	6.10	25	55	75	4
Outcome measures	Self-report performance	21.5	3.01	7	25	718	
	Employed	Yes = 498, No = 220					

H3 was partially supported. The video-based assessment correlated weakly with Conscientiousness ($r = 0.21$; convergent validity), and was not significantly correlated with the remaining Big Five traits (discriminant validity) (see Table 3).

Predictive validity of Conscientiousness

H4 was supported. Scores on the video-based and traditional Conscientiousness assessments correlated significantly and positively with self-reported performance ($r = 0.15$ and $r = 0.53$, respectively; calculated using the sub sample of $N = 498$ respondents currently in work), and with being employed ($r = 0.12$ and $r = 0.11$, respectively).

User experience

H5 was supported. Emotion Management and Recognition GBAs were experienced as more immersive and better designed than traditional measures (see Figure 2).

Adverse impact

Adverse impact was absent for all protected groups, except Hispanic and Latino on the Emotion Management measure (see Table 4). This group had a small group differences effect

	1	2	3	4
1. Emotion Recognition	–	–	–	–
2. Emotion Management	0.26	–	–	–
3. GERT-S	*0.42*	0.33	–	–
4. STEM	0.25	*0.39*	0.30	–
5. COPE	0.28	*0.31*	0.20	0.23
Conscientiousness video and Big Five	*0.21*			

Note(s): Target convergent validity coefficients between the novel and their established counterpart are highlighted in italic
All correlations are significant at $p < 0.01$
The correlation between the novel and established assessment of Conscientiousness is calculated in a separate sample, correlations with the remaining constructs are therefore not available

Table 2. Convergent validity of the novel assessments measured as correlation coefficients with established measures of the respective trait

	Emotion Management	Emotion Recognition		Conscientious-video
TEIQue	0.16**	0.15**	Conscientiousness	0.21**
N	634	634	*N*	718
TEIQue- Self-Control	0.03	0.07	Agreeableness	0.07
N	604	604	*N*	718
TEIQue- Well-Being	0.1**	0.08*	Extraversion	0.01
N	613	613	*N*	718
TEIQue- Sociability	0.09*	0.08*	Openness	−0.03
N	614	614	*N*	718
TEIQue- Emotionality	0.22**	0.16**	Neuroticism	−0.07
N	597	597	*N*	718

Note(s): *Correlations significant at $p < 0.01$ **Correlations significant at $p < 0.05$
Correlations between the novel game-based assessments and the Big Five, as well as the novel Conscientiousness assessment and Emotional Intelligence did not reach sufficient statistical power to detect significant relationships because only 70 participants completed all measures in the study, and are therefore omitted (194 participants were needed to detect an expected correlation of 0.2 between the novel and traditional assessment scores, $p = 0.05$, $\beta = 0.20$)

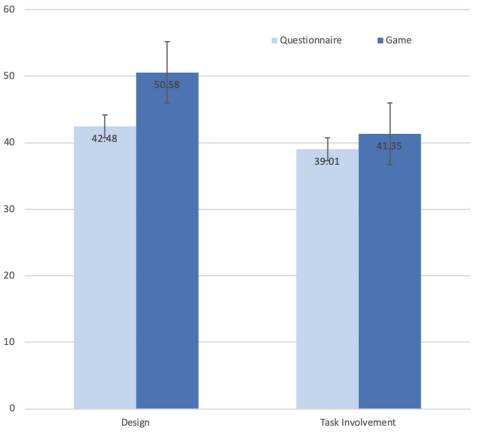
Table 3. Convergent and discriminant validity correlations between novel game-based assessments and related personality constructs, as well as between novel video-based assessments and related personality constructs

compared to the majority group (White) as measured by Cohen's D, but not as measured by the 2 SD statistic.

