Game Based Assessments
Are they really the future?

12 May, 2019

Prepared by:

STEN10
Ben Williams
Business Psychologist

Kings Head House, 15 London End,
Beaconsfield HP9 2HN

+44 (0)1494 412 861
+44 (0)7939 156 708
ben@sten10.com/amy@sten10.com
Who I am

- Chartered Psychologist
- Managing Director of Sten10 Ltd. / Chair of ABP
- Publisher-independent
- (Was an) avid gamer
Agenda

LEVEL 1 - Introduction to Game Based Assessment

• Key parameters of a GBA
• Four types of GBA

LEVEL 2 - Evidence Base

• Types of Evidence
• Reliability / Validity / Adverse impact / Engagement

LEVEL 3 - Conclusions
Level 1
Introduction to GBA
Key Parameters of a GBA

• Nature: Gamification vs. Game Based Assessment
• Type: Custom-built vs. pre-existing vs. gamified traditional vs. VR
• Measures: performance, behavioural choice and / or ‘meta-data’ to assess:

  • Abilities:
    • Cognitive processing speed
    • Attention span
    • Working memory
    • V, N, A reasoning

  • Personality traits:
    • Persistence
    • Risk propensity
    • Emotional Intelligence

  • ‘Role-Fit’ – A.I. % match
Gamification in Recruitment

If you have what it takes to be an engineer in the Air Force call the number below.

\[ 9^{\frac{1}{2}} + 2^{13} + 14^3 + 5^7 + (\sinh x - \sum_{k=4}^{\infty} \frac{x^{2k+1}}{(2k+1)!}) = \]

AIR FORCE ACCOMPLISHED
Types of GBA

1. Custom-Built GBA’s
Arctic Shores
Knack
HireVue (formerly MindX)
Quest
Revelian
Pymetrics

Play fun neuroscience games

You will be presented with a series of balloons. You can click on the "Pump" button to increase the size of the balloon, receiving 5 cents in a temporary bank for each pump. At any point, you can stop pumping the balloon and click on the "Collect" button to transfer the accumulated money from your temporary bank to your permanent bank labeled "Total Earned". You will then start on the next balloon.
Types of GBA

2. Pre-existing
‘Pre-Existing’ Games

![Game Images]

Table 2: IPIP Correlations Total Sample

<table>
<thead>
<tr>
<th>Statement</th>
<th>Game Variable</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>I love children</td>
<td>KillsPerSecond</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>HitsPerSecond</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td>HeadShotsPerSecond</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>UnlockScorePerSecond</td>
<td>-0.14</td>
</tr>
</tbody>
</table>
Types of GBA

3. Tailored Traditional
Gamified Assessments (Not ‘Games’?)
Types of GBA

4. Virtual Worlds, Virtual Reality
Level 2
Evidence Base
The Challenges

The challenges of establishing psychometric properties:

- **A New Market** - GBA Test publishers are quite young meaning evidence of predictive power is limited by necessity.

- **Generalisations** about the evidence base are difficult compared to ‘traditional’ psychometrics due to the variety of design.

- **Objectivity** - Investigating GBAs objectively is problematic as commercial IP is tied up in the algorithms used. Also, most research being funded and facilitated by the publishers themselves.

- **Common method variance** – using GBAs changes the way constructs are measured (construct validity).

- **Complex** – not only raw score but thousands of meta-data points are measured.
Reliability and Validity

Reliable not Valid
Precise not Accurate

Not Reliable but Valid
Not Precise but Accurate

Not Reliable and not Valid
Not Precise and not Accurate

Reliable and Valid
Precise and Accurate
Reliability

- Consistency over time
- Internal consistency
- Sources of measurement error
Consistency
(All from test GBA test publishers)

Internal consistency
• 0.6 – 0.9 (n = 6,000)
• 0.51 – 0.96 (n = < 100)
• 0.84 (n = 500)
(n.b. typical vs maximum ideal values)

Consistency over time
• 0.57 – 0.82 test-retest

Parallel form
• 0.44 – 0.79 for subtests
• >0.9 for app version vs laptop version
Sources of Measurement Error

Length of assessment

Distortion
- GBA assesses behaviour directly, not through self report: more resistant to distortion? (Landers, 2015) Scores modified on self-report PQs for extraversion and agreeableness, but unable to in a GBA (Montefiori, 2016)

Irrelevant Factors
- Potential reliance on irrelevant factors such as hand-eye co-ordination. Highly interactive games may create unnecessary cognitive load. (Zapata-Rivera & Bauer, 2012)
Validity

Face / Engagement

Criterion

Construct
Face Validity / Engagement
- Selected studies
Face Validity / Engagement
- Selected studies

Intention to accept job offer

Animated characters = positive attitude towards hiring company, stronger intention to accept a job offer (e.g. Motowidlo et al., 1990; Richman-Hirsch et al., 2000; Bruk-Lee et al., 2012)
Face Validity / Engagement - Selected studies

Enjoyment

+ve
• A test publisher found 94.3% of ppts (N = 1747) reported enjoyed playing a GBA

• Another test publisher found 90% of candidates feel that GBAs are the same or better than traditional assessments

-ve
• Candidates value ease of use and usability more than enjoyment. Most candidates would prefer job relevant test (e.g. work sample) over fun games. (Laumer et al. 2012)

Enjoyment mediated by individual differences:
• Oostrom et al (2011): candidate perceptions positively correlated with personality traits of Openness and Agreeableness
Face Validity / Engagement
- Selected studies

Gaming Expertise
A test publisher (2014) found 80% ‘enjoyed’ gamified learning tool BUT ‘hard-core gamers’ disengaged. Millennials most likely to logon, but quickest to drop out. Also found males more likely to engage with the game.

