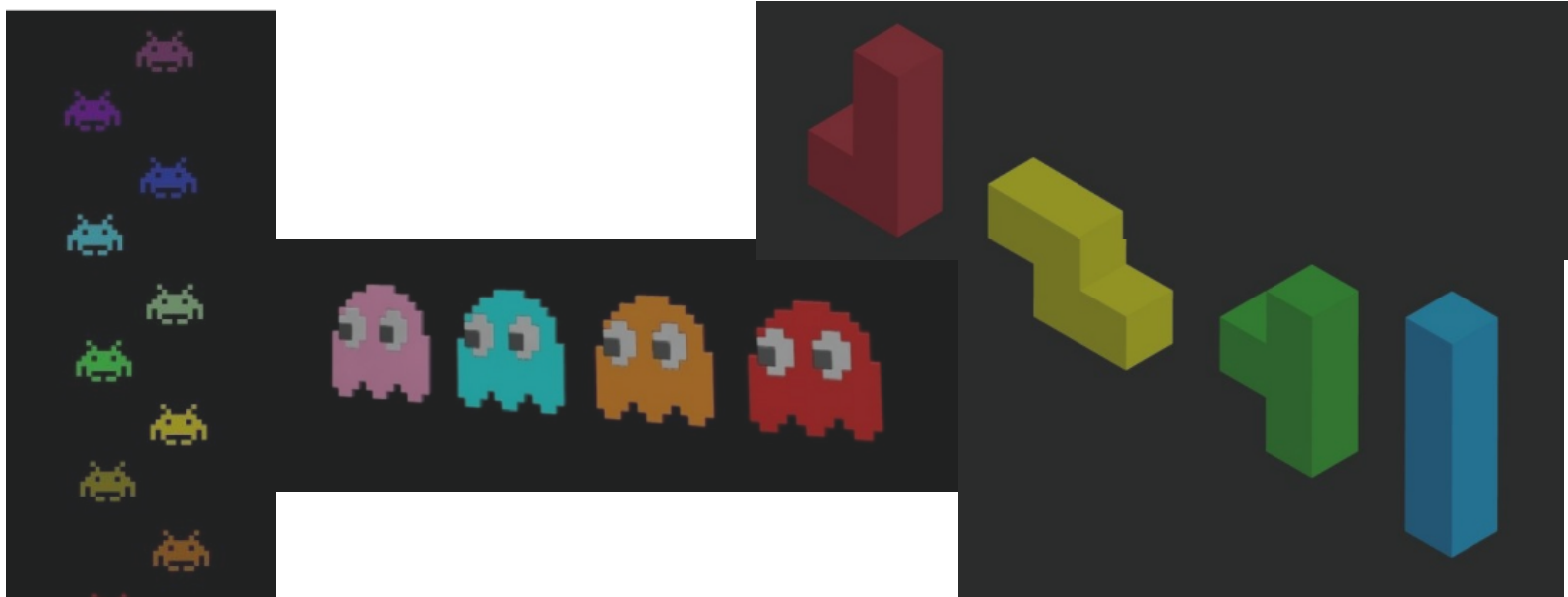


# Assessing Game-Based Assessments: Prospects Meet Principles

---

Fred Oswald

Department of Psychological Sciences - Rice University



## Medical video game could help astronauts diagnose and treat problems on way to Mars

Alex Stuckey | Aug. 16, 2019 | Updated: Aug. 19, 2019 10:19 a.m.



FILE -- An image provided by NASA shows an artist's rendering of the Opportunity rover on the surface of Mars, which landed in January 2004, was designed for 90 days of exploration but remained functional for nearly 15 years. (NASA via The New York Times) -- EDITORIAL USE ONLY --

1,213 views | Aug 21, 2019, 04:22am

## Five Companies Using Virtual Reality To Improve The Lives Of Senior Citizens



Sol Rogers Contributor @ Consumer Tech

Virtual reality is emerging as a useful tool to bring about positive change for many, including the elderly. From reducing loneliness to transporting the infirm to far-flung places, with VR is enhancing the lives of senior citizens across the globe.