Discussion

This study investigated the psychometric properties of a video-based Conscientiousness assessment and two GBAs measuring Emotion Recognition and Emotion Management. Results indicated that these novel assessment formats show some good evidence of convergent and discriminant validity (H1–H3), low adverse impact for age, gender and ethnicity, and deliver a favorable assessment experience compared with traditional formats (H5). Preliminary results indicate that the predictive validity for job performance (being employed) is similar for Conscientiousness assessed by video interview and traditional measures (H4) (Avis *et al.*, 2002; Barrick and Mount, 1991). However, job performance data in this study are self-reported and has limited reliability. Of note, the self-report Conscientiousness measure showed higher correlations with self-report work performance than the video assessment, which may be due to common method variance.

Several factors contribute to the strength of between-measure correlations. Lower correlations are expected for measures that are shorter (e.g. Herzberg and Brähler, 2006) or use different assessment methods (Campbell and Fiske, 1959). Common-method variance inflates correlations between variables that are measured using the same methodology (see Podsakoff *et al.*, 2003 and Spector, 2006 for a discussion of common method variance). Indeed, convergent validity correlations varied in strength across the social attributes measured ($r = 0.24$ to $r = 0.42$) and were lower than those observed between traditional and GBA-based cognitive ability measures ($r = 0.93$ in Quiroga *et al.*, 2015 versus $r = 0.37$ in Atkins *et al.*, 2014), which have similar formats. The highest method variance in this study was present for the video and Likert scale



Note(s): Error bars show the standard error. Mean differences are significant for both design ($t(150) = -5.63$, $p < 0.001$)and task involvement ($t(150) = -2.001$, $p = 0.047$) Mean differences are stronger for the design component, indicating that better design and ease of use are the most appealing aspect of GBAs, followed by an advantage in task involvement compared with questionnaires

Figure 2.
Mean user experience scores for the novel game-based assessments and traditional questionnaires

	Total	Mean	SD	Cohen's D	95% CI lower	95% CI upper
Emotion Management						
Male	318	−0.06	0.99	−0.22	−0.37	0.07
Female	405	0.05	1.01		−0.14	0.14
White	457	0.01	0.95		−0.13	0.13
East Asian/other Asian	51	0.08	1.01	−0.18	−0.47	0.10
Black or African American	79	0.01	1.02	−0.16	−0.40	0.08
Hispanic or Latino	72	−0.07	1.13	−0.49	−0.74	−0.24
<= 40	594	0.00	1.02	−0.09	−0.28	0.09
>40	134	0.00	0.91		−0.24	0.24
Emotion Recognition						
Male	318	−0.13	1.02	−0.12	−0.26	0.03
Female	405	0.11	0.97		−0.14	0.14
White	457	0.07	0.96	−0.04	−0.32	0.25
East Asian/other Asian	51	−0.11	1.09		−0.39	0.39
Black or African American	79	−0.08	0.97	−0.04	−0.39	0.31
Hispanic or Latino	72	−0.41	1.15	−0.05	−0.40	0.31
<= 40	594	−0.02	1.00		−0.11	0.11
>40	134	0.08	0.95	−0.04	−0.23	0.14
Conscientiousness						
Male	*351*	0.02	1.04			
Female	*363*	0.01	0.95	−0.05	−0.20	0.10
White	398	0.12	0.96			
East Asian/other Asian	210	−0.12	1.03	−0.24	−0.41	−0.07
Black or African American	31	−0.14	0.91	−0.29	−0.66	0.07
Hispanic or Latino	75	−0.20	1.06	−0.37	−0.63	−0.13
<= 40	587	−0.04	1.03	−0.28	−0.47	−0.08
>40	127	0.22	0.83			

Note(s): For gender, males are the reference group, for age, under 40s, and for ethnicity, white
+/− 2 SD or higher indicates a statistically significant difference between expected and observed selection rates for the different groups; Cohen's D > |0.2| is considered a small effect, > |0.5| moderate and > |0.8| large

Table 4. Adverse impact measured through tests of mean differences and standard deviation as recommended in the Uniform Guidelines on Employee Selection Procedures (1978)

Conscientiousness assessments. It is also where the weakest convergent validity was observed ($r = 0.21$, $p < 0.001$). For comparison, two Likert scale Conscientiousness assessments will typically correlate around $r = 0.60$ (Gosling *et al.*, 2003). However, the convergent validity correlations observed in this study are in line with other alternative assessments of Conscientiousness; interviewers' ratings of Conscientiousness correlate at $r = 0.26$ with self-report assessments (Cortina *et al.*, 2000). Behavioral and performance measures of persistence (a facet of Conscientiousness) correlate more strongly ($r = 0.22$) than behavioral and self-reported measures ($r = -0.06$ to 0.09) (Ventura and Shute, 2013).