Technology
Preuss (2017) found that 60% of candidates prefer using Gamified SJT over a traditional SJT.

However, technological difficulties for some candidates resulted in lower perception of gamified SJT.
Face Validity / Engagement
- Selected studies

**Perception of ‘fairness’**

- A quarter of candidates believe completing an assessment on a mobile device would provide a ‘fair’ testing experience (Fursman & Tuzinski, 2015)

- Landers (2017) found test takers consider GBA ‘fairer’ than general cognitive ability tests

- Different publisher’s manual showed 40% saw it as more fair, 40% less fair

**Anxiety**

- 74% (n=200) felt less anxiety for GBA, 89% enjoyed the selection process, 81% felt more excited about the prospect of working for the firm (test publisher research)

- Geimer et al (2015) found Candidates experienced higher levels of anxiety when feedback is given in game
Construct Validity
-Selected research

Big Five Personality

Van Lankveld (2011) 275 individual metrics in ‘Neverwinter Nights’ and found 1,375 correlations with Big 5 traits. However, some of these could be spurious. (n.b. n=44)

Short et al (2017) found no links to Big 5 using World of Warcraft. Fairly consistent support for preference for virtual teamwork and technology readiness.
Construct Validity

-Selected research

**Working Memory/Fluid Intelligence**

Baniqued et al (2013) found performance on games that required working memory and reasoning significantly correlated with performance on working memory and fluid intelligence tasks.
Construct Validity

-Selected research

Correlations with established measures of same constructs:

Test provider 1*: 0.24 to 0.44

Test provider 2*: 0.2 to 0.26

Test provider 3*: 0.3 to 0.54
Construct Validity cont.

Figure 1 below for results. Personality constructs were found to be partly similar. There were varying results for cognitive abilities (divergent – different, convergent – similar).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>conscientious</td>
<td>.03</td>
<td>.04</td>
<td>-.10</td>
<td>.01</td>
<td>-.18*</td>
<td>.05</td>
<td>.04</td>
<td>-.07</td>
<td>-.05</td>
</tr>
<tr>
<td>energetic</td>
<td>-.01</td>
<td>-.02</td>
<td>-.19*</td>
<td>.16</td>
<td>.02</td>
<td>.04</td>
<td>-.15</td>
<td>-.12</td>
<td>-.04</td>
</tr>
<tr>
<td>fx</td>
<td>.28**</td>
<td>.15</td>
<td>.14</td>
<td>.08</td>
<td>.07</td>
<td>.05</td>
<td>.09</td>
<td>.21*</td>
<td>.23**</td>
</tr>
<tr>
<td>lct</td>
<td>.42**</td>
<td>.28*</td>
<td>.07</td>
<td>.03</td>
<td>-.08</td>
<td>.08</td>
<td>.22**</td>
<td>.25**</td>
<td>.30**</td>
</tr>
<tr>
<td>sparks_fluency</td>
<td>.04</td>
<td>.03</td>
<td>.02</td>
<td>-.18*</td>
<td>.07</td>
<td>.10</td>
<td>.02</td>
<td>.01</td>
<td>-.03</td>
</tr>
<tr>
<td>sparks_flexibility</td>
<td>.09</td>
<td>.04</td>
<td>-.01</td>
<td>-.11</td>
<td>.10</td>
<td>.06</td>
<td>.02</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>sparks_originality</td>
<td>.05</td>
<td>-.09</td>
<td>.05</td>
<td>.15</td>
<td>.04</td>
<td>-.12</td>
<td>.01</td>
<td>.05</td>
<td>.10</td>
</tr>
</tbody>
</table>

Note. ** p < .01. * p < .05. N = 149. Expected and significant correlations shaded in green. Not expected but significant correlations shaded in grey. Expected but not significant correlations shaded in yellow.
Landers (2017) aimed to validate a cognitive ability GBA through comparison with a traditional test battery and found:

- The game predicted ‘grade point average’ outcome measure better than 15 separate Spearman’s g measures (Spearman’s g provided no ‘unique’ prediction).

Other case studies from GBA publishers:

- Prediction of selection success for air traffic controllers (2017). Significant difference between successful and unsuccessful applicants’ mean scores on GBA (p > 0.001)
- Overall AC pass rate in 2016 = 24% Now in 2017 = 40% (60% for some Business Areas)
- Hi / low manager rating versus GBA performance: 0.019 sig.
- Global Tech Co.: Quality of Hire survey: .162 and .220
- Prediction of competency scores in AC for sales roles ranged between .135 to .347.
- Prediction of competency performance at a retail company – Multiple R .539
- High performance contact centre agents made 66% more bookings in value than the lowest performers, 10% more calls in a month on average
Adverse Impact

Case study 1 (2016): 5,000+ participants, no adverse impact for:
Age, Gender, Ethnicity, Disability (after WM adjustment for dyslexia), Gaming experience, Handedness, Screen size

Case Study 2 (2017): 1,054 candidates, no adverse impact for:
Age, Gender, Race

Case Study 3 (2016): 155 participants, no gender differences on:
“cognitive style”, “information processing competencies”

Case Study 4 (2018): No gender differences on personality responses

BUT, SHOULD there be group differences to reflect what we know about human nature?
Level 3
Conclusions
Summary

‘The practice of gamification has far outpaced researcher understanding of its processes and methods’ (Landers et al, 2015).

- Relative lack of peer-reviewed, academic (non-vendor-led) research.

- Of the evidence there is, reliability (internal consistency and over time), engagement and adverse impact data looks promising. Construct validity and parallel form reliability is positive, with caveats. Validity on later-assessment stages and on the job looks good, although more academic-led research would be beneficial.
Thank you!

Any Questions?