Unproctored Internet Testing
Issues and an Applied Example

Rod McCloy

HumRRO

Fourth Annual River Cities I-O Conference

University of Tennessee - Chattanooga

November 8, 2008

HumRRO
Human Resources Research Organization

Overview

- **UIT Defined**
- **Issues Concerning UIT**
  - Advantages
  - Disadvantages
- **An Applied Example**
- **Final Points**

HumRRO
Human Resources Research Organization

UIT Defined

"Internet-based testing, completed by a candidate without a traditional human proctor"

- Tippins (in press)

- **Not the same as "proctor-free testing"**
  - Non-traditional/alternative proctoring might be in effect
    - Video cameras
    - Verification testing
    - Analysis of response patterns

HumRRO
Human Resources Research Organization

Issues Concerning UIT

Advantages

- **$aving$!!**
  - No travel costs for candidates
  - No proctoring costs (hiring, training, travel)
  - No hardware costs (purchase, distribution, maintenance)
  - Lower test maintenance costs

- **Administration advantages**
  - Consistency, accuracy
  - Detailed response information (e.g., item timing, accurate test timing and scoring)

HumRRO
Human Resources Research Organization

Advantages (cont.)

- **Other advantages**
  - Speed (no waiting for appointments)
  - Company Image (cutting-edge technology)
  - Expanded applicant base
    - Applicant tests at home on own time
  - Proctors ain't all they're cracked up to be
    - Sometimes unskilled/untrained/unmotivated

Disadvantages

- **Technology?**
  - Not so much these days, but some issues still (especially with clocks/timing)

- **CHEATING!**

- **Verification testing?**
  - Evidence of cheating suggestive only, not certain
  - Most depend on equivalent tests, adaptive testing, large item pools with known item parameters
  - Sensitivity paramount with cheating notification

- **Company image?**
  - Might applicants question the image of a company that offers such programs, given cheating potential is so obvious?

Disadvantages (cont.)

- **Might not those prone to cheat . . .**
  - . . . do better on the test and thus get hired . . .
  - . . . only to cheat on the job?

- **Test environment**
  - Should provide environment for optimal performance
  - But on-demand testing often occurs in environments full of distractions, leading to reduced test performance

Disadvantages (cont.)

- **Ethics**
  - Some believe UIT is unethical (Pearlman, in press; Tippins, in press)
    - Cite Principle 9.09 of the Ethics Code (2002)

"Psychologists who offer assessment or scoring services to other professionals accurately describe the purpose, norms, validity, reliability, and applications of the procedures and any special qualifications applicable to their use."

  - Relation of given score to norms?

Disadvantages (cont.)

- **Ethics**
  - Some cite Principle 9.01 of the Ethics Code to *support* use of UIT

  > "…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."

  - In essence, they interpret this to say, "If you've documented/explained UIT's effects, you're golden!"

Disadvantages (cont.)

- ## What do the Standards (1999) and Principles (2003) say?

Standard 5.2: "Modifications or disruptions of standardized test administration procedures or scoring should be documented."

Surprisingly little

Standard 5.7: "Test users have the responsibility of protecting the security of test materials at all times."

"For security reasons, the identity of all candidates should be confirmed prior to administration. Administrators should monitor the administration to control possible disruptions, protect the security of test materials, and prevent collaborative efforts by candidates. **The security provisions, like other aspects of the Principles, apply equally to computer and Internet-administered sessions**."

HumRRO
Human Resources Research Organization

Disadvantages (cont.)

- **International Testing Commission Guidelines on Computer-Based and Internet-Delivered Testing**
  - More specific
  - Guideline 45.3 recommends verification testing for unproctored assessments

"For moderate and high stakes assessment (e.g., job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e. at their convenience in non-secure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores (p. 55)."

Disadvantages (cont.)