Tests of construct validity are equally affected by common method variance, such that correlations between the novel and traditional measures of unrelated constructs are lower than those typically observed if the novel assessment uses a different method (Bagozzi *et al.*, 1991). This is the case for the video Conscientiousness assessment, where we observe no significant correlation with agreeableness. Conscientiousness and agreeableness typically correlate moderate to high, with a correlation of $\rho = 0.39$ reported in meta-analysis (Mount *et al.*, 2005). Correlations are lower in same method studies for extraversion (at $\rho = 0.17$), and openness (at $\rho = 0.09$; Mount *et al.*, 2005). Contradictory to existing research, we observed no significant correlation between Conscientiousness and Emotional Stability. Meta-analysis of same method assessments of the two traits detects high correlations at $\rho = 0.52$ (Mount *et al.*, 2005). This might be a result of method variance, or the video assessment not measuring aspects of Conscientiousness related to Emotional Stability.

Practical implications, limitations and future research

Organizations are increasingly using novel assessment formats. This trend is likely to continue and expand, given significant venture capital investment in assessment companies (Bersin, 2018). This study offers a resource for researchers, assessment developers and organizations wishing to utilize novel assessments as part of their selection processes.

In the applied context, predictive validity is key to confirm the suitability of a measure for application in selection. This study gives preliminary indication that predictive validity of Conscientiousness assessments is retained in the novel video-based format. However, this result is based on a small sample of self-reported performance data. No predictive validity data were available for the game-based measures. Further research testing predictive validity of novel assessment formats using objective performance measures such as supervisor ratings of performance, income or career progression are needed.

A limitation of the user experience data in this study is that it was collected in low stakes settings and based on self-report. Future research should investigate how novel assessment formats affect the user experience in high stakes settings, in particular in regards to reducing test taker anxiety and improving data quality. Think-alouds and user interviews should be used in addition to self-reported data.

Novel assessment formats also raise validity and reliability considerations in addition to those typically addressed for psychometric assessments, which were not addressed in the scope of this study. Adverse impact data in this study were limited to specific protected groups (age, gender, typical US ethnicity groups). Further research is needed to confirm how the novel assessment formats impact additional groups, in particular those likely affected by the respective formats.

For example, the video interview format might impact those with differing speech and language abilities, accents or dialects. There is little research on the impact of using novel assessment methods on adverse impact. Artificial intelligence algorithms are facing increasing scrutiny for the discriminatory effects they might have on minority groups. Despite the use of such algorithms in novel assessment formats, such as digital footprints, free text or video analysis to predict personality, adverse impact analysis is not typically included in published research. This study offers a first evaluation of adverse impact in alternative assessment formats. However, it is limited to a small sample and specific ethnic and protected groups.

Similarly, game-based formats could disadvantage groups with low digital fluency, limited dexterity or different device types. One of the GBAs showed small group differences in scores for Latino participants, and further investigation is needed to establish whether this effect is retained in larger samples, and what might cause group differences.

Conclusions

Novel assessment formats present an exciting opportunity for the wider application of assessments, with modern user experience and shortened testing times. This study describes the first evaluation of novel and machine learning-based assessment formats, evaluating the adverse impact and psychometric properties of video- and game-based assessments. Results are encouraging for practitioners wishing to apply novel assessments, showing preliminary evidence of construct validity, a weak relationship with self-reported performance and an improved assessment experience. Further research is needed to address adverse impact, in particular group differences that might arise from novel assessment formats. This study may serve as a reference point for practitioners and researchers developing novel assessment formats and evaluating their convergent and predictive validity.

References

Allen, V., Rahman, N., Weissman, A., MacCann, C., Lewis, C. and Roberts, R.D. (2015), "The situational test of emotional management–brief (STEM-B): development and validation using item response theory and latent class analysis", *Personality and Individual Differences*, Vol. 81, pp. 195-200.