Virtual reality has the power to give the elderly the freedom they never thought possible. MYND VR

## Depressed and Anxious? These Video Games Want to Help



In the video game Sea of Solitude, the main character, a young woman named Kay, navigates a partly submerged city and fights to overcome loneliness. Electronic Arts

- **selection (stable IDs)**
  - cognitive ability
  - personality
  - motivation/interest
  - stable behaviors (e.g., teamwork)
  - changing behaviors (e.g., adaptability)
- **training/development (changing IDs)**
  - knowledge (e.g., learn R/Python)
  - technical skills (e.g., code R/Python)
  - interpersonal skills (e.g., best/worst interactions)
  - intrapersonal skills (e.g., mental and physical health, STEM interests)

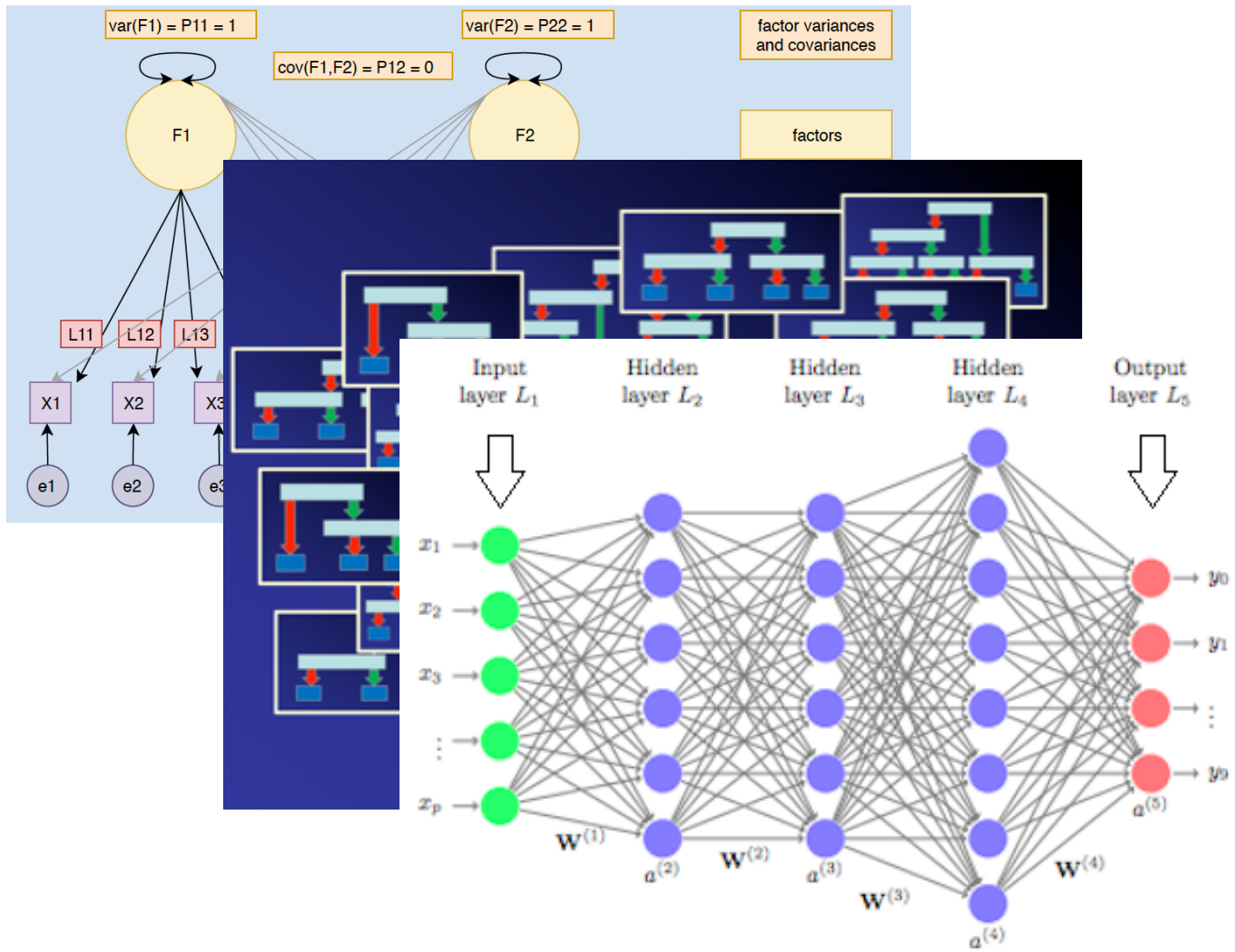


## Why games, why GBA?

Compared with traditional approaches:

