An Update on Computerized Testing: Boon and Boondoggle

Fritz Drasgow

University of Illinois at Urbana-Champaign

Department of Psychology
University of Illinois at Urbana-Champaign

Outline

- Computers, the Internet, and testing
- APA Taskforce on Internet Testing
- Authentic assessment
- Proctored vs. unproctored Internet testing
- Summary

Computers, the Internet, and Testing

Department of Psychology
University of Illinois at Urbana-Champaign

Computers, the Internet, and Testing

- Clinical psychologists initiated the use of computers for assessment in the early 1960s…using teletypes and mainframes.
- The US Navy R&D Lab pioneered computerized adaptive testing (CAT) in the late 1970s and 1980s…Sands, Waters, & McBride (1997), *Computerized adaptive testing: From inquiry to operation*.

Computers, the Internet, and Testing

- The 1990s saw greatly increased use of computerized assessment
  - CAT-ASVAB implemented
  - National Board of Medical Examiners case simulation
  - GRE-CAT
  - Computerized administration of employment tests

Association of Test Publishers' Conference

- Focuses on computer-based testing (CBT).
- Has grown enormously in the past five years, from less than 150 in 2000 to 550 in 2004.
- These are the most responsible test developers, but it's apparent that they are businesses, dominated by sales and IT, not psychometricians and I/O psychologists.
- Internet is the "wild west of testing."

APA Taskforce on Internet Testing

Department of Psychology
University of Illinois at Urbana-Champaign

APA Taskforce on Internet Testing

- In 2000, APA's Committee on Psychological Tests and Assessments (CPTA) discussed issues related to testing via the Internet.
- CPTA was concerned about
  - Reliability
  - Validity
  - Item security and copyrights
  - Administrative procedures

APA Taskforce on Internet Testing

- CPTA took these concerns to its parent boards, the Board of Scientific Affairs and the Board of Professional Affairs.
- BSA and BPA endorsed the creation of a taskforce to examine CPTA's issues.
- The Taskforce on Internet Testing was therefore created in 1991.

APA Taskforce on Internet Testing

- Jack Naglieri and I were asked to co-chair.
- Taskforce members were chosen for their expertise across a wide range of areas (educational, I/O, forensic, clinical, career/vocational, cross-cultural, neuropsychological).
- The Taskforce met twice and ultimately produced a Whitepaper for APA. An article appeared in the American Psychologist in April 2004.

APA Taskforce on Internet Testing

- The Taskforce report contains five main sections:
  - Background and Context
  - New Problems Yet Old Issues
  - Issues for Special Populations
  - Ethical and Professional Issues
  - Recommendations for the Future

APA Taskforce: Background

- Internet testing has grown exponentially because it is better, faster, and cheaper.
  - A new or updated test can be quickly rolled out nationally or internationally.
  - Scores and interpretive reports can be available almost instantly.
  - Testing programs can be scaled up at little additional cost.

New Problems Yet Old Issues

- Jack, a 50 year old white male, found several tests on the Internet purporting to measure intelligence.
- Jack took these tests, and obtained scores ranging from 80 to 145.
- All of Jack's answers were random.

New Problems Yet Old Issues

- Many of the groups putting tests on the Internet are composed of IT and sales, but not I/O, psychometricians, or other measurement experts.
- BUT changing the medium of assessment does not change the basic requirements of testing:
  - Reliability
  - Validity
- See the AERA/APA/NCME *Standards for Educational and Psychological Testing*…sales and IT folks don't really understand the *Standards*, they need a psychometrician.

Issues for Special Populations

- Culturally and linguistically diverse groups may also need some type of accommodation.
- You can't just translate a test into a second language and expect scores to be equivalent to the original test.
- The Taskforce's review of various webpages found many poorly translated tests and translations of old/outdated versions of tests.

Ethical and Professional Issues

- APA Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 2003), specifically Section 9, Assessment, states

"psychologists provide opinions of the psychological characteristics of individuals only after they have conducted an examination of the individuals adequate to support their statements or conclusions. When, despite reasonable efforts, such an examination is not practical, psychologists document the efforts they made and the result of those efforts, clarify the probable impact of their limited information on the reliability and validity of their opinions, and appropriately limit the nature and extent of their conclusions or recommendations."

- Internet dating services and their "dimensions of compatibility" based on psychological profiles.