Alter, A.L., Aronson, J., Darley, J.M., Rodriguez, C. and Ruble, D.N. (2010), "Rising to the threat: reducing stereotype threat by reframing the threat as a challenge", *Journal of Experimental Social Psychology*, Vol. 46 No. 1, pp. 166-171.

Atkins, S., Sprenger, A., Colflesh, G., Briner, T.L., Buchanan, J.B., Chavis, S.E., Chen, S.Y., Iannuzzi, G.L., Kashtelyan, V., Dowling, E. and Harbison, J.I. (2014), "Measuring working memory is all fun and games", *Experimental Psychology*, Vol. 61 No. 6, pp. 417-438.

Avis, J.M., Kudisch, J.D. and Fortunato, V.J. (2002), "Examining the incremental validity and adverse impact of cognitive ability and conscientiousness on job performance", *Journal of Business and Psychology*, Vol. 17 No. 1, pp. 87-105.

Bagozzi, R.P., Yi, Y. and Phillips, L.W. (1991), "Assessing construct validity in organizational research", *Administrative Science Quarterly*, Vol. 36, pp. 421-458.

Bänziger, T., Grandjean, D. and Scherer, K.R. (2009), "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT)", *Emotion*, Vol. 9 No. 5, pp. 691-704.

Barrick, M.R. and Mount, M.K. (1991), "The big five personality dimensions and job performance: a meta-analysis", *Personnel Psychology*, Vol. 44 No. 1, pp. 1-26.

Bersin, J. (2018), "AI in HR: a real killer app", *Forbes*, available at: https://www.forbes.com/sites/joshbersin/2018/06/18/ai-in-hr-a-real-killer-app/#457ed80348f1.

Biel, J.I., Teijeiro-Mosquera, L. and Gatica-Perez, D. (2012), "Face tube: predicting personality from facial expressions of emotion in online conversational video", *Proceeding of the 14th ACM International Conference on Multimodal Interaction*, pp. 1-4, 2012, d.

Biel, J.I., Tsiminaki, V., Dines, J. and Gatica-Perez, D. (2013), "Hi youtube! Personality impressions and verbal content in social video", *Proceeding of the 15th ACM International Conference on Multimodal Interaction*, 2013, pp. 119-126.

Bledow, R. and Frese, M. (2009), "A situational judgment test of personal initiative and its relationship to performance", *Personnel Psychology*, Vol. 62 No. 2, pp. 229-258.

Burgers, C., Eden, A., van Engelenburg, M.D. and Buningh, S. (2015), "How feedback boosts motivation and play in a brain-training game", *Computers in Human Behavior*, Vol. 48, pp. 94-103.

Campbell, D.T. and Fiske, D.W. (1959), "Convergent and discriminant validation by the multitrait-multimethod matrix", *Psychological Bulletin*, Vol. 56 No. 2, pp. 81-105.

Carbonneau, M.A., Granger, E., Attabi, Y. and Gagnon, G. (2016), "Feature learning from spectrograms for assessment of personality traits", *IEEE Transactions on Affective Computing*, Vol. 1, pp. 1-12.

Carver, C.S., Scheier, M.F. and Weintraub, J.K. (1989), "Assessing coping strategies: a theoretically based approach", *Journal of Personality and Social Psychology*, Vol. 56 No. 2, pp. 267-283.

Cavazotte, F., Moreno, V. and Hickmann, M. (2012), "Effects of leader intelligence, personality and emotional intelligence on transformational leadership and managerial performance", *The Leadership Quarterly*, Vol. 23 No. 3, pp. 443-455.

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D. and Sherman, R. (2017), "The datafication of talent: how technology is advancing the science of human potential at work", *Current Opinion in Behavioral Sciences*, Vol. 18, pp. 13-16.

Chen, J. (2007), "Flow in games (and everything else)", *Communications of the ACM*, Vol. 50 No. 4, pp. 31-34.

Christian, M.S., Edwards, B.D. and Bradley, J.C. (2010), "Situational judgment tests: constructs assessed and a meta-analysis of their criterion-related validities", *Personnel Psychology*, Vol. 63 No. 1, pp. 83-117.

Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T. and Boyle, J.M. (2012), "A systematic literature review of empirical evidence on computer games and serious games", *Computers and Education*, Vol. 59 No. 2, pp. 661-686.

Cortina, J.M., Goldstein, N.B., Payne, S.C., Davison, H.K. and Gilliland, S.W. (2000), "The incremental validity of interview scores over and above cognitive ability and conscientiousness scores", *Personnel Psychology*, Vol. 53 No. 2, pp. 325-351.

Ekman, P. (1992), "An argument for basic emotions", *Cognition and Emotion*, Vol. 6 Nos 3-4, pp. 169-200.

Ekman, P. and Rosenberg, E.L. (Eds) (1997), *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, Oxford.

Elfenbein, H.A. and Ambady, N. (2003), "When familiarity breeds accuracy: cultural exposure and facial emotion recognition", *Journal of Personality and Social Psychology*, Vol. 85 No. 2, pp. 276-290.

Elfenbein, H.A., Der Foo, M., White, J., Tan, H.H. and Aik, V.C. (2007), "Reading your counterpart: the benefit of emotion recognition accuracy for effectiveness in negotiation", *Journal of Nonverbal Behavior*, Vol. 31 No. 4, pp. 205-223.

Farh, C.I., Seo, M.G. and Tesluk, P.E. (2012), "Emotional intelligence, teamwork effectiveness, and job performance: the moderating role of job context", *Journal of Applied Psychology*, Vol. 97 No. 4, pp. 890-900.

Friedman, H.S. and Miller-Herringer, T. (1991), "Nonverbal display of emotion in public and in private: self-monitoring, personality, and expressive cues", *Journal of Personality and Social Psychology*, Vol. 61 No. 5, pp. 766-775.

Friesen, E. and Ekman, P. (1978), *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto.

Gamberini, L., Cardullo, S., Seraglia, B. and Bordin, A. (2010), "Neuropsychological testing through a Nintendo Wii console", *Annual Review of Cybertherapy and Telemedicine*, Vol. 8, pp. 22-25.

Goldberg, L.R. (1999), "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models", *Personality Psychology in Europe*, Vol. 7 No. 1, pp. 7-28.

Gosling, S.D., Rentfrow, P.J. and Swann, W.B. Jr (2003), "A very brief measure of the Big-Five personality domains", *Journal of Research in Personality*, Vol. 37 No. 6, pp. 504-528.

Heggestad, E.D. and Morrison, M.J. (2008), "An inductive exploration of the social effectiveness construct space", *Journal of Personality*, Vol. 76 No. 4, pp. 839-874.

Herzberg, P.Y. and Brähler, E. (2006), "Assessing the Big-Five personality domains via short forms", *European Journal of Psychological Assessment*, Vol. 22 No. 3, pp. 139-148.

Hirevue, Inc (2018), *HireVue for Candidates (4.0.5)*, [Mobile application software], available at: https://itunes.apple.com/gb/app/hirevue-for-candidates/id492573020?mt=8.

Hogan, R., Chamorro-Premuzic, T. and Kaiser, R.B. (2013), "Employability and career success: bridging the gap between theory and reality", *Industrial and Organizational Psychology*, Vol. 6 No. 1, pp. 3-16.

Jimison, H., Pavel, M., McKanna, J. and Pavel, J. (2004), "Unobtrusive monitoring of computer interactions to detect cognitive status in elders", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 8 No. 3, pp. 248-252.

John, O.P. and Srivastava, S. (1999), "The big five trait taxonomy", in Pervin, L.A. and John, O.P. (Eds), *Handbook of Personality: Theory and Research*, 2nd ed., Guilford, New York, NY, pp. 102-138.

Joseph, D.L. and Newman, D.A. (2010), "Emotional intelligence: an integrative meta-analysis and cascading model", *Journal of Applied Psychology*, Vol. 95 No. 1, pp. 54-78.

Kaplan, S., Cortina, J., Ruark, G., LaPort, K. and Nicolaides, V. (2014), "The role of organizational leaders in employee emotion management: a theoretical model", *The Leadership Quarterly*, Vol. 25 No. 3, pp. 563-580.