- **Larger-scale** assessment  
(e.g., rich and diverse talent pools)
- **Safe environments for 'unsafe' mistakes**  
(e.g., saying the wrong things in conversation, medical errors)
- **Overlearning** in rare environments  
(e.g., nuclear power plant emergencies)
- **Rich** forms of interaction (...thus, rich/multilevel constructs)  
(e.g., gamification/VR, video interviews, biometrics, social networks, adaptive testing)
- **Rapid/automated** decisions  
(e.g., selection, acquire talent before others do; learning, provide feedback in real time)
- **Enhanced** prediction  
(e.g., ML algorithms applied to a massive number of features/predictors)

- **Engagement = increase the volume of people who decide to play**
  - e.g., games + neuro  
= fun/engagement + sophistication/science
  - ...self-selection effect?
- **Engagement = increase persistence of players within a game**
  - heighten motivation to perform:  
escape, esthetic, interests, challenge, social connection
  - get more (big) data



Big data (game-based, otherwise)  
is also facing a **replication crisis**:

- **ML / deep learning methods** have been

**Are We Really Making Much Progress? A Worrying Analysis of  
Recent Neural Recommendation Approaches**

Maurizio Ferrari Dacrema  
Politecnico di Milano, Italy  
maurizio.ferrari@polimi.it

Paolo Cremonesi  
Politecnico di Milano, Italy  
paolo.cremonesi@polimi.it

Dietmar Jannach  
University of Klagenfurt, Austria  
dietmar.jannach@aau.at

<https://arxiv.org/pdf/1907.06902.pdf>

**AAAS: Machine learning 'causing science  
crisis'**

By Pallab Ghosh  
Science correspondent, BBC News, Washington

© 16 February 2019

    Share

<https://www.bbc.com/news/science-environment-47267081>

PUBLIC RELEASE: 15-FEB-2019

Can we trust scientific discoveries made  
using machine learning?

*Rice U. expert: Key is creating ML systems that question their own predictions*

RICE UNIVERSITY

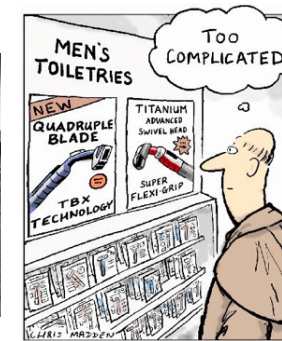
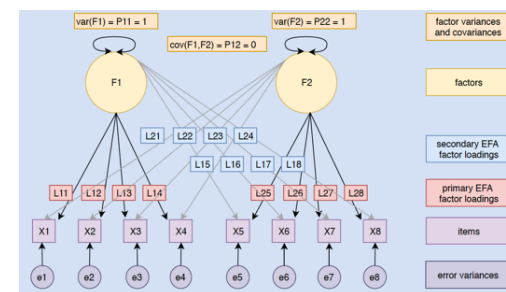
[https://eurekaalert.org/pub\\_releases/2019-02/ru-cwt021119.php](https://eurekaalert.org/pub_releases/2019-02/ru-cwt021119.php)

