Ch. 7: Test Administration

Psyc 402 - Midterm 2 Study Guide

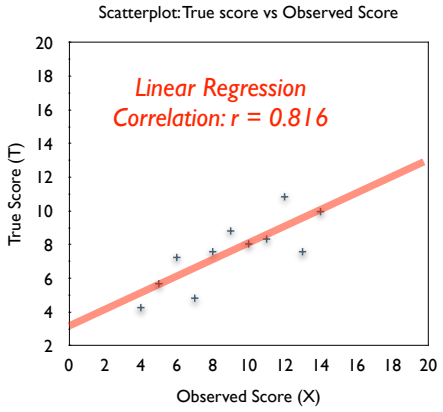
- 50 questions total, multiple choice
- Topics:
  - K04 - Reliability (12 questions)
  - K05 - Validity (14 questions)
  - K06 - Test Development (12 questions)
  - K06 - Test Administration (12 questions)

Reliability (Ch 4)

- Constructs, definitions and difficulties with
- Classical Test Score Theory
- 4 kinds of Reliability
- Common R values for reliability
- SEM and Confidence Intervals
- Increasing Reliability

Classical Test Score Theory

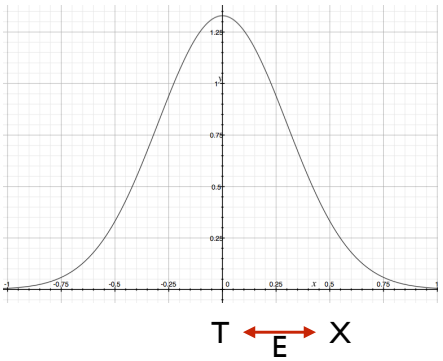
X	T
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



Classical Test-Score Theory

- T= True Score
- X = Observed
- E = Error

- X = T+E
- E = X-T



4 Kinds of Reliability

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach's Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

Validity (Ch. 5)

- Griggs v. Duke Power, Ricci v. DeStefano
- History of Validity
- 4 kinds of Validity
- Common R and R^2 values for validity
- Reliability and Validity

4 Kinds of Validity

	Description	Notes	Statistic
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific event?	requires a well-defined criteria	Pearson's R (correlation) between Test and Criteria
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson's R)

Test Development (Ch. 6)

- Definitions for several kinds of questions
- Pros & Cons of different kinds of questions
- Probability of Guessing
- Item Difficulty, Discriminability
- ICC

Test Administration (Ch. 7)

- Assumptions of Classical Test Score theory
- Bias
- Ways of administering tests / the protocol
- Bias / demographic variables
- Reducing bias
- Computerized assessment, issues with
- Polygraph exams
- Low Base Rate / False Positive Paradox

Review

- Test Items
  - Common formats
  - Alternative formats
- Item Analysis
  - Item Difficulty
  - Discriminability
- Item Response Theory
  - Computer Adaptive Testing

Test Administration

- Theory - what affects test scores?
- The Examiner and the Subject
  - Relationships between Examiner and Subject
  - Race, Language of subject
  - Examiner Training
  - Expectancy effects / Reinforcement
  - Computer-administered testing
  - Subject Variables
- Behavioral Assessment Methodology
  - Reactivity, Drift, Expectancy
- Deception & Detection of Malingering

What affects test scores?

- Simple View / Classical Test theory
  - X = T + E
  - Observed Score = True Score + Random Error
- Modern View
  - Error - is not always random
  - Error - comes from both subject AND protocol

Test protocol

- Simple view :
  - A test is just the collection of test items and grading rules and norms. The "What"

Test protocol

- Simplistic
  - **What**: the test items & scoring.
- More Realistic - the entire situation:
  - Set: **Why**?
  - Setting: **Where/When**
  - Examiner: **Who**?
  - Method of administration: **How**?
- Often the "**What**" is specified carefully but the others are not.

