Psychometrics 101: Engaging your board and new staff members

June 12, 2019

Lawrence J. Fabrey, PhD

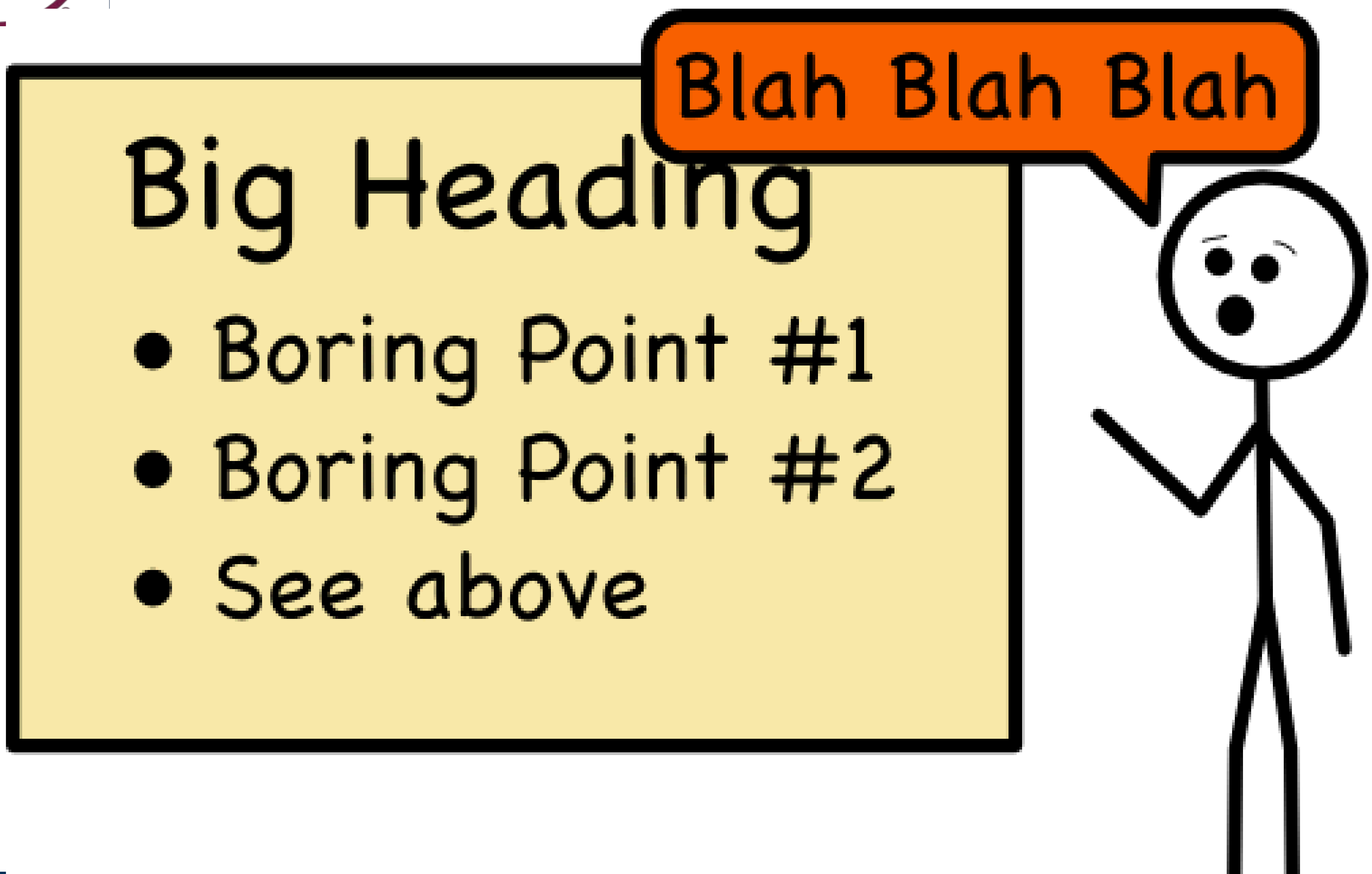
Outline for today

- Identifying the Intent of the Measurement Process
- Psychometric standards
- Theories Of Measurement
- Reliability
- Validity
- Questions for clarification are invited as we go, planning time for other Q&A at the end

What is psychometrics?