"It is paradoxical that the concerns raised over risks of cheating in UIT, despite the fact that cheating is also an issue in proctored assessment, have resulted in the development of technologies, policies and procedures that potentially make online UIT more secure that traditional paper-and-pencil proctored tests"

-- Bartram (in press)

Disadvantages per Pearlman

- **Verification testing**
  - Adds steps (time/cost)
  - Basically tells examinee, "You can't trust UIT results"

- **Validity**
  - If cheating reduces it, reduced applicant pool contains fewer people but not better ones

- **Standardization**
  - UIT violates the daylights out of this
  - Is it a test?

> "a measure or procedure in which a sample of an examinee's behavior in a specified domain is obtained, evaluated, and scored <u>using a standardized process</u>" (SIOP, 2003, p. 71, emphasis added)].

Test Compromise

"Any Internet test that administers the same set of items to all examinees is asking to be compromised. At the very least, the items should be administered in a randomized order. It would be better yet to sample items from a reasonably large item pool."

-- Tippins et al. (2006)

Tippins's Five Camps

- **UIT Never Acceptable**
  - ID of test taker, likelihood of cheating, validity of inferences from unproctored score, ethics

- **UIT OK for Some Tests/Purposes**
  - Non-cognitive; personal development, practice testing
  - Cheating can also occur in proctored setting

Tippins's Five Camps (cont.)

- **Prevent Cheating before It Happens**
  - Ways to prevent it (warnings, threats of retesting, honor statements), or to make UIT score equal to proctored scores
  - Ways to authenticate examinee identity, monitor their behavior, or halt testing if cheating indications occur

- **Detect Cheating via Verification Testing or Statistical Means**
  - Multinomial logit, compromise IRT, IPD

Tippins's Five Camps (cont.)

- **Accept UIT without Extraordinary Measures to Prevent/Detect Cheating**
  - Argue for its utility
  - "Cheaper to accept costs of hiring a few cheaters than to spend the money to ensure accurate individual assessment"

- **Camps not mutually exclusive!**

An Applied Example

Client

- **Procter & Gamble (P&G)**
- **Research-driven company**
- **More than 138,000 employees in 80 countries**

Current P&G Selection System

- **Step 1: Requirements-Based Prescreening**
  - Questions
    - Based on job requirements
    - Developed by the hiring manager and recruiter
    - Delivered online as part of initial job application

- **Step 2: Online, Unproctored Measures**
  - Biographical data assessment
    - Measures KSAs for Management, Researcher, and Office Administration job families
  - **Adaptive cognitive ability screen**

Current P&G Selection System (cont.)

- **Step 3: Proctored Measures**
  - Paper-and-pencil cognitive ability measures
  - Three structured behavioral interviews conducted by trained, calibrated interviewers

Project Goal

- **Develop a new cognitive ability test**
  - Given via Internet
  - Unproctored, on-demand
  - Computer-adaptive
  - New content