Rapport & IQ scores

- Feldman & Sullivan (1960) : children taking WISC
- Neutral condition
- High rapport condition
- Results:
  - Grades 1-4: little or no effect
  - Grades 5-9: IQ scores increased (122 vs 109)
- Age x Rapport interaction?

Interaction Effects

Rapport & IQ scores

- Review by Fuchs & Fuchs (1986)
  - 22 studies
  - 1500 students
  - Familiar examiners vs. strangers
    - IQ increased 0.28 SD overall
    - IQ increased up to 0.5 SD in lower SES students
- SES x rapport interaction

- Question: given these results, is cross-cultural testing fair?

Race of Tester & Subject

- Common belief: when race of tester and subject differ, testing is biased
- Satler (2002, 2004) review found little evidence (only 4 of 29 studies)
- Concludes it is a "myth"

Race of Tester & Subject 2

- Some studies do show effects
- Effects tend to be larger when
  - testing protocol is more flexible
  - testers are less-well trained

- Explanation?

Stereotype Threat Example

- Emerging research in late 2010s
- Subject's own beliefs about group performance affects individual performance
- Example:
  - Test of math
  - subjects told "men score higher"
  - subjects equally capable
  - result: men score higher

Stereotype Threat Review

- Studies show between 17% to 80% of test differences (SAT, IQ) due to S.T.
- Theory: self-defeating cognitions increase load on Working Memory, lower engagement, motivation, etc.

- Effect reduced when subjects told
  - Intelligence is malleable
  - Environment can affect test scores
  - Individuals may outperform group averages

Stereotype Threat - Reducing

- Problem:
  - Many tests begin with demographics questions (age, gender, race…)
  - Triggers stereotype threat

- Solution:
  - move demographics questions to the end?

Training of Testers

- Administering some tests properly is difficult
- Patterson et al. (1995):
  - new testers made numerous errors administering the WAIS-R
  - tester improved after 10 administrations. (Graduate students often get only 4 practice)
- Behavior of Testers is largely unspecified.
- Example : Do you say "yes" or "good job" when an answer is correct? Do you say "keep trying" if a person appears to be giving up easily?

Expectancy Effects

- Definition: finding evidence biased towards a pre-existing hypothesis…

Expectancy Effects

- Two kinds
  - (A) selection bias in collecting data -- ignore data that seems wrong, accept data that fits your theory
  - (B) actually changing the environment -- encourage desired behavior by subtle or overt prompting

Expectancy Effects: Rosenthal (1966)

- Rate faces on "Success" or "Failure"
- All subjects get same faces, but
- Half told faces were successful people.
- Result: were about 1 point higher (on a 20 point scale)
- Conclusion: expectation influences judgement
- Effects even seen when rating non-humans (e.g. rats)

Expectancy Effects: Testing

- Sattler et al. (1970): expectancy effects when rating ambiguous response on an IQ test.
  - same response given to raters
  - Half told it is a "smart" child.
  - Results: "smart" children scored better.
- Sattler (1998)
  - same even when the test answer was not ambiguous.
- Literature Review: inconsistent results, which are small
- Conclusion: small, but real problem, design tests with clear scoring rules

Lawyers Guns and Money

- Reinforcements shape behavior. Can reinforcement change test results? IQ tests?
- Theoretical issue: if money, candy or praise can improve IQ score, what exactly is IQ a measurement of?
- *Is IQ a valid measure if Intelligence?*
- Review suggests complex effects
  - In one study: boys responded to tokens, response to praise was mixed (girls: performance improved but slowed; boys went faster.)

Are your intestines too long?

- Expectancy effects higher on tests with subjective scoring
  - Cannell (1974) : "yes" answers to physical symptoms increased when interviewer gave approving nod.
  - Yes answers increased to nonsense questions: "Do the ends of your hair itch" and "Are your intestines too long?"

