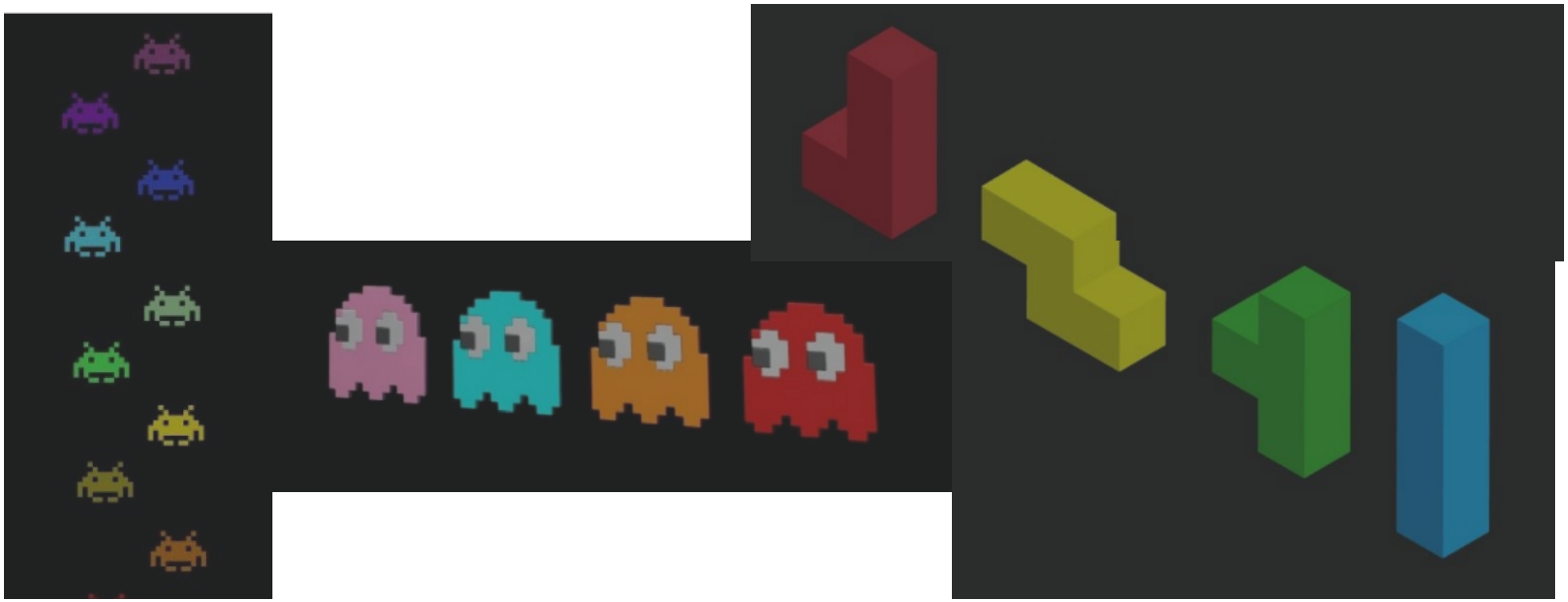
RICE UNIVERSITY

Assessing Game-Based Assessments:

Prospects Meet Principles

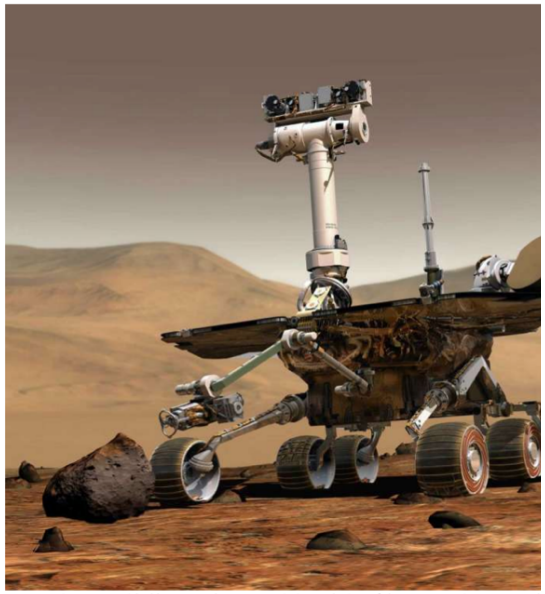
Fred Oswald

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 th which landed in January 2004, was designed for 90 days of exploration but remained functio days. (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, wit
VR is enhancing the lives of senior citizens acr