HumRRO
Human Resources Research Organization
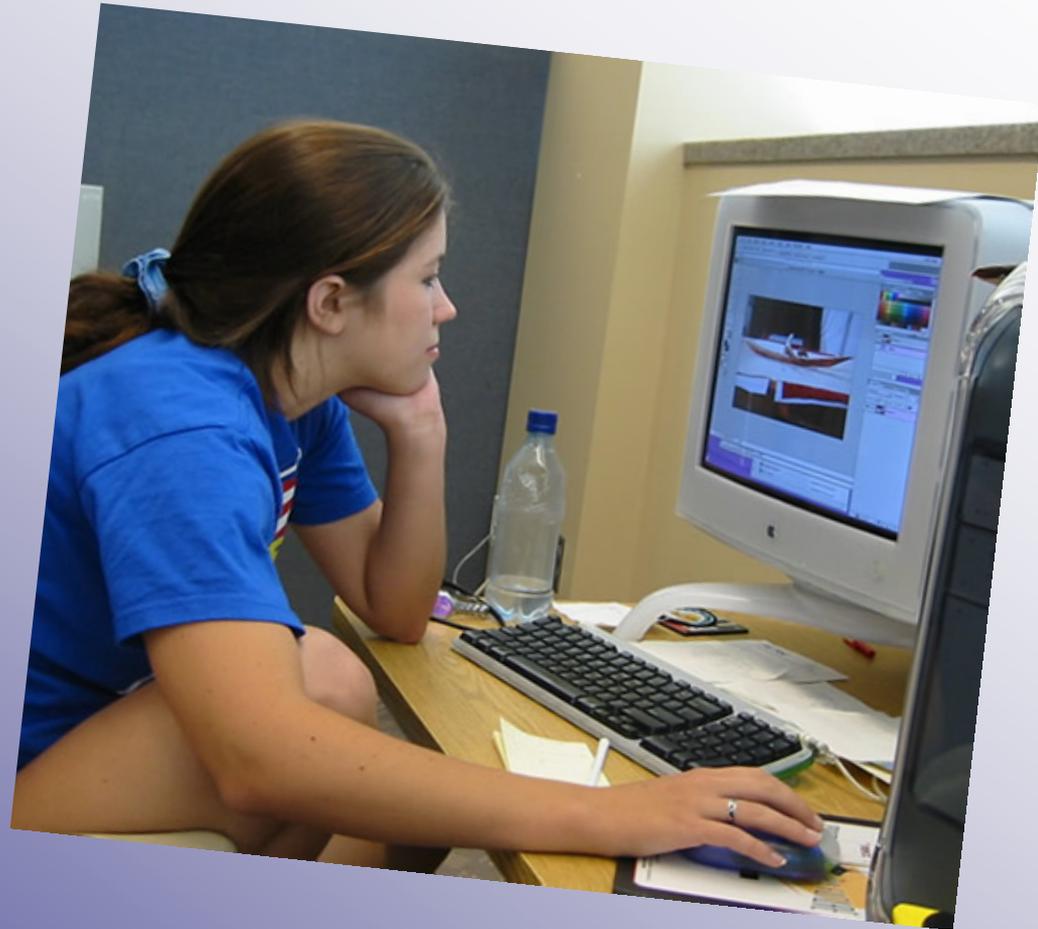
Why an Internet-Based CAT?

- **Convenience and security**
  - Ever-increasing candidate volumes
    - 500,000+ applicants per year
  - Item exposure control

- **Greater accuracy per unit time**

- **Updated models of cognitive ability**
  - (e.g., Carroll, 1993)

- **Applicant reactions**
  - Convenience of on-demand online administration
  - Reduced assessment times

HumRRO
Human Resources Research Organization

How Adaptive Question Administration Works

Start ; **question right**

Low   **High**

We then move them up or down the continuum based on whether they get the question right.

question right

Low   **High**

They get a harder question if they get it right and an easier question if they get it wrong.

question wrong

Low   **High**

We then keep presenting questions in this way until we are certain of the candidate's true level of cognitive ability.

Stop

Low   **High**

CAT-ASVAB Engine

- **A major drawing point**
  - Developed by Irwin Hom and colleagues
  - Undergirds DoD's accessioning system
  - Selects items, updates ability estimates, monitors exposure

- **P&G was prepared to license it**

- **Required modification to support iCAT demands**
  - Calculating information on demand
  - Priority given to precision *vs*. item exposure
  - Resulted in psychometric support contract with DDI

To UIT or Not to UIT?

- **Sources for determining decision**
  - Candidates
  - Other companies and consultancies
  - Scientific literature

Candidate Reactions

- **Studies of Potential/Actual Candidates**
  - <u>Study 1</u>: 1,000+ university students, 20+ countries
    - Did not know P&G would use the process

  - <u>Study 2</u>: Reactions from several thousand candidates via the career site

  - <u>Study 3</u>: Part of test calibration research with 150,000+ P&G candidates

Candidates: Results

- **UIT > > P&P, provided . . .**
  - UIT created greater flexibility in taking tests
  - Faster responses on their progress/status

Industry

- **Extant practice for UIT**

- **Not *whether* to use UIT, but *how* to . . .**
  - develop
  - validate
  - use
  - **. . . it most effectively**

Literature: Best Practices

- Use of <u>multiple supervised assessments</u> to verify scores on our UIT biodata and cognitive assessments

- <u>Partnership</u> with consultancy and academic leaders in assessment research, development, and practice

- Extensive/rigorous <u>translation</u> of all assessments (40+ for the cognitive assessments) to ensure construct measurement rather than English language ability

- <u>Concurrent validation</u> of the tests with several thousand employees worldwide under **unsupervised conditions for our UIT tools** and **supervised conditions for our supervised tests**

- <u>Cut scores</u> that allow proper candidate flow to reduce false negative/positive results

- Development of <u>advanced features</u> for UIT biographical data assessment, including
  - randomized item delivery
  - cultural scoring algorithms that enable appropriate measurement of candidates' fit with KSAs

Literature: Best Practices (cont.)