The Humans are Dead

- Problem: human interviewers bias performance of subjects
- Solution: use robots instead?
- Pros:
  - complete standardization
  - adaptive testing
  - precision of timing
  - cost effective
  - patience
  - bias reduced
  - encourage socially undesirable responses?

Robots make better testers?

- Resenfedl et al. (1989): subjects preferred computer-test to paper-and-pencil test
- Lock & Gilbrt (1995) subjects took MMPI via computer, paper & pencil, or interview. More undesirable information revealed with computer version (and subjects rated the computer version as more pleasant)
- Studies show computers at least as reliable as humans
- Issues about validity:
  - administration vs. scoring vs interpretation. Testing vs. Assessment

The Robot will see you now...

Therapeutic Implications

Patients with the 2-7/7-2 profile type are often very motivated for psychotherapy because of their high level of subjective distress. Many are also sufficiently psychologically minded to make good use of therapy. They often seek support and reassurance, and can be overly guilt-ridden and self-critical in therapy. Their tendency to obsesses (particularly when coded 7-2) can make for unproductive periods during sessions. They do not tend to terminate therapy prematurely. In fact, some may become dependent on the therapeutic relationship and be reluctant to terminate. The prognosis for therapy is usually quite good. A variety of cognitive/behavioral techniques are particularly likely to be of value including cognitive restructuring (e.g., learning to dispute the unreasonable demands they make on themselves) and relaxation training. If their level of anxiety or depression becomes disabling, a pharmacologic component to treatment may prove beneficial in not only relieving stress, but in permitting the patient to maximize the value of psychotherapeutic interventions

Automated Testing Issues

- Boundaries of competence? "Psychologists provide services...only within the boundaries of their competence…"
- Scientific Basis? "Psychologist's work is based on established scientific […] knowledge of the discipline"
- Delegation of Work: "Psychologists who delegate work...authorize only those responsibilities that such persons can be expected to perform competently…"
- Use of Assessments: "Psychologists use assessment instruments whose validity and reliability have been established for use...

Automated Testing Issues 2

- Use of Assessments: "Psychologists use assessment instruments whose validity and reliability have been established for use with members of the populations tested."
- Assessment by Unqualified Persons: Psychologists do not promote the use of psychological assessment techniques by unqualified persons, except when such use is conducted for training purposes with appropriate supervision.

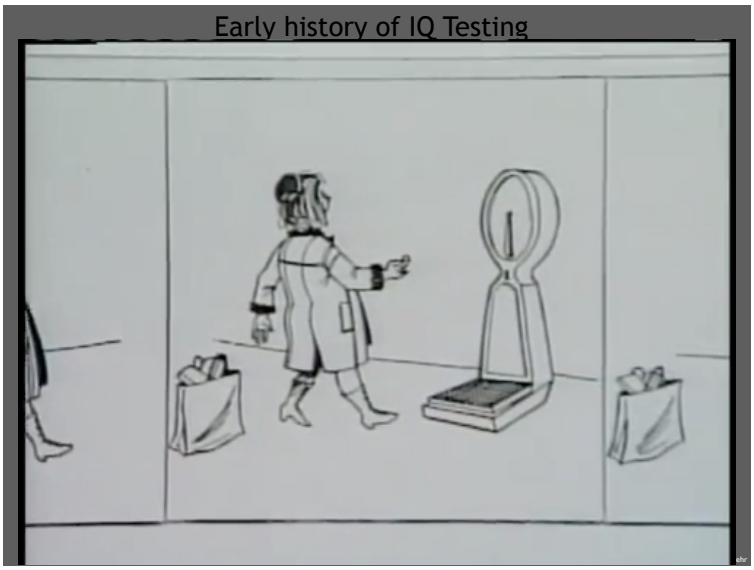
Automated Testing Issues 3

- "When interpreting assessment results, including automated interpretations, psychologists take into account the purpose of the assessment as well as the various test factors, test-taking abilities, and other characteristics of the person being assessed, such as situational, personal, linguistic, and cultural differences, that might affect psychologists' judgments or reduce the accuracy of their interpretations

"Subject Variables"

- Motivation
- Anxiety
- Illness
- Medications
- Hormones
- Sleep
- etc...