1. The word is formed by two parts; the parts convey the basic meaning: ‘metric’ refers to measurement and ‘psycho’ to the mind, so psychometrics could be said to refer to measurement of the mind.
2. *Guide to Understanding Credentialing Concepts* (NOCA/ICE; Durley, 2005):
   - psychometrician is “…an individual who normally holds a doctoral degree in measurement or a discipline of psychology (such as educational or industrial/organizational psychology) who can understand, apply, and describe the science and technology of mental measurement.”
   - Psychometrics: “The science and technology of mental measurement, including psychology, behavioral science, education, statistics, and information technology.”
3. *New York Times* article: “Psychometrics, one of the most obscure, esoteric and cerebral professions in America, is also one of the hottest.” (Herszenhorn, 2006)

Typical psychometrics presentation



5. BASIC PSYCHOMETRIC PRINCIPLES

Lawrence J. Fabrey
Applied Measurement Professionals, Inc.

Carol Hartigan
AACN Certification Corporation

Introduction

What do we want to measure? And how precise do we need to be? These two questions address the most basic descriptions of validity and reliability.

Reliability and validity are terms that, because they are often used together, have been assumed to have an identical meaning, but one of the first things that students may learn in a measurement class is that reliability and validity are really not the same. The phrase "reliable and valid examination" is so indelibly etched in our minds that it would be easy to assume that it is actually a singular concept. After these same students have learned that reliability and validity are different,

***Certification: An ICE Handbook* 2nd Edition**

3rd Ed. is due for publication later this year

psi®

Identifying the Intent of the Measurement Process

1. What do we want to measure?
2. And how precise do we need to be?

- “Reliability and Validity” are often noted together
- They are separate concepts:
  - # 1 addresses validity, # 2 addresses reliability
- Why is validity more important than reliability?

Can we claim "This is a valid examination!"

No -- but why not?

- The question suggests that validity can be evaluated with only one method,
- The inferences made about examination results are not addressed,
- A yes or no response would fail to account for other psychometric issues, and
- The purpose for which examination results are to be used must be described.

Validity Evidence for Certification

- What do we want to measure?
  - Knowledge, skills, abilities, attitudes…
- Why do we want to measure?
  - Protect the public or inform the public

- More on validity later.

Can we Claim a Test is Reliable?

- Maybe, but reliability is a matter of degree

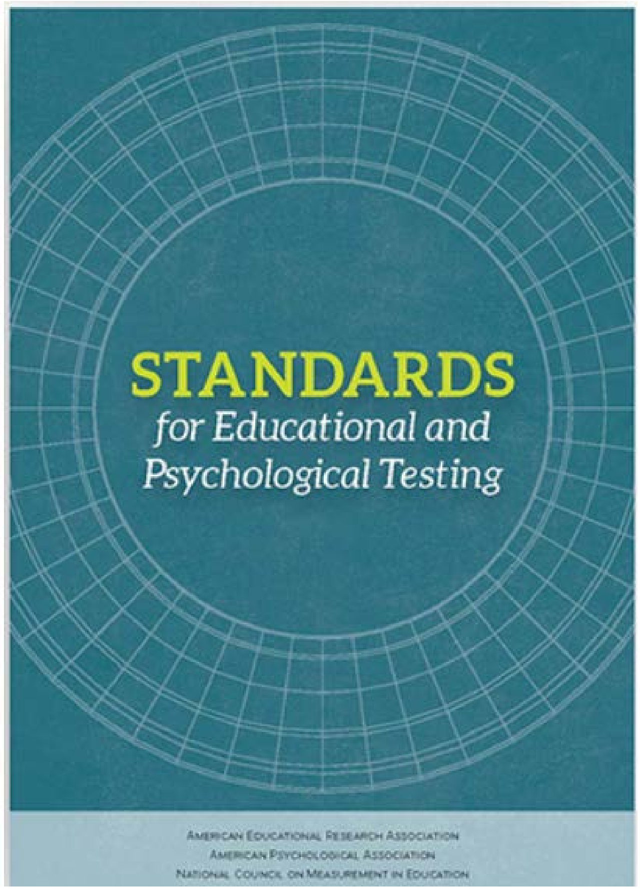
- More on reliability later.

psi

Psychometric Standards

- Several “standards” documents
- *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014).
- *NCCA Standards for the Accreditation of Certification Programs* (ICE, 2014), and
- ISO/IES 17024 (and others)