- Development of advanced features to <u>maximize the precision and security</u> of the UIT cognitive ability assessment
  - <u>Adaptive item delivery</u> of a small subset of the total item pool
  - <u>Extensive item pools</u> globally developed/screened with more than 150,000 candidates via IRT calibration and DIF assessment
  - Item-level, globally calibrated <u>timing</u> to make cheating more difficult
  - Ongoing <u>live-item research</u> to understand and detect item parameter drift as an indication of item compromise
  - Ongoing <u>item development and calibration</u>
  - Item <u>exposure control</u> algorithms
  - <u>Parallel item pools</u> that can be interchanged, making it difficult for candidates to acquire test content

Literature: Best Practices (cont.)

- Instructions to candidates that
  - all responses should be made <u>without any help</u> from others
  - online test <u>results would be verified</u> with additional assessments under supervised conditions

- Development of an <u>assessment portal system</u> to manage access, consistent messaging, and delivery of multiple UITs to all candidates globally

- Development of <u>enhanced web-based delivery systems</u> that minimize the ability of candidates to game the assessments for more time or capture the test questions

Conclusion

- **This set of activities/requirements demonstrates the rigor, resources, and cost required to effectively research, develop, and validate noncognitive and cognitive UIT measures**

- **Reproduction of this effort would require**
  - Access to tens of thousands of respondents (both candidates and employees)
  - A multi-million dollar budget
  - Extensive I-O psychology and technology resources for specifying, designing, programming, and maintaining the online systems

- **4 years from design to implementation**

Benefits to Candidates

- **On-Demand, 24/7 Access to Assessments**
  - Candidates manage the testing environment with administration timing that works for them

- **Faster Status Updates**
  - Candidates want to know if they are moving forward in the application process
  - Primary request of P&G in the candidates' reactions research

- **Standardized Assessment Delivery**
  - No issues related to human proctoring of assessments
  - Ensures all candidates receive the same test instruction, preparation

- **Source-Independent Application Consideration**
  - Source = university campus, career conference, etc.
  - Not true for paper-based system (access to candidates primarily based on source

- **Language Choice**
  - Not managed by recruiter who may default to English

Benefits to P&G

- **Consistent Candidate Management Across Geographies, Business Units**
  - Increasingly important requirement as applications have grown from 25,000 in 1999 to 500,000+ in 2007

- **Casting of Wider Talent Net**
  - Permitting outreach to more diverse candidate pool

- **Management of Multi-Hurdle, Multi-Assessment Selection System**
  - Provides more holistic, complete assessment of candidates' fit with KSAs

- **Controlled Messaging**
  - Deliver messaging to candidates (and consumers) in a way that builds overall Company equity

- **Assessment Efficiency**
  - Reduced screening times and costs for up-front assessments
  - Real-Time score results delivered to recruiters

Benefits to P&G

- **Role Enrichment**
  - For recruiters and employees who otherwise would have spent time delivering tests

- **Improved Cognitive Test Item Security**

- **Consideration of Candidates' Cultural Backgrounds**
  - Used when scoring noncognitive assessments

Payoff

- **No formal utility analysis yet**

- **Have seen significant cost reductions, improved process efficiencies**

- **Example:** P&G will deliver approx. 10,000 fewer supervised p-&-p cognitive tests in Japan alone this year
  - New process isolates top talent faster with less cost and fewer resources (internal, external)

- **Favorably reviewed in audits by government agencies**
  - Have stood up to legal challenges

- **Bottom Line:** In an era of reduced availability of talent and greater competition for it, we believe this system offers a defensible competitive advantage for P&G.