Early history of IQ Testing



Army Alpha and Beta

- Sample questions
- Sample administration protocol
- Results

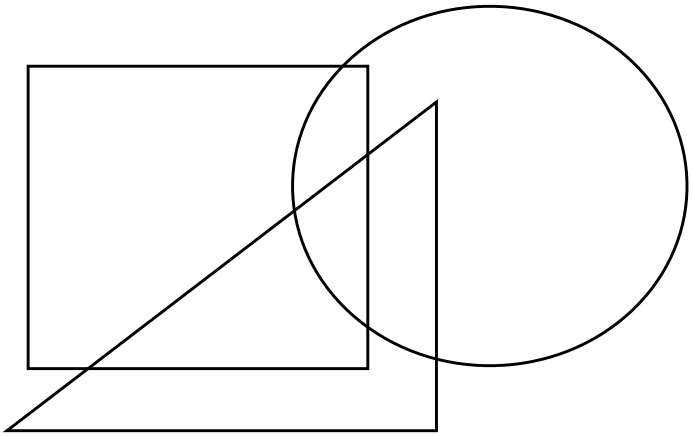
Army Alpha

- Washington is to Adams as First is to _____
- Crisco is a : patent medicine, disinfectant, toothpaste, food product
- The number of a Kaffir's legs is : 2, 4, 6, 8
- Christy Mathewson is famous as a : writer, artist, baseball player, comedian

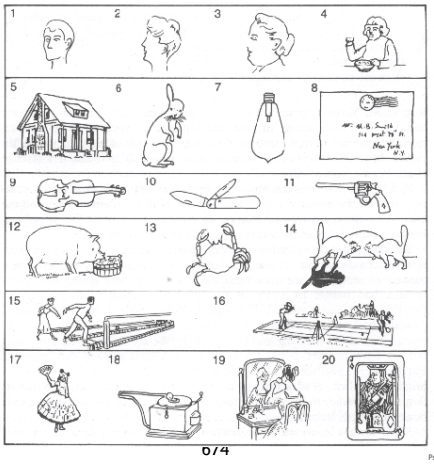
Army Alpha

- Attention! Look at 4. When I say "go" make a figure 1 in the circle but not in the triangle or square, and also make a figure 2 in the space which is in the triangle and circle, but not in the square. Go!

Army Alpha

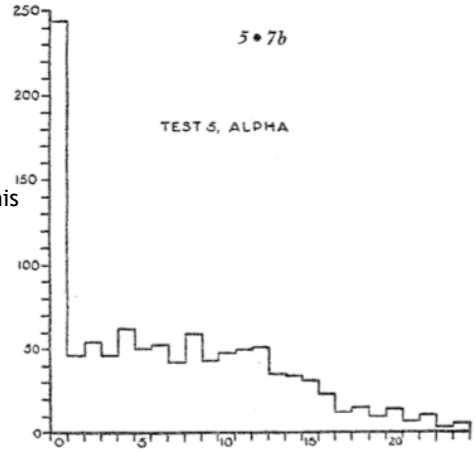


Army Beta



Army Alpha Results

- What is a test like this measuring?



Behavioral Assessment

- aka “Functional Assessment”
- Work samples, on-the-job testing, “in situ” / “in vivo”
- More active role of psychologist / observer / rater can lead to bigger problems with accurate measurement
- Reactivity…
- Drift…
- Expectancies…

Reactivity

- Reliability of observers is highest when the *observers* are being observed
- Reid (1970) : observer accuracy dropped 25% when told their work would not be measured
- Methods: random sampling, covert sampling

- Measures of test Reliability are often done in ideal situation, not everyday situation.