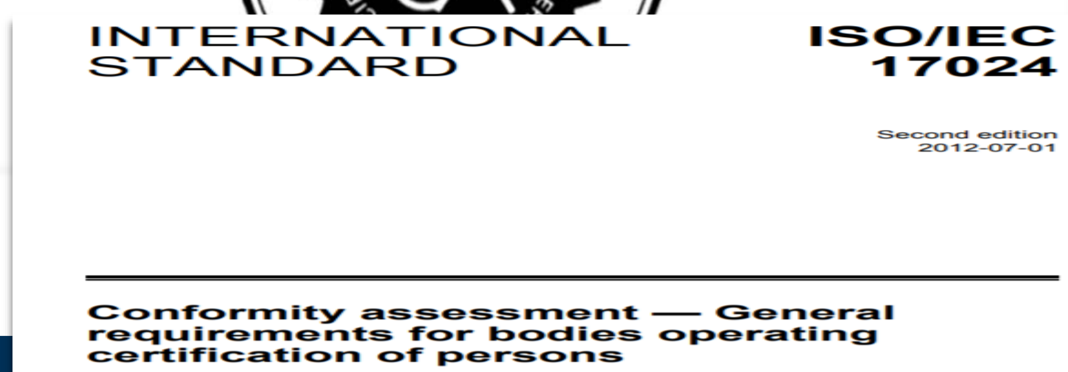
Standards


STANDARDS
for Educational and
Psychological Testing

American Educational Research Association
American Psychological Association
National Council on Measurement in Education


National Commission for Certifying Agencies

Standards for the
Accreditation of Certification Programs


INTERNATIONAL STANDARD

ISO/IEC 17024

Second edition
2012-07-01

Conformity assessment — General requirements for bodies operating certification of persons

Theories Of Measurement

- Two Models:
- **Classical Measurement Theory**
- **Item Response Theory**
- Statistics used for each model
- Choosing a Measurement Model

Classical Measurement Theory Model

- Usually called CTT (Classical Test Theory)
- Sampling items from a domain
- Each item assumed to make an equal contribution to measurement of the domain
- Scores: counting the number of correct answers
- Sometimes transformed to another scale, but
- Highest number correct will yield the best result

Classical Measurement Theory Model Statistics

- Key Item Statistics
  - Item difficulty identified by p value (proportion correct)
  - Item discrimination identified by a correlation (e.g., point-biserial)
- Test level statistics
  - Mean, standard deviation, etc.
  - Measures of reliability to be discussed later