Participating Parties

Procter & Gamble
- **Robert E. Gibby**
- **Andrew M. Biga**
- **Angela K. Pratt**
- **Jennifer L. Irwin**

Michigan State University
- **Anthony S. Boyce**

HumRRO
- **Rod McCloy**
- **Suzanne Tsacoumis**
- **Dan J. Putka**
- **Irwin Hom**
- **Andrea Sinclair**
- **Matt Trippe**

University of Minnesota
- **Stephan Dilchert**
- **Deniz S. Ones**

Logos Corporation
- **Magda Colberg**

DDI
- **Kirk Fischer**
- **Evan Sinar**

Bowling Green State University
- **Alison Broadfoot**

Project Activities

- **Analyze data from extant P&G tests**
- **Identify content domains for tests**
- **Develop test content**
  - NR – drawn from current tests, WestEd
  - FR – new test; Ones/Dilchert; review
  - LBR – new test; Colberg/HumRRO
- **Identify scoring/stopping/exposure rules**
- **Conduct calibration study (numerous 9-item forms)**
- **Establish item timings**
- **Construct item pools**
- **Support parallel DDI platform effort (Irwin)**
- **Conduct QA of testing platform**
- **Contribute enhanced functionality (DIF, IPD)**
- **Develop criterion scales for validation study**

Test Content

$$p \rightarrow q$$
$$\ldots$$
$$p$$
$$\ldots$$
$$\therefore q$$

$$p \rightarrow q$$
$$\ldots$$
$$\sim q$$
$$\ldots$$
$$\therefore \sim p$$

(1)   (2)   (3)   (4)   (5)
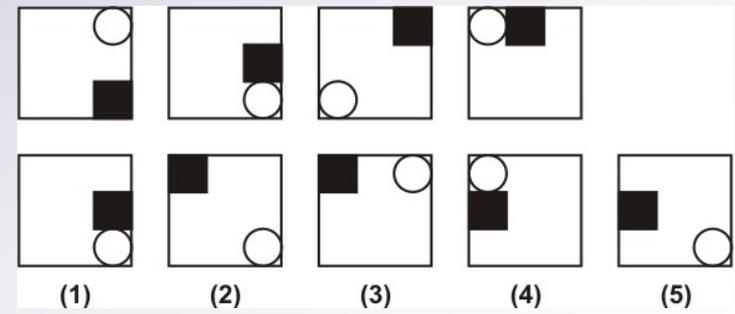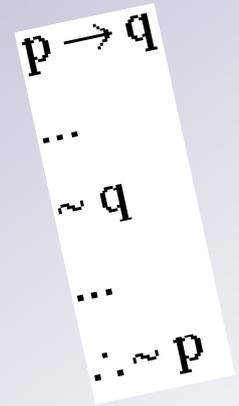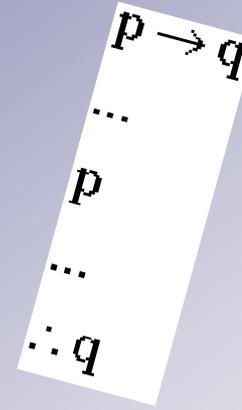
Numerical Reasoning

Sample Questions:

- **Analyze and solve business-related problems** involving numerical information and data

- Strongest historical P&G subtest

- Displayed best IRT properties among extant content

A mixture is created by combining chemicals A, B, and C in the proportion 3:1:2. In combining the chemicals, 2 additional gallons of chemical C were accidentally added to 15 gallons of chemical A and 5 gallons of chemical B. How many additional gallons of chemicals A and B must be added to accurately produce the mixture?

- ● 3 gallons of A and 1 gallon of B
- ○ 2 gallons of A and 1 gallon of B
- ○ 3 gallons of A and 2 gallons of B
- ○ 5 gallons of A and 3 gallons of B
- ○ 4 gallons of A and 2 gallons of B

The annual sales target in millions of units sold is 2.500 for Product B. The sales target for Product A is 90% that for Product B. This year, the two Products together exceeded their combined sales targets by 30%. What were the combined sales of Products A and B this year in millions of units?