Drift

- Observers can be trained to certain level of accuracy, but their performance tends to change slowly over time.
- A 9/10 rating when you first started may only be an 8/10 now.
- Drift can happen on individual or group basis.
- Group drift especially hard to counteract, since the group members tend to support each other’s ratings.
- Drift is frequently ignored: 17% of studies even report the # of raters. 10% documented the training of raters. 5% tested for drift.

Expectancies in Beh. Obs.

- Expectancy effects, in Behavioral Observation situations, are similar to those seen in Testing
- Effects are subtle, small, but real and can be a significant problem in some contexts
- Effects seem to be largest when Observer is rewarded for reporting certain behaviors.

Review

Ch. 7 - Part 2

Behavioral Observation

- Deception and Malingering…
- The Low Base Rate /False Positive Paradox…

Deception

- People are very poor at detecting deception
- Polygraph "Lie Detector" tests
  - invented in 1921 by John Larson (medical student and police officer in Berkeley CA)
  - developed to scientifically measure deception
- Theory: physiological responses happen when subject lies, and can be measured objectively.

Bad Words

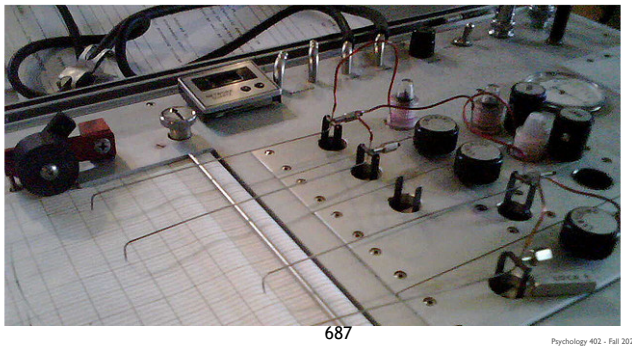
Lie Detectors: Video

"The government doesn't like that…"

Polygraph.com owner pleads guilty to training customers to beat polygraph

"Lying, deception, and fraud" influenced hiring of national security officials.

by **David Kravets** - May 14, 2015 8:52am PDT



Polygraph Examination

- Lie detector tests:
  - poor reliability
  - poor validity
  - Example: correlation between honesty test and thefts : r= 0.13, (r^2 = .02) meaning about 2% of variance is explained.
  - Over 95% false positive rate
- Two thirds of experts call "pseudoscience"
- Some belief that participation by Psychologists in such testing is violation of ethical principles (Camara & Schneider 1994)
- Few countries use them (e.g. not used in Europe)

U.S. DEPARTMENT OF LABOR

EMPLOYMENT STANDARDS ADMINISTRATION

Wage and Hour Division
Washington, D.C. 20210

NOTICE

EMPLOYEE POLYGRAPH PROTECTION ACT

The Employee Polygraph Protection Act prohibits most private employers from using lie detector tests either for pre-employment screening or during the course of employment.

PROHIBITIONS

Employers are generally prohibited from requiring or requesting any employee or job applicant to take a lie detector test, and from discharging, disciplining, or discriminating against an employee or prospective employee for refusing to take a test or for exercising other rights under the Act.

EXEMPTIONS*

Federal, State and local governments are not affected by the law. Also, the law does not apply to tests given by the Federal Government to certain private individuals engaged in national security-related activities.

The Act permits *polygraph* (a kind of lie detector) tests to be administered in the private sector, subject to restrictions, to certain prospective employees of security service firms (armored car, alarm, and guard), and of pharmaceutical manufacturers, distributors and dispensers.

The Act also permits polygraph testing, subject to restrictions, of certain employees of private firms who are reasonably suspected of involvement in a workplace incident (theft, embezzlement, etc.) that resulted in economic loss to the employer.