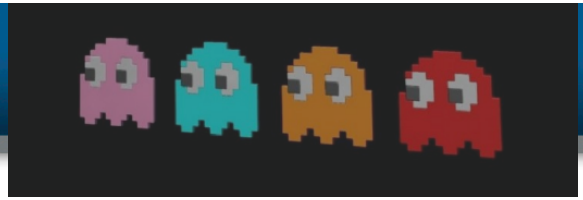
Virtual reality has the power to give the elderly the fre never thought possible. MYND VR

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

purposes of GBA

- **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?

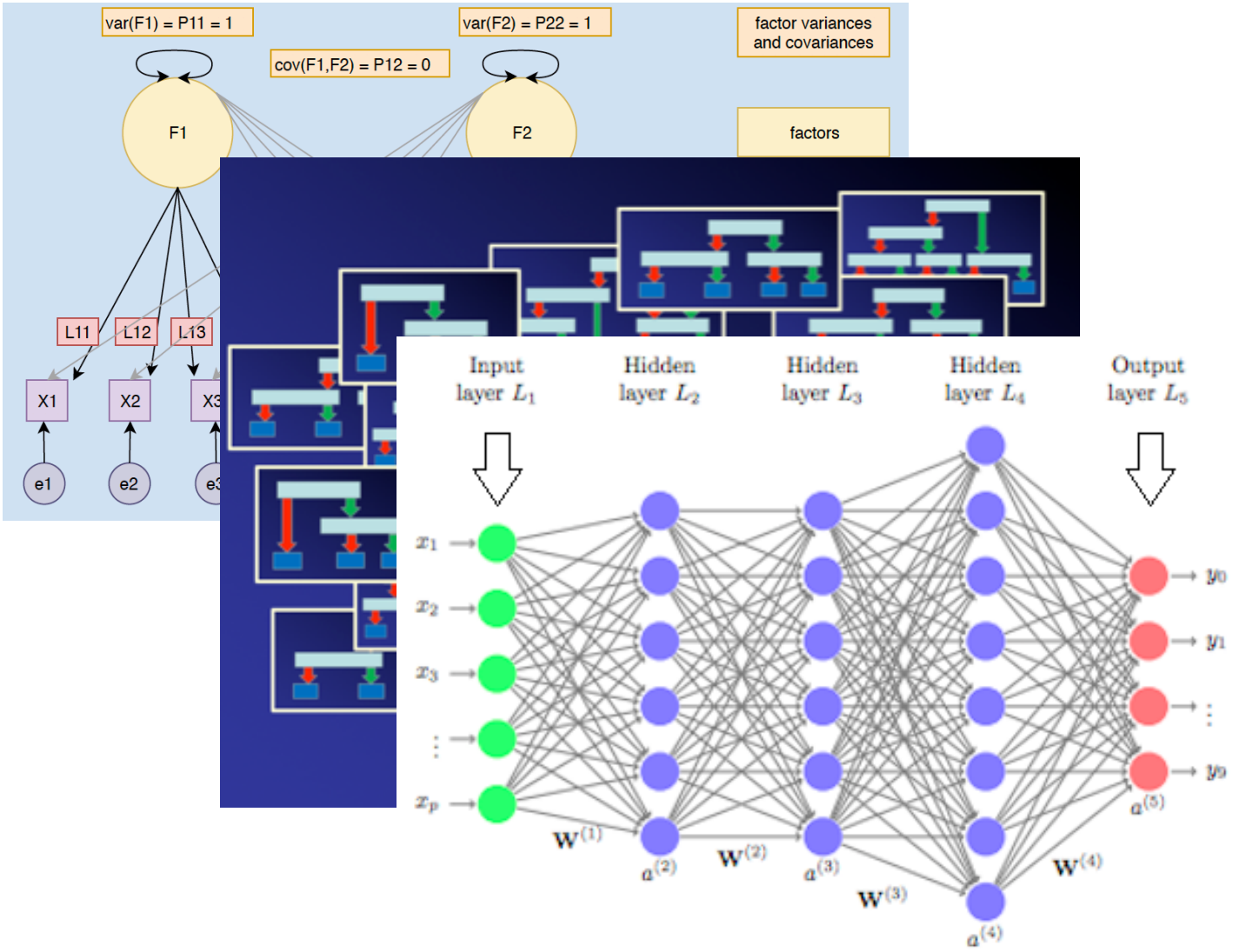
Compared with traditional approaches:

- **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

- **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

models/algorithms



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

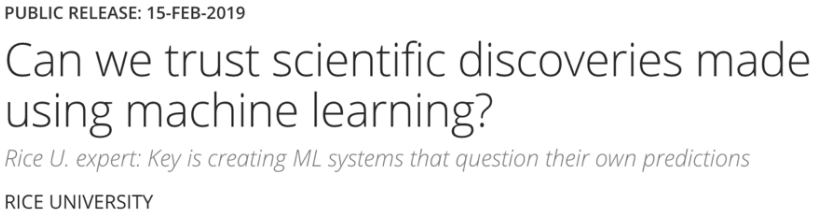
https://arxiv.org/pdf/1907.06902.pdf

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?

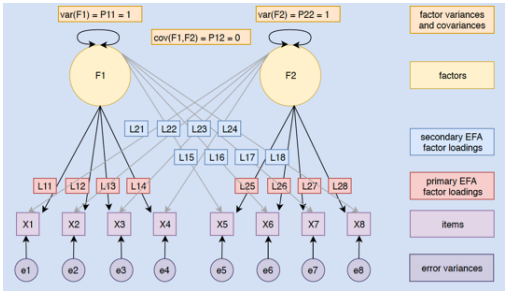
Rice U. expert: Key is creating ML systems that question their own predictions

RICE UNIVERSITY

https://eurekalert.org/pub_releases/2019-02/ru-cwt021119.php

Computational Psychometrics:

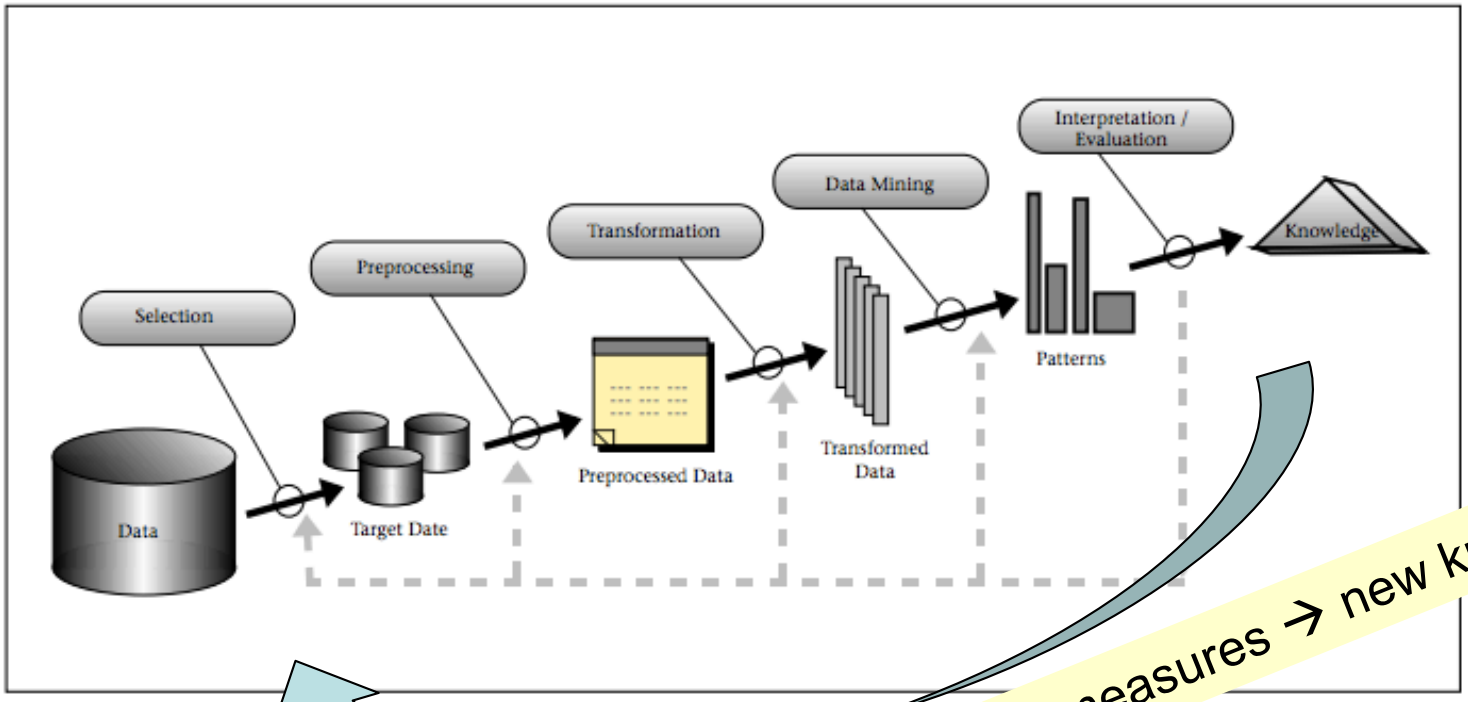
Measurement, Modeling, and Meaning in the Big Data Era
[Rice University + Army Research Institute
for the Behavioral and Social Sciences]



- **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)

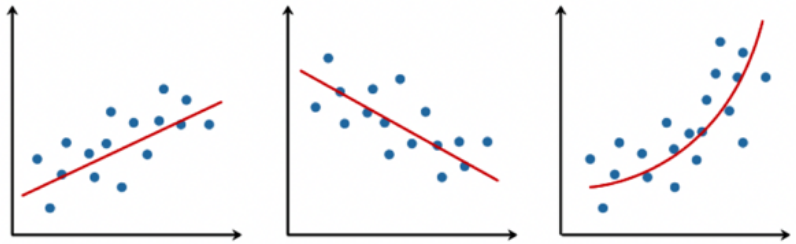
big data / induction

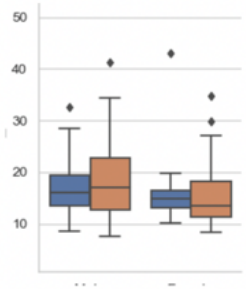
Useful 'signals' in data discovered through predictive modeling could be amplified by developing measures that collect more data (given enough development time, testing time, $...).

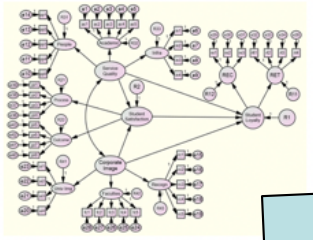


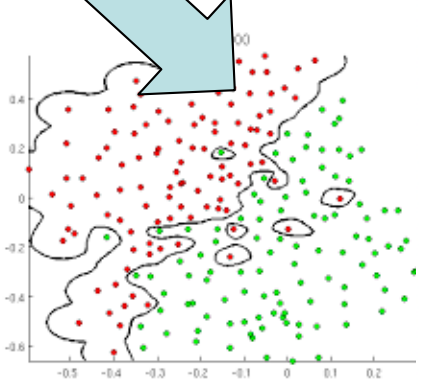
(knowledge → new measures → new knowledge)

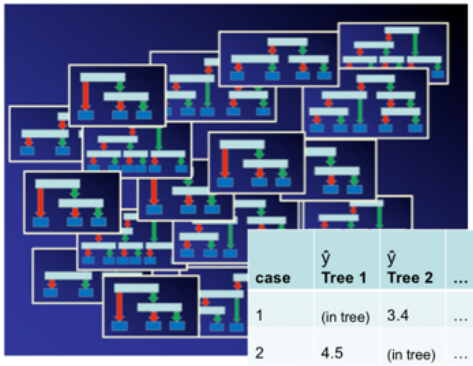
(Fayyad et al., 1996 + exciting arrow by Oswald)



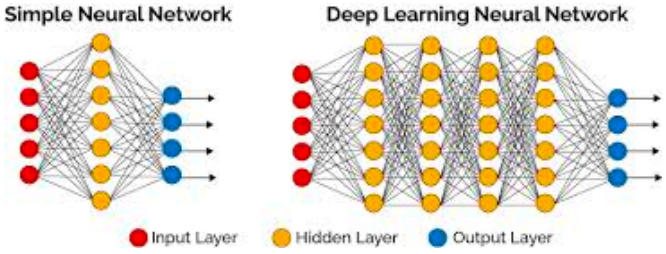


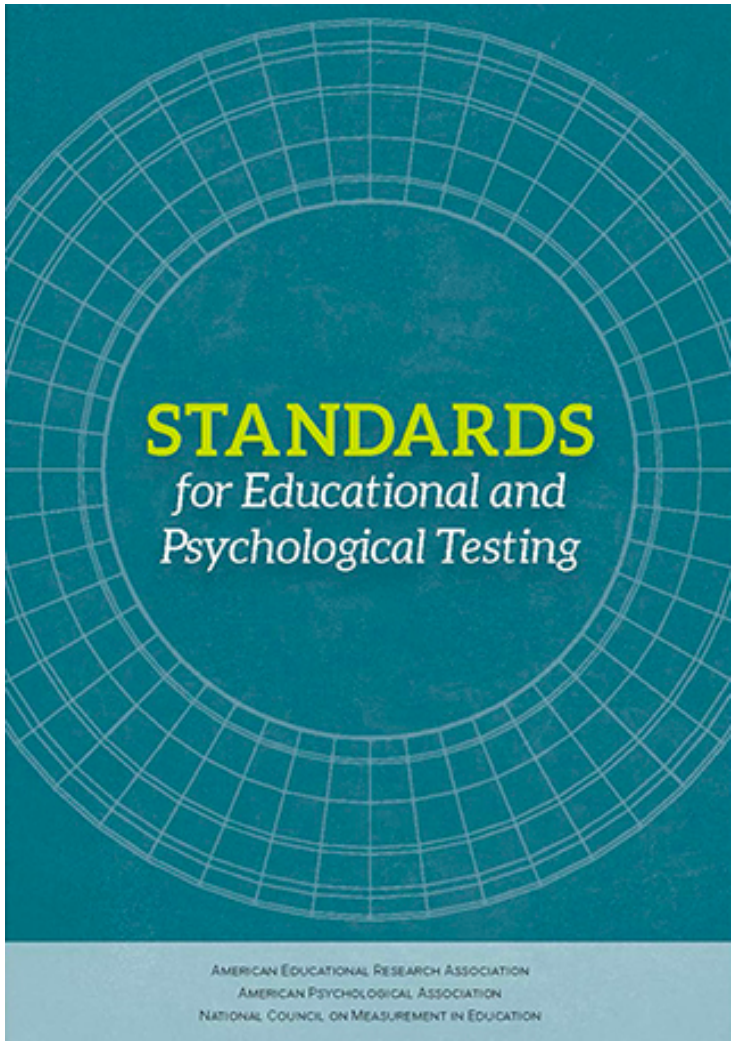






case	ŷ Tree 1	ŷ Tree 2	...	ŷ Tree *k*	*Average Predicted ŷ*
1	(in tree)	3.4	...	2.1	3.554
2	4.5	(in tree)	...	(in tree)	4.312
...	...	...	...	...	...
N	2.6	(in tree)	...	2.4	2.561





Overview of Organization and Content

Part I: Foundations

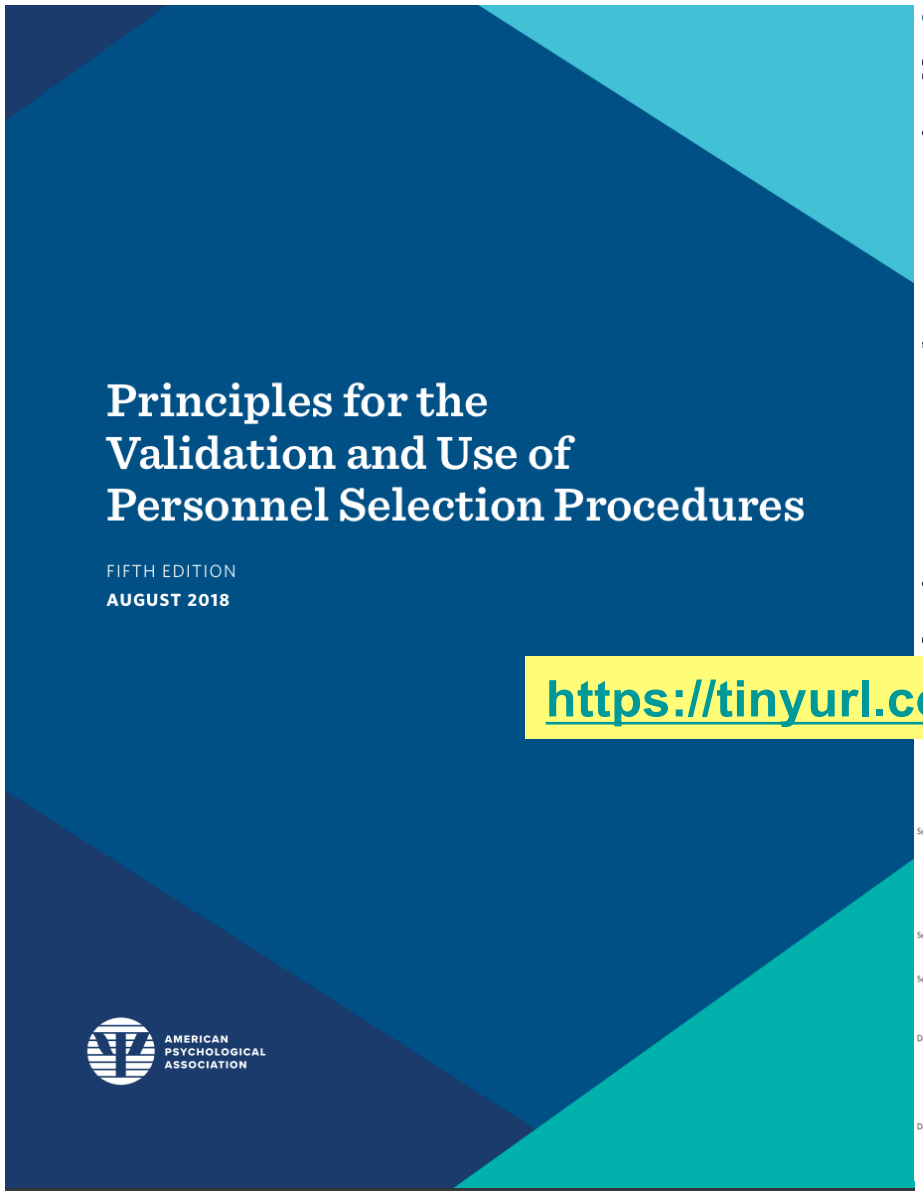
1. Validity.
2. Reliability/precision and errors of measurement.
3. Fairness in testing.

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.



CONTENTS

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

silly-not-silly question…

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)	✗	?

future directions

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)

RICE UNIVERSITY

thanks and discussion

foswald@rice.edu

https://workforce.rice.edu

RICE UNIVERSITY