- ○ 4.275
- ○ 5.175
- ○ 5.700
- ○ 5.850
- ● 6.175

Deductive Thinking Style – Logic-Based Reasoning

- Using information known to be true to solve problems by logically deducing the valid conclusions

- Directly measures a candidate's ability to **quickly synthesize information**

- **Top-down process**

Intermediaries play a critical role in the distribution of a product. Intermediaries are classified as either wholesalers or retailers. Wholesalers sell a product to another business, which in turn resells the product to the final consumer. Retailers, on the other hand, sell a product directly to the final consumer. A company's decision to use wholesalers or retailers depends on a variety of factors, including cost, target markets, and the nature of the product.

From the information given above, it can be validly concluded that:

❍ There are at least some intermediaries that are neither wholesalers nor retailers.

❍ A wholesaler does not sell a product directly to the final consumer.

● Products directly sold to the final consumer are not sold via retailers.

❍ If an organization is not an intermediary, then it is not a wholesaler.

Inductive Thinking Style – Figural Reasoning

- Solving problems by determining what information is known when presented **novel problems**

- Assesses capacity to:
    - think outside of the box
    - uncover relations
    - apply learning of new relations to solve the novel problem at hand

- **Culture-fair test** content to drive global diversity

- Bottom-up process

- **Serves as Reasoning Screen**
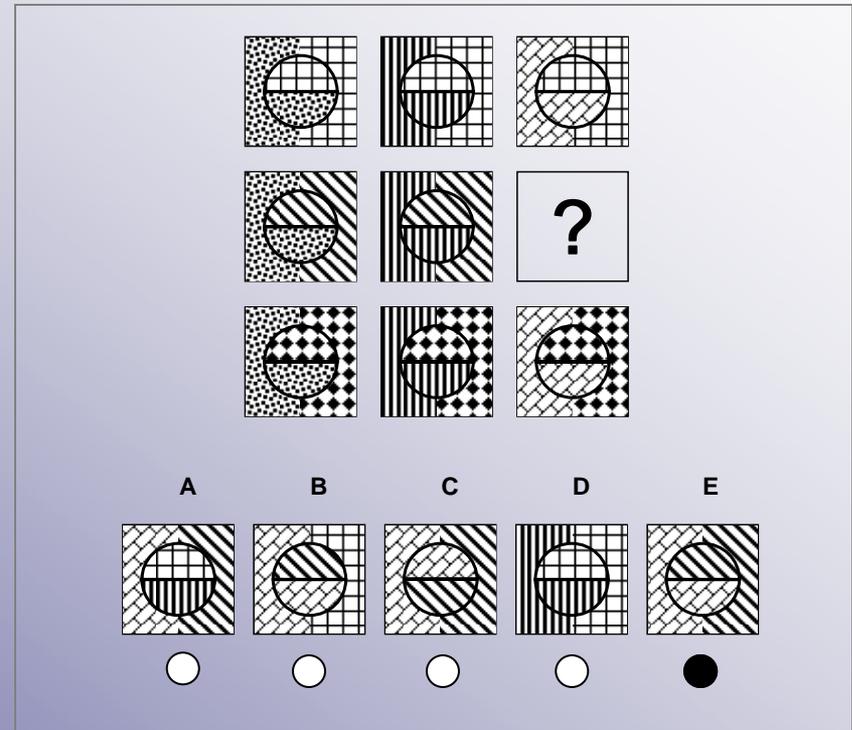
Final Points

- **Pandora's out and dancing**
- **Proper use could save much money**
- **Proper development will cost much money**
- **Savings**
  – Remember that utility a function of validity as well as cost
- **User buy-in important component**

Final Points (cont.)

- **Multiple-hurdle systems likely more hospitable environs for cognitive UITs**

- **Yet to be legal challenge**
  - No precedent
  - Updated Guidelines, Standards, and Principles need to address UIT
  - Arm yourselves for bear
    - Many potential issues (test conditions, verification decisions)

- **Documentation critical**

- **Not for those who want something "quick and dirty"**

Thank you!!