## Computational Psychometrics:

Measurement, Modeling, and Meaning in the Big Data Era

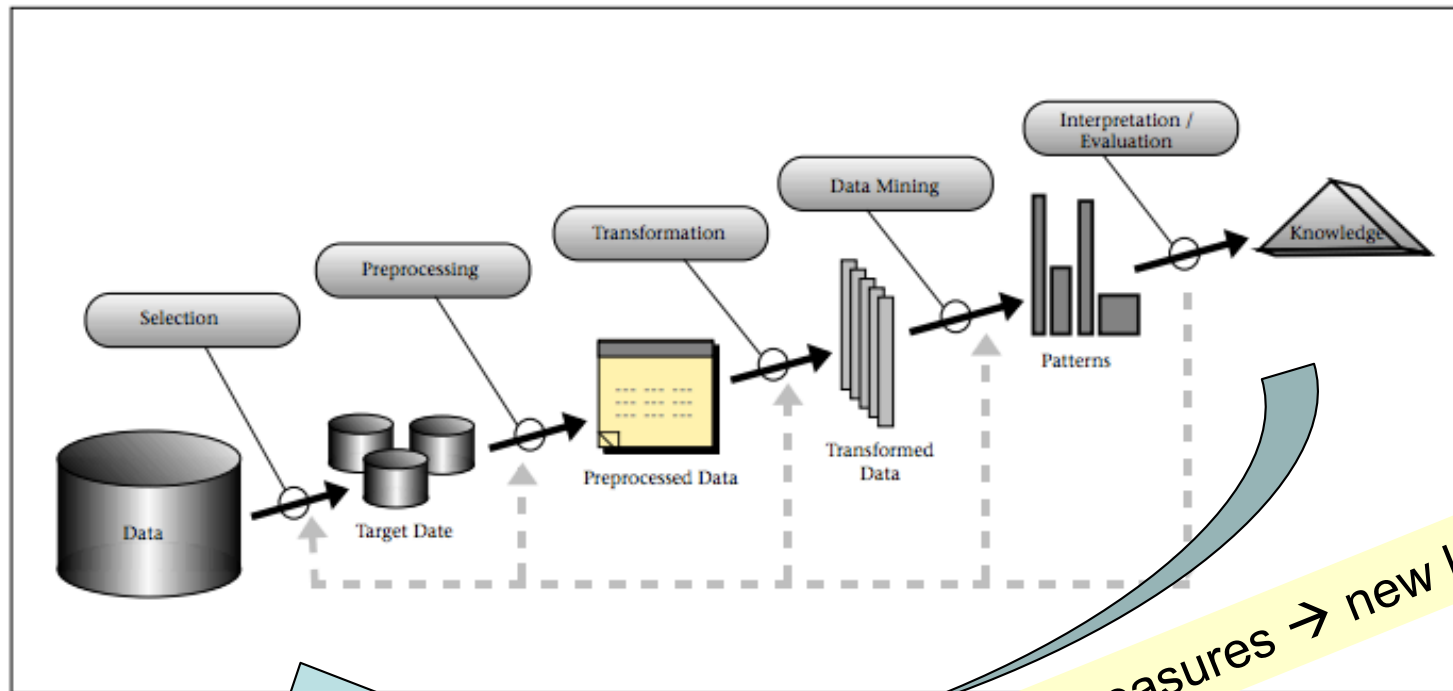
[Rice University + Army Research Institute  
for the Behavioral and Social Sciences]

- **reliability beyond alpha and CFA**  
given large-scale ‘messy’ data  
(missing, text-based, game-based, temporal)
- **explore multiple methods for establishing reliability and construct relevance** (vs. **algorithmic bias**)  
network psychometrics, dynamic modeling, merging *incidental* data (bottom-up, unstructured/activities) with *intentional* data (top-down, traditional/items)
  - exploratory/inductive surprises = apply big data algorithms to the “data firehose”
  - cross-validated EFA/CFA/SEM vs. predictive models (random forest, SVM, elastic net...)
  - how do we know when we need complexities ...vs. when we don’t (Occam)