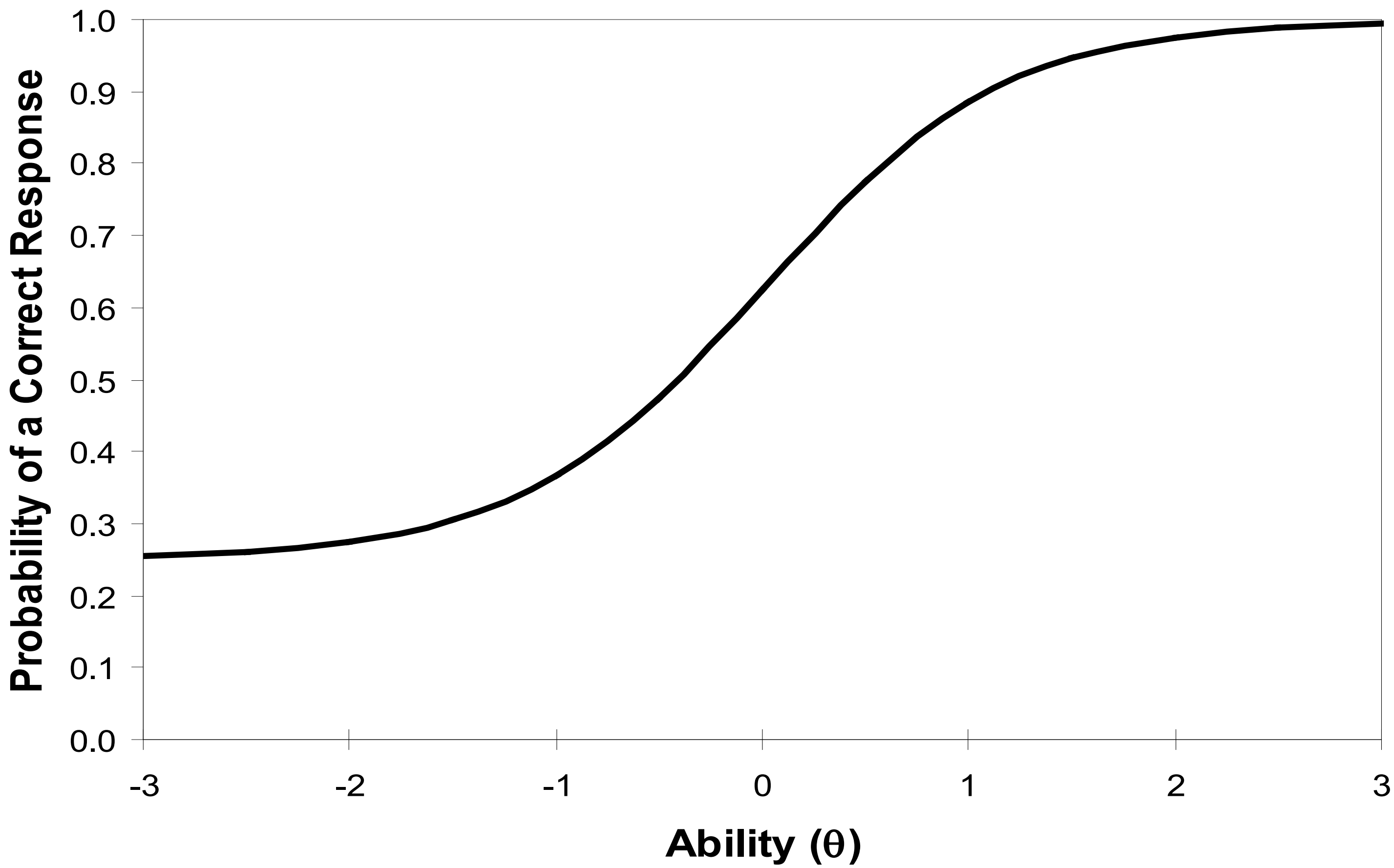
Item Response Theory Model

- Commonly called IRT
  - Sometimes called Latent Trait Theory
- Ability of examinee estimated based on the difficulty of the items with correct responses
- Items are calibrated
- Probability of a correct response is a function of the underlying ability of the examinee and the statistical characteristics of the item.

Item Response Theory Model Statistics

- Key Item Statistics – up to three parameters
  - Item discrimination = the a parameter
  - Item difficulty = the b parameter
  - Guessing = the c parameter
  - Depicted with an ICC
- Test level statistics
  - Test Characteristic Curves
  - Information functions

Sample Item Characteristic Curve (ICC)



Choosing a Measurement Model

- Both provide valuable information
- Interpretations for either benefit from larger candidate volume
  - Candidate volume is more critical when choosing number of parameters for IRT

Reliability - Overview

- "The degree to which the results of testing are free from errors of measurement"
- Methods of assessment
- Contributing factors
- How to promote higher reliability
  - Item writing
  - Statistical analysis of items
- Scaling and equating

Reliability - Methods of Assessment

- CTT:
  - Internal Consistency (Coefficient α, KR-20)
  - SEM
  - Generalizability (G) theory (based on ANOVA)
- IRT:
  - SEM
  - Fit
- Decision Consistency – for both models

Reliability – Contributing Factors

1. Homogeneity of content
2. Heterogeneity of examinees
3. Number and quality of items

How reliable is reliable enough?

NCCA Standard 20: "scores are sufficiently reliable for the decisions that are intended"

How to Promote Higher Reliability - Item Writing

- Training Item Writers to Ensure:
  - Pertinence to the examination
  - Clearly written stems
  - Avoidance of extraneous clues
  - Plausible, high quality distractors
  - Absence of bias
- Thorough review by SMEs

How to Promote Higher Reliability - Statistical Analysis of Items

- Pretest before using items to compute scores
- The goal is generally:
  - Higher positive discrimination
  - Moderate difficulty

How to Promote Higher Reliability - Example of Item Analysis

	Overall	Omits	A (True)	B (False)	C	D
Item	**Admins**					
Type	**p**					
Pts	**Avg**					
	rpb					
2 2						
1	164	0	44	7	**101**	12
MCS	0.62	0.00	0.27	0.04	**0.62**	0.07
1.00	74.00	0.00	69.47	64.91	**77.01**	70.67
	+0.390	-	-0.282	-0.197	**+0.390**	-0.096
1 54						
2	164	0	7	11	29	**117**
MCS	0.71	0.00	0.04	0.07	0.18	**0.71**
1.00	74.00	0.00	64.69	68.65	67.03	**76.79**
	+0.452	-	-0.202	-0.147	-0.332	**+0.452**
9 16						
3	164	0	14	**146**	2	2
MCS	0.89	0.00	0.09	**0.89**	0.01	0.01
1.00	74.00	0.00	68.40	**74.71**	67.20	68.80
	+0.205	-	-0.176	**+0.205**	-0.078	-0.059

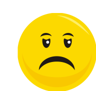
Appropriately difficult item, discriminates well (could be similar to item depicted by the previous ICC)