Kern, M.L., Eichstaedt, J.C., Schwartz, H.A., Dziurzynski, L., Ungar, L.H., Stillwell, D.J., Kosinski, M., Ramones, S.M. and Seligman, M.E. (2014), "The online social self: an open vocabulary approach to personality", *Assessment*, Vol. 21, pp. 158-169.

Loper, E. and Bird, S. (2002), "NLTK: the natural language toolkit", *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Vol. 1, Association for Computational Linguistics, pp. 63-70.

Lumsden, J., Skinner, A., Woods, A.T., Lawrence, N.S. and Munafò, M. (2016), "The effects of gamelike features and test location on cognitive test performance and participant enjoyment", *PeerJ*, Vol. 4, e2184.

MacCann, C. and Roberts, R.D. (2008), "New paradigms for assessing emotional intelligence: theory and data", *Emotion*, Vol. 8 No. 4, pp. 540-551.

McPherson, J. and Burns, N.R. (2008), "Assessing the validity of computer-game-like tests of processing speed and working memory", *Behavior Research Methods*, Vol. 40 No. 4, pp. 969-981.

Mislevy, R.J., Steinberg, L.S. and Almond, R.G. (2003), "Focus article: on the structure of educational assessments", *Measurement: Interdisciplinary Research and Perspectives*, Vol. 1 No. 1, pp. 3-62.

Motowidlo, S.J., Dunnette, M.D. and Carter, G.W. (1990), "An alternative selection procedure: the low-fidelity simulation", *Journal of Applied Psychology*, Vol. 75 No. 6, pp. 640-647.

Mount, M.K., Barrick, M.R., Scullen, S.M. and Rounds, J. (2005), "Higher-order dimensions of the big five personality traits and the big six vocational interest types", *Personnel Psychology*, Vol. 58 No. 2, pp. 447-478.

Newman, D.A. and Lyon, J.S. (2009), "Recruitment efforts to reduce adverse impact: targeted recruiting for personality, cognitive ability, and diversity", *Journal of Applied Psychology*, Vol. 94 No. 2, pp. 298-317.

Ng, T.W., Eby, L.T., Sorensen, K.L. and Feldman, D.C. (2005), "Predictors of objective and subjective career success: a meta-analysis", *Personnel Psychology*, Vol. 58 No. 2, pp. 367-408.

Nguyen, L.S. and Gatica-Perez, D. (2016), "Hirability in the wild: analysis of online conversational video resumes", *IEEE Transactions on Multimedia*, Vol. 18 No. 7, pp. 1422-1437.

O'Brien, H.L. and Toms, E.G. (2010), "The development and evaluation of a survey to measure user engagement", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 1, pp. 50-69.

Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H. and Seligman, M.E. (2015), "Automatic personality assessment through social media language", *Journal of Personality and Social Psychology*, Vol. 108 No. 6, pp. 934-952.

Pennebaker, J.W. and King, L.A. (1999), "Linguistic styles: language use as an individual difference", *Journal of Personality and Social Psychology*, Vol. 77 No. 6, pp. 1296-1312.

Petrides, K.V. and Furnham, A. (2006), "The role of trait emotional intelligence in a gender-specific model of organizational variables 1", *Journal of Applied Social Psychology*, Vol. 36 No. 2, pp. 552-569.

Podsakoff, P.M. and MacKenzie, S.B. (1989), *A Second Generation Measure of Organizational Citizenship Behavior*, Indiana University, Bloomington, Unpublished manuscript.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y. and Podsakoff, N.P. (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88 No. 5, pp. 879-903.

Quiroga, M.Á., Escorial, S., Román, F.J., Morillo, D., Jarabo, A., Privado, J., Hernández, M., Gallego, B. and Colom, R. (2015), "Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can!", *Intelligence*, Vol. 53, pp. 1-7.

Quiroga, M.A., Román, F.J., De La Fuente, J., Privado, J. and Colom, R. (2016), "The measurement of intelligence in the XXI century using video games", *Spanish Journal of Psychology*, Vol. 19, pp. 1-13.