Ethical and Professional Issues

- These issues are especially salient for psychological assessments by clinicians…
- Suppose a clinician is licensed in Illinois but has a client who takes an Internet assessment while living in Minnesota…

Taskforce Recommendations for the Future

#1. Traditional psychometric standards (reliability & validity) apply to Internet tests.

#2. The validity of inferences from the many diverse types of Internet tests must be demonstrated…opportunity for exciting new research.

#3. Internet site authors should be accountable so that users receive the same type of protection as in traditional testing (e.g., need test manuals, norming studies, etc.).

Authentic Assessment

Introduction

- Computerized testing allows dynamic interactions between the test and the examinee, thus facilitating assessments of process.
- Computerized testing can also incorporate rich multimedia stimuli.
- Together, these capabilities provide opportunities for improving the accuracy and construct validity of measurement.

Innovation

- An important direction: *authenticity.*
- A recurring criticism of MC tests is that they are so artificial…they assess "granules of knowledge"
- But…in real world situations, people do not answer MC items.
- Many innovative assessments emulate real world tasks. This is a continuum, with varying degrees of real world emulation.

Some Examples

- CPA Licensing Exam
- National Board of Medical Examiners' case simulation
- National Council of Architect Registration Boards (NCARB) site design test
- Conflict Resolution Skills Assessment

Improving Authenticity

- The American Institute of Certified Public Accountants (AICPA) licensing exam began using simulations in April 2004… Richard DeVore is a key player in developing these simulations.
- Motivation: When clients meet with accountants, they don't ask multiple-choice questions; instead, they pose open-ended, vague questions.

AICPA Simulations

- For example, a client may ask the accountant various questions about a bond her company issued.
- AICPA simulates such interactions.

In the following simulation, you will be asked various questions regarding accounting for bonds. The simulation will provide you with all of the information necessary to address the questions. For full credit, be sure to answer all questions. Resources are available to you under the tab marked RESOURCES.

The Jastore Company issued $1,000,000 of 3-year bonds dated January 1, 2000. Jastore had an option that allowed the company to call the bonds at 105 of face value after one year. The coupon (stated) annual interest rate for the bonds was 10%, with semi-annual interest payments to be made on June 30 and December 31 of each year. The market interest rate at the time the bonds were issued was 7% per annum.

Jastore used the effective interest method for amortizing bond discounts and premiums.

The bonds were term bonds that were to mature on December 31, 2002.

Jastore's fiscal year for financial reporting purposes is December 31.

Using the proceeds received from a new bond issue, Jastore exercised its option to call the bonds at 105 of face value on June 30, 2001. The new bonds were 5-year, 6% bonds, issued at face value $1,000,000, interest payable annually on June 30 of each year.

Use the following spreadsheet to calculate Jastore's bond issue price. From the selection lists, select for each shaded cell in columns B, D, E, and F the appropriate value or formula and drag it to the cell. You may also enter the cell contents yourself, but all formulas must begin with "=". For cell C2, enter the amortization interest rate (as a decimal fraction). For column E find the appropriate factors for time value of money in the references under the RESOURCES tab and enter those values in the appropriate cell. The spread sheet will automatically calculate the bond issue price based on your entries. (Formula symbol definitions: * = multiplied by; / = divided by)

	A	B	C	D	E	F
1	Cash Flow Type	Time Period (n)	Amortization Interest Rate	Cash Flow Amount	Time Value of Money Factor	Cash Flow Value
2	Principal					
3	Interest					
4	Bond Issue Price					

Cells B2 and B3

- 1
- 3
- 5
- 6
- 12

Cells D2 and D3

- $1,000,000
- $100,000
- $70,000
- $50,000
- $35,000

Cells E2 and E3

See References under 'Resources' Tab

Cells F2 and F3

- D2*E2
- D2/E2
- D3*E3
- D3/E3

Exit Section

AICPA Simulations

- Central to the development of the simulations has been the notion of “measurement opportunities”.
- I.e., the simulations are designed to require many scoreable actions by the examinee.
- So, a 25-minute simulations should provide at least 10-15 measurement opportunities.

AICPA Simulations

- The simulations have the “look and feel” of actual client encounters and so have considerable face validity.
- But, do they assess skills beyond those already tested via MC questions?
- Along the lines of a suggestion by Randy Bennett, it is important to examine the disattenuated r between the MC questions and the simulations.

National Board of Medical Examiners (NBME)

- The *computer-based case simulation* provides a high fidelity emulation of diagnosis and treatment of patients.
- The NBME software simulates a patient with an unknown problem and allows the candidate physician to care for him/her.