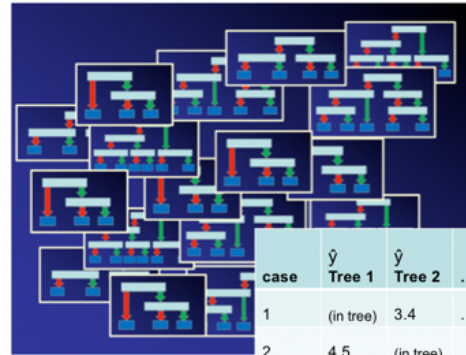
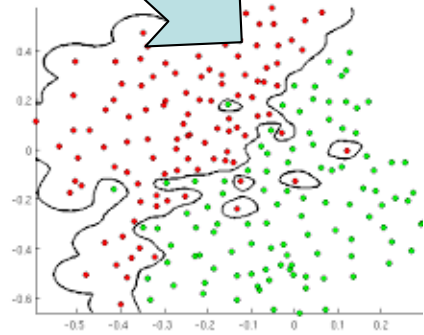
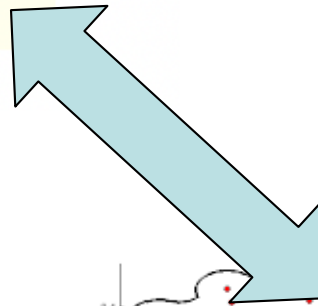
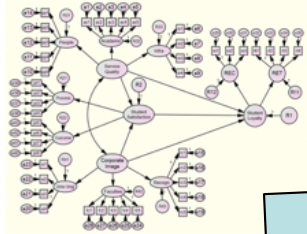
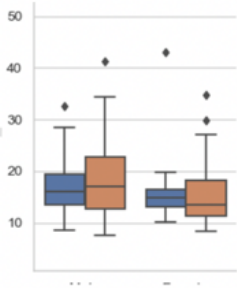
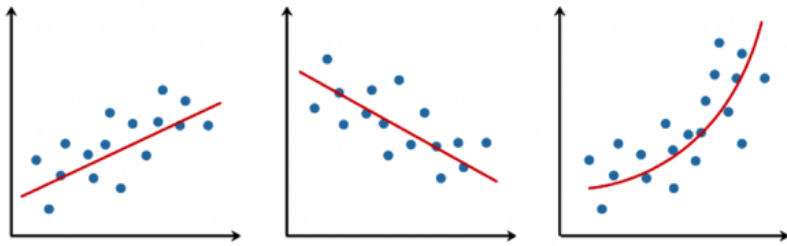


Useful 'signals' in data discovered through predictive modeling could be amplified by developing measures that collect more data (given enough development time, testing time, \$...).



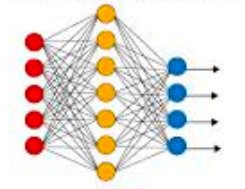
(Fayyad et al., 1996 +  
exciting arrow by Oswald)

(knowledge → new measures → new knowledge)

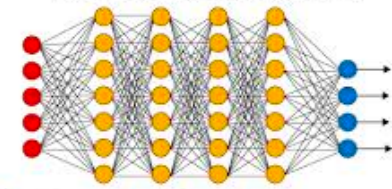


case	$\hat{y}$ Tree 1	$\hat{y}$ Tree 2	...	$\hat{y}$ Tree k	Average Predicted $\hat{y}$
1	(in tree)	3.4	...	2.1	3.554
2	4.5	(in tree)	...	(in tree)	4.312
...	...	...	...	...	...
N	2.6	(in tree)	...	2.4	2.561