Robinson, D. and Bellotti, V. (2013), "A preliminary taxonomy of gamification elements for varying anticipated commitment", *ACM CHI 2013 Workshop on Designing Gamification: Creating Gameful and Playful Experience*.

Rubin, R.S., Munz, D.C. and Bommer, W.H. (2005), "Leading from within: the effects of emotion recognition and personality on transformational leadership behavior", *Academy of Management Journal*, Vol. 48 No. 5, pp. 845-858.

Rupasinghe, A.T., Gunawardena, N.L., Shujan, S. and Atukorale, D.A.S. (2016), "Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis", *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, IEEE, pp. 288-295.

Schlegel, K. and Scherer, K.R. (2016), "Introducing a short version of the geneva emotion recognition test (GERT-S): psychometric properties and construct validation", *Behavior Research Methods*, Vol. 48 No. 4, pp. 1383-1392.

Spector, P.E. (2006), "Method variance in organizational research: truth or urban legend?", *Organizational Research Methods*, Vol. 9 No. 2, pp. 221-232.

Stone, P.J., Dunphy, D.C. and Smith, M.S. (1966), *The General Inquirer: A Computer Approach to Content Analysis*, M.I.T. Press, Oxford.

Tlili, A., Essalmi, F., Jemni, M. and Chen, N.S. (2016), "Role of personality in computer based learning", *Computers in Human Behavior*, Vol. 64, pp. 805-813.

Uniform Guidelines on Employee Selection Procedures (1978), *Federal Register*, Vol. 43, pp. 38290-3831.

Ventura, M. and Shute, V. (2013), "The validity of a game-based assessment of persistence", *Computers in Human Behavior*, Vol. 29 No. 6, pp. 2568-2572.

Verhaegh, J., Fontijn, W.F., Aarts, E.H. and Resing, W.C. (2013), "In-game assessment and training of nonverbal cognitive skills using tag tiles", *Personal and Ubiquitous Computing*, Vol. 17 No. 8, pp. 1637-1646.

Weekley, J.A., Ployhart, R.E. and Holtz, B.C. (2006), "On the development of situational judgment tests: issues in item development, scaling, and scoring", in Weekley, J.A. and Ployhart, R.E. (Eds), *Situational Judgment Tests: Theory, Measurement, and Application*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 157-182.

Weekley, J.A., Hawkes, B., Guenole, N. and Ployhart, R.E. (2015), "Low-fidelity simulations", *Annual Review of Organizational Psychology and Organizational Behavior*, Vol. 2 No. 1, pp. 295-322.

Wood, R.T., Griffiths, M.D., Chappell, D. and Davies, M.N. (2004), "The structural characteristics of video games: a psycho-structural analysis", *CyberPsychology and Behavior*, Vol. 7 No. 1, pp. 1-10.

Further reading

Borkenau, P., Mauer, N., Riemann, R., Spinath, F.M. and Angleitner, A. (2004), "Thin slices of behavior as cues of personality and intelligence", *Journal of Personality and Social Psychology*, Vol. 86 No. 4, pp. 599-614.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R.A. and Hogan, R. (2016), "New talent signals: shiny new objects or a brave new world?", *Industrial and Organizational Psychology*, Vol. 9 No. 3, pp. 621-640.

Hogan, J.W., Roy, J. and Korkontzelou, C. (2004), "Handling drop-out in longitudinal studies", *Statistics in Medicine*, Vol. 23 No. 9, pp. 1455-1497.

Hoppe, S., Loetscher, T., Morey, S.A. and Bulling, A. (2018), "Eye movements during everyday behavior predict personality traits", *Frontiers in Human Neuroscience*, Vol. 12 No. 105, pp. 1-8.

Kosinski, M. and Behrend, T. (2017), "Editorial overview: big data in the behavioral sciences", *Current Opinion in Behavioral Sciences*, Vol. 18, pp. Iv-Vi.

Mischel, W. (1996), *Personality and Assessment*, Wiley, New York.

Corresponding author
Franziska Leutner can be contacted at: f.leutner@ucl.ac.uk