EXAMINEE RIGHTS

Where polygraph tests are permitted, they are subject to numerous strict standards concerning the conduct and length of the test. Examinees have a number of specific rights, including the right to a written notice before testing, the right to refuse or discontinue a test, and the right not to have test results disclosed to unauthorized persons.

ENFORCEMENT

The Secretary of Labor may bring court actions to restrain violations and assess civil penalties up to $10,000 against violators. Employees or job applicants may also bring their own court actions.

ADDITIONAL INFORMATION

Additional information may be obtained, and complaints of violations may be filed, at local offices of the Wage and Hour Division. To locate your nearest Wage-Hour office, telephone our toll-free information and help line at 1 - 866 - 4USWAGE (1 - 866 - 487 - 9243). A customer service representative is available to assist you with referral information from 8am to 5 pm in your time zone; or if you have access to the internet, you may log onto our Home page at www.wagehour.dol.gov.

THE LAW REQUIRES EMPLOYERS TO DISPLAY THIS POSTER WHERE EMPLOYEES AND JOB APPLICANTS CAN READILY SEE IT.

**The law does not preempt any provision of any State or local law or any collective bargaining agreement which is more restrictive with respect to lie detector tests.*

U.S. DEPARTMENT OF LABOR
EMPLOYMENT STANDARDS ADMINISTRATION
Wage and Hour Division
Washington, D.C. 20210

WH Publication 1462
June 2003

Lie Detector Tests

- Prohibited by employee Polygraph Protection Act of 1988 (EPPA).
- "Employers generally may not require or request any employee or job applicant to take a lie detector test, or discharge, discipline, or discriminate against an employee or job applicant for refusing to take a test or for exercising other rights under the Act."
- Exceptions -- security firms and pharmaceutical manufacturers, and government.
- Not admissible in court of law (Frye Rule from 1923)

The low base rate / False Positive Problem

Decision Making & Errors

		The Real World	
		Guilty	Innocent
You Decide	Guilty	True Positive 1-β Power	False Positive Type I Error α Alpha
	Innocent	False Negative Type II Error β	True Negative 1-α

The low base rate / False Positive Problem

- Scenario
  - 10,000 people tested
    - 10 are actually spies
  - Lie Detector Test
    - 84% accuracy (theoretical)
    - 16% false positive rate

Example

		The Real World	
		Guilty	Innocent
You Decide	Guilty	True Positive 8	False Positive 1598
	Innocent	False Negative 2	True Negative 8392

False Positives

- NAS concluded that if 10,000 employees (of whom 10 were spies) were given a polygraph:
  - 8 spies would fail the test
  - 1598 non-spies would fail the test
  - Roughly 99.6% of those failing the test would be False Positives
  - This assumes a very optimistic 84% accuracy (actual accuracy much worse)

False Negatives

- Notorious people not being caught by polygraphs:
- Aldrich Ames (CIA agent, KGB spy) passed twice
  - His advice? “Get a good night’s sleep…”
- Gary Ridgway (the “Green River Killer”)
  - another suspect failed test (but was innocent)
  - Ridgway passed polygraph test in 1984
  - Killed more people after passing the test

Lie Detector Tests as Coercion

- Poor reliability & validity
- Widely prohibited by law
- Why used?
  - Common belief of accuracy: test is punishment and an inducement to confession.
  - Fear of being caught causes such severe anxiety that a person may choose to confess (even sometimes, to a crime not committed).
  - Test can be fairly easily beat with simple training.
  - Which groups of people lack knowledge and are susceptible to these tests?

Detection of Malingering

- Sometimes there are benefits to performing poorly on a test (disability, forensic, military, etc.)
- Often called “faking bad”
- On some tests, an untrained person can’t know what “normal” performance is.
- Malingering tests give false feedback, which can encourage a person faking bad to perform worse than people with actual injury. In some cases, perform worse than chance.
- Note: not all such performance is intentional. Possible for patient to believe in their illness.

Hiscock Forced-Choice Procedure