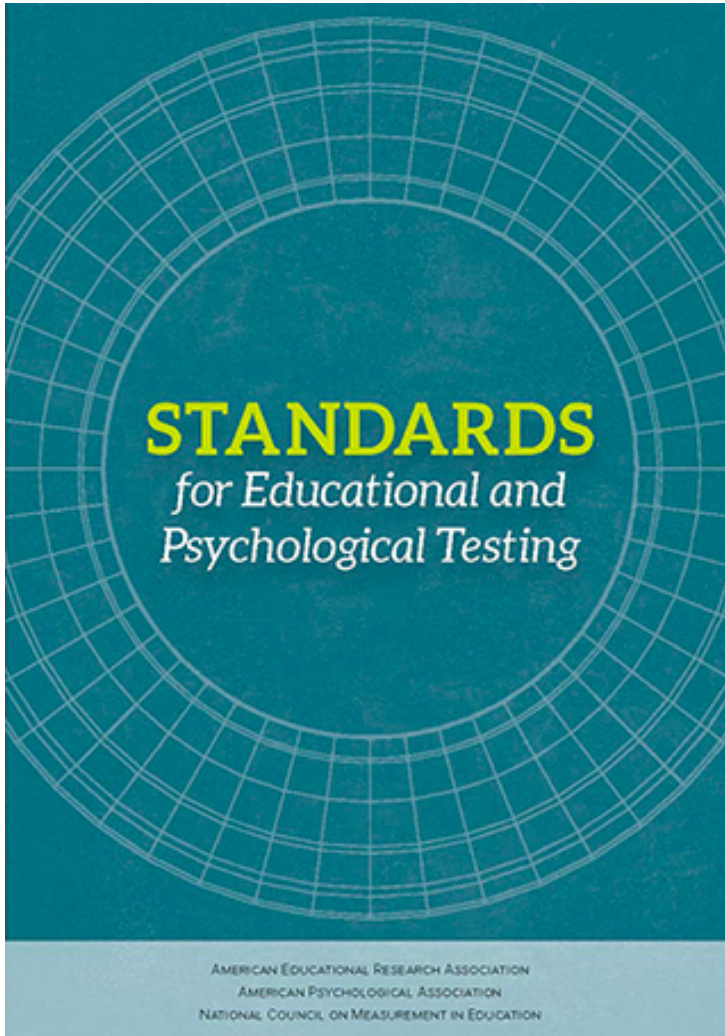
Simple Neural Network



Deep Learning Neural Network



● Input Layer    ● Hidden Layer    ● Output Layer



## Overview of Organization and Content

### Part I: Foundations

1. Validity.
2. Reliability/precision and errors of measurement.
3. Fairness in testing.

### Part II: Operations

1. Test design and development.
2. Scores, scales, norms, score linking and cut scores.
3. Test administration, scoring, reporting and interpretation.
4. Supporting documentation for tests.
5. The rights and responsibilities of test takers.
6. The rights and responsibilities of test users.

### Part III: Testing Applications

1. Psychological testing and assessment.
2. Workplace testing and credentialing.
3. Educational testing and assessment.
4. Uses of tests for program evaluation, policy studies and accountability.

## Principles for the Validation and Use of Personnel Selection Procedures

FIFTH EDITION  
AUGUST 2018



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

### CONTENTS

Foreword .....	vi	Appropriate analysis .....	31
Introduction .....	1	Differential prediction .....	31
Statement of Purpose .....	1	Combining selection procedures into a selection system .....	31
Principles as guidance .....	2	Multiple hurdles versus compensatory models .....	32
Selection Procedures Defined .....	2	Cutoff scores versus rank orders .....	32
<b>Overview of the Validation Process .....</b>	<b>4</b>	Bands .....	32
Sources of Evidence .....	5	Norms .....	33
Evidence based on the relationship between scores on predictors and other variables .....	5	Communicating the effectiveness of selection procedures .....	33
Content-related evidence .....	5	Appropriate Use of Selection Procedures .....	33
Evidence based on the internal structure of the test .....	5	Combining selection procedures .....	33
Evidence based on response processes .....	5	Using selection procedures for other purposes .....	33
Evidence for validity and consequences of personnel decisions .....	6	Recommendations .....	33
Planning the Validation Effort .....	6	Technical Validation Report .....	33
Existing evidence .....	6	Identifying information .....	34
Proposed uses .....	7	Statement of purpose .....	34
Requirements of sound inference .....	7	Analysis of work .....	34
Feasibility .....	7	Search for alternative selection procedures .....	34
Analysis of Work .....	7	Identification or development of selection procedures .....	34
Purposes for conducting an analysis of work .....	7	Establishing validity .....	34
Level of detail .....	7	Research sample .....	34
Results .....	7	Scoring and transformation of raw scores .....	34
<b>Sources of Validity Evidence .....</b>	<b>9</b>	Normative information .....	35
Evidence of Validity Based on Relationships With Measures of Other Variables .....	9	Recommendations .....	35
Criterion-Related Evidence of Validity .....	10	Caution regarding interpretations .....	35
Feasibility of a criterion-related validation study .....	10	Technology-enabled selection procedures .....	35
Design and conduct of criterion-related studies .....	10	References .....	35
Criterion development .....	11	Administration Information .....	35
Choice of predictor .....	12	Introduction and overview .....	35
Choice of participants .....	13	Contact information .....	36
Data analysis for criterion-related validity .....	13	Selection procedures .....	36
Evidence for Validity Based on Content .....	15	Applicability .....	36
Feasibility of a content-based validation study .....	15	Administration responsibilities .....	36
Design and conduct of content-based strategies .....	16	Information provided to candidates .....	36
Defining the content domain .....	16	Guidelines for administration of selection procedures .....	36
Developing or choosing the selection procedure .....	16	Administration environment .....	37
Procedural considerations .....	16	Scoring instructions and interpretation guidelines .....	37
Evaluating content-related evidence .....	17	Test score databases .....	37
Evidence of Validity Based on Internal Structure .....	17	Reporting and using selection procedure scores .....	37
<b>Generalizing Validity Evidence .....</b>	<b>19</b>	Candidate feedback .....	37
Transportability .....	19	Minimum test administration conditions .....	38
Synthetic Validity/Job Component Validity .....	19	Validation Effort and Use of Selection Procedures .....	38
Meta-Analysis .....	20	Individual demands .....	38
<b>Fairness and Bias .....</b>	<b>22</b>	Review of validation and need for updating the validation effort .....	39
Fairness .....	22	Assessing Candidates With Disabilities .....	39
Defining the organization's needs, objectives, and constraints .....	23	Responsibilities of the selection procedure developers, testing professionals, and users related to accommodation .....	39
Communicating the validation plan .....	27	Candidate Linguistic and Cultural Background .....	40
Understanding Work and Worker Requirements .....	27	<b>References .....</b>	<b>41</b>
Strategies for analyzing the work domain and defining worker requirements .....	27	<b>Glossary of Terms .....</b>	<b>46</b>
Considerations in specifying the sampling plan .....	27		
Documentation of the results .....	27		
Selecting Assessment Procedures for the Validation Effort .....	27		
Review of research literature and the organization's objectives .....	27		
Psychometric considerations .....	27		
Scoring considerations .....	27		
Format and medium .....	28		
Acceptability to the candidate .....	28		
Alternate forms .....	28		
Selecting the Validation Strategy .....	28		
Fit to objectives, constraints, and selection procedures .....	29		
Individual assessments .....	29		
Selecting Criterion Measures .....	29		
Performance-oriented criteria .....	29		
Other indices .....	29		
Relevance and psychometric considerations .....	29		
Data Collection .....	30		
Communications .....	30		
Pilot testing .....	30		
Match between data collection and implementation expectations .....	30		
Confidentiality .....	30		
Quality control and security .....	30		
Data Analyses .....	30		
Data accuracy and management .....	30		
Missing data and outliers .....	31		
Descriptive statistics .....	31		

<https://tinyurl.com/siop-standards-5th>

Is an AI game better at hiring  
than...a coin flip? Ask your vendor!



### Coin flip

### AI game

**fun/engaging**

**make quick decisions**

**affordable**

**fair**  
(e.g., no adverse impact)

**reliable**  
(e.g., similar score  
retaken 1 week later)

**valid**  
(e.g., predicts employee  
performance)

? 😊

✓

✓

✓

✗

✗

✓

✓

?

?

?

?

## 1. Keep going beyond “<game> works!”

- company / games / algorithms ≠  
constructs → measures → decisions → outcomes

## 2. Improve the conceptualization and measurement of goals and criteria

- what is a “successful” employee or student  
(teamwork, taskwork, engagement, low turnover)
- how are multiple criteria related? how does prediction work?
- what about GBA predicting intervention success + criteria:  
i.e., GBA → training → outcome criteria

3. **cultivate community**: develop an extended and engaged network of expertise and collaboration around GBA (project-driven, profession-driven, listserv driven, etc.) – mentor others (world is small)
4. **develop collaborative strategies and goals**: yes even between vendors; communication through this community (advisory board, publication plans, conference presence, etc.)
5. **develop and share innovative research and tools** that could not have happened otherwise

## 6. Work toward GBAs being more transparent, replicable, generalizable

- yes, there are proprietary issues
- yes, there are lawyers
- yes, there need to be profits (no, really)
- but compete on your science as a differentiator
- make headway as a “thought leader” via sharing your findings for the community, for discerning consumers

## 7. Provide clear indices of reliability, validity, and fairness

- whether through traditional methods or non-traditional analogs
- stakeholders will demand and push on improved reliability/validity/fairness data (practice, science, legal...ethics...fronts)



## ORGANIZATION & WORKFORCE LABORATORY (OWL)

*Dr. Fred Oswald, Rice University*

Home

Who We Are

Current Projects

Publications

Resources

Partners

About Fred

Contact

[foswald@rice.edu](mailto:foswald@rice.edu)

<https://workforce.rice.edu>



RICE UNIVERSITY