	Overall	A	B	C	D
n	192	**137**	13	19	23
p	0.76	**0.71**	0.07	0.10	0.12
Mean	74.29	**77.63**	63.57	62.07	70.56
Disc	+0.412	**+0.412**	-0.156	-0.317	-0.231

Easy item that does not discriminate well

	Overall	A	B	C	D
n	192	20	2	**166**	4
p	0.87	0.10	0.01	**0.87**	0.02
Mean	75.41	75.50	69.05	**75.61**	69.73
Disc	+0.049	+0.03	-0.059	**+0.049**	-0.086

Scaling and Equating

- **Scaling**
- A linear transformation of a number or score from one scale (usually a raw score or number correct) to another
- Examples: temperature, currency
- Simplest is calculating a percentage 
- Other scaling methods
- Advantages and disadvantages of scaling

Scaling and Equating

- **Equating**
- Statistical process for determining comparability of score interpretations based on different examination forms
- CTT: usually through common items (anchor test)
- IRT: usually through placing all parameter estimates from different samples of examinees on a common scale

psi®

Validity - Overview

- Evidence of validity based on content
- Criterion-Related validation strategies
- Item and test bias
- Establishing cut scores
- Summary of validity as applied to credentialing examinations

Types of Validity Evidence

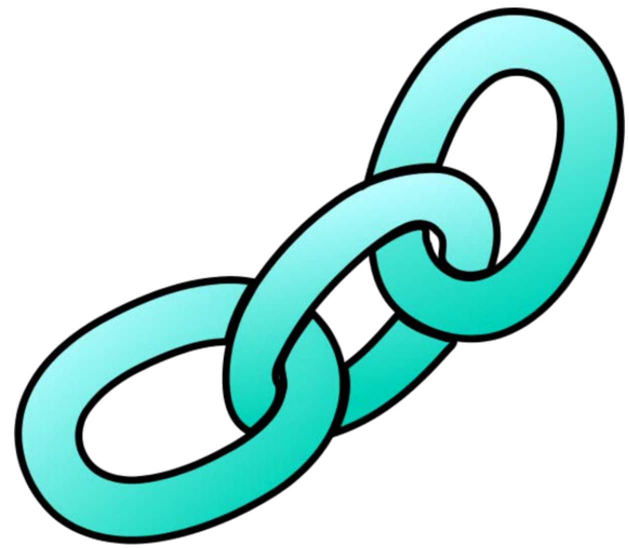
- Traditional: Content, Construct, Criterion-related
- 1999 *Standards:* “Validity is a unitary concept”
- 2014 *Standards*
- Sources of validity evidence based on:
  - Test Content
  - Response Processes
  - Internal Structure
  - Other Variables (i.e., convergent, discriminant, test-criterion relationships, generalization)
  - Consequences of testing

Evidence of Validity Based on Content

- Job analysis
  - Also known as practice analysis, role delineation study, or other terms
- The goal: determine what a practitioner must know and do in the role/job
- Collection of data (e.g., survey)
- Interpretation of data
- Leads to development of specifications

Links in the Chain of Evidence…

- Job Analysis (aka Practice Analysis, RDS)
- Examination Specifications
- Item Writing
- Examination Development
- Standard Setting
- Scoring



Criterion-Related Validation Strategies

- Less common for credentialing
- Relationships of scores with other measures
  - For example: job performance ratings or other assessments
- Challenges

Item and Test Bias

- Item bias
- Differential item functioning (DIF)
- Detection at two points in the testing process:
  - during examination development or
  - during the analysis of examination results.
- Prevention is the key

Establishing Cut Scores

- “Perfect” test; no value without appropriate passing point
- *NCCA Standard 17*: “standard setting study that relates performance on the examination to proficiency, so that the program can set a passing score appropriate for the certification.”
- Criterion related (and **not** norm-referenced)

Establishing Cut Scores

- Different methods may be considered
  - Principles can apply regardless of format
- Commonalities among usual methods:
  - Selection of judges
  - Agreement on MCP definition
  - Judgments about items
  - Reasonability check

Establishing Cut Scores

- Examples of different methods:
  - Angoff
  - Bookmark
  - Ebel, Jaeger, Nedelsky (not often used)
- Demonstration of a modified-Angoff method

Establishing a Cut Score - Angoff Method

- Let's pretend:
  - You are SMEs
  - A four item test has been approved*
  - We just discussed and agreed on the definition of an MCP (or minimally qualified candidate)
- Let's set a cut score for the Certified Chocolate Consumer (CCC) examination

*Items adapted from Hogan, Waters, Nettles, & Breyer (ATP, 2008)

Rate, then we'll check key

1. Which of the following foods would be the best to use as a palate cleanser before