Computer-based Case Simulation

- Clyman, Melnick & Clauser (1999). In Tekian, McGuire, & McGahie (Eds), *Innovative simulations for assessing professional competence*. Chicago, IL: University of Illinois, Department of Medical Education.

Computer-based Case Simulation

- The candidate physician can
  - request a medical history or physical exam;
  - order medical tests and procedures;
  - request consultations;
  - order treatments.

Computer-based Case Simulation

- A very interesting component of this assessment is the simulation of the passage of time.
- The examinee can allow time to move forward (but not backward). For example, the results of a medical test might be available 3 hours after the test is ordered. Once the test is ordered, the candidate must move the clock ahead 3 hours to get the test's results
- *Simultaneously, the patient's condition also changes!*

Computer-based Case Simulation

- This assessment is highly authentic because:
  - Patient care strategies are assessed in realistic simulations.
  - No artificial cues (a.k.a. multiple-choice options) are provided.
  - The process of care is not artificially broken up into segments…time and a patient's condition progress continuously.

Computer-based Case Simulation

- Adding to authenticity:
  - The assessment is dynamic…the patient's condition changes as a function of treatments ordered by the examinee.
  - "The complex interplay of clinical information about the patient with time and physician action" provides the essence of this assessment.

Computer-based Case Simulation

- Is the case simulation assessment redundant with MC tests?
- Corrected r's range between .35 to .55, so clearly the simulation adds value to the assessment of competence.
- But 4-8 hours of testing are required for adequate reliability.

Computer-based Case Simulation

- Scoring has been a challenge; Brian Clauser et al. have conducted a series of studies.
- Basically, all of the examinee's actions are recorded in a log.
- Each action is classified as
  - Beneficial…three levels (most, more, least important).
  - Harmful…three levels (nonharmful, risky, extremely dangerous).
- And points are allocated accordingly.

National Council of Architect Registration Boards (NCARB)

- The site design subtest asks the candidate to design a facility (e.g., buildings for a country club) with certain features (e.g., all trees are to be preserved, a swimming pool should be adjacent to the clubhouse, a tennis court should be aligned north-south).
- The candidate uses software tools to create a design.

NCARB

- Clearly, the work of architects involves creating designs, not answering MC questions about pre-existing designs.
- And modern architects use AUTOCAD software for designs. So, the computerized assessment is authentic.
- An important challenge for this exam involves scoring: Can the computer be used to automate scoring or are human judges required?

NCARB

- Isaac Bejar (1991, *J Applied Psychology, 76*, 522-532) describes a computerized scoring method that extracts from a candidate's design:
  - features (e.g., clubhouse location) and
  - relations among features (e.g., bleachers on side of tennis court).

NCARB

- Then the features and relations of features are compared to experts' ratings of the candidates' designs to determine critical design elements such as:
  - best bleacher location;
  - trash on right of clubhouse.

NCARB

- For the country club design, Bejar found his computerized score and one expert agreed on 85% of the scorable cases; the computer score agreed with another expert 81%; the experts agreed on 85% of the cases.

Interpersonal Skills

- Researchers have been interested in assessing social and interpersonal skills since at least 1920:
  - E. L. Thorndike (1920). Intelligence and its use. *Harper's Magazine, 140*, 227-235.

Interpersonal Skills

- But early attempts,
  - Moss (1926). Do you know how to get along with people? *Scientific American, 135*, 26-27.

  as well as recent attempts have been criticized as lacking construct validity:
  - Thorndike & Stein (1937). An evaluation of the attempts to measure social intelligence. *Psychological Bulletin, 34*, 275-285.
  - Davies, M., Stankov, L., & Roberts, R.D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology, 75,* 989-1015.

Interpersonal Skills

- Presenting video clips, rather than text-based descriptions of social situations, may, finally, provide a means of assessing these skills in a way that is not so highly related to cognitive ability or personality.
- Chan & Schmitt (1997). *J Applied Psychology, 82*, 143-159. Found a paper-and-pencil situational judgment test of work habits and interpersonal skills correlated .45 with reading comprehension, but a parallel video assessment correlated only .05.

Interpersonal Skills

- Graduate students and I have developed the following computer-administered video assessments:
  - Conflict Resolution Skills Assessment (CRSA)
  - Leadership Skills Assessment (LSA)
  - Teamwork Skills Assessment (TSA)

Interpersonal Skills

- In each case, the assessee views a video clip depicting a workplace scene. For example, in CRSA, there are clips depicting conflicts among co-workers, one worker taking credit for another's work, an arrogant worker, etc.
- After each clip, a multiple-choice question asks "If you were in this situation, what would you do?" and provides four options.

Interpersonal Skills

- In a series of studies, we've found that the video assessments correlate most strongly with ratings of the assessee's supervisor on the corresponding dimension of job performance (e.g., conflict resolution job performance)
- The video assessments correlate significantly, albeit more modestly, with ratings of overall job performance.

Interpersonal Skills

- Correlations with cognitive ability are small or nonsignificant.
- Correlations with personality dimensions are also small.

Proctored vs. Unproctored Internet Testing

Department of Psychology
University of Illinois at Urbana-Champaign

Proctored vs. Unproctored Internet Testing

- It would be convenient and cost-effective to have job applicants take an unproctored screening test to qualify for serious consideration.
- But…who is actually taking the test? The applicant? Or his/her smart friend?
- Nancy Tippins is honchoing a paper discussing issues regarding unproctored Internet testing.

Proctored vs. unproctored Internet testing

My personal view:

- The likelihood of cheating is related to the stakes of a test…
  - Low stakes test ➔ little incentive for cheating
  - High stakes test ➔ great incentive for cheating

- Exactly how this relationship maps out is an empirical question.

Two Empirical Studies

- Ben-Roy Do, a U of IL doctoral student, William Shepherd, and I have conducted two types of studies.
- Random assignment of Psych 100 students to a proctored test in a computer lab vs. unproctored Internet test.
- Field study of an organization where some sites used proctored sessions and some used unproctored Internet testing.

Lab Study

- Random assignment of 252 students to the proctored lab and 163 students to the unproctored Internet condition.

- Incentive: Students were told they would be entered into a lottery for a $100 prize based on the number of items correctly answered.

Assessments Included:

- Biodata items measuring experience
- Personality items assessing
  - Conscientiousness
  - Customer service
  - Sales
- Cognitive ability

Lab Study Findings:

	Proctored	Unproctored	
Scale	Mean	Mean	t
Biodata	17.79	18.00	-0.53
Consci.	34.33	33.62	1.23
Cust. Ser.	47.87	45.79	2.11
Sales	43.09	41.29	1.61
Cognitive	25.25	24.35	1.92

Student Study Conclusion

Students had higher performance in the proctored sessions than in the unproctored sessions!

Field Study

- Had extensive data, which we analyzed by quarter. Results were similar across quarters, so I'll just describe Spring 2002 results.
- N = 1502 in the proctored group, N = 2628 in the unproctored group.
- Assessments included Conscientiousness, Leadership, and Problem Solving.

Field Study Findings

	Proctored	Unproctored		
Scale	Mean	Mean	t	d
Conscie.	7.92	7.51	7.62	0.26
Leader-ship	9.68	9.25	5.37	0.18
Problem Solving	6.84	7.16	-3.73	-0.13

Field Study Conclusions

- Differences were statistically significant due to the large sample sizes.
- Effect sizes for medium of administration were nugatory.
- For this organization, at least, it appears that unproctored Internet testing has not led to cheating.

Overall Summary of Findings

- No evidence that unproctored testing leads to inflated scores over the range of conditions we examined.
- Note: Both studies were low stakes, one giving $100 prizes and the other led to a low paying hourly job.
- Nonetheless, this suggests that *unproctored Internet testing may be viable under some circumstances.*

How to Use Unproctored Internet Tests?

- Dan Segall (2001) has suggested unproctored Internet administration of the ASVAB for military enlistment.
- Applicants could take the ASVAB via the Internet at their own convenience.
- Individuals with qualifying scores would then travel to secure testing centers and take a short confirmation test.

Segall (2001)

- Segall described a statistical procedure based on item response theory that would be used to see if the examinee's answers to the confirmatory test were consistent with their responses to the Internet ASVAB.
- A simulation study found this analysis to be very effective in detecting cheating.

Summary

Summary

- Testing is testing, whether delivered via paper & pencil or computer.
- AERA/APA/NCME *Standards* tell us how to develop and validate tests.
- The Wild West of testing needs to be tamed, using the *Standards*, and being responsible to test consumers.
- Many opportunities for new research.

New Tests To Improve Authenticity

- AICPA simulations
- Medical case management exam
- Interpersonal skills
- Many others

New Items and New Tests

- Finally, I've only discussed a small portion of the new items and new tests today.
- And there are opportunities for many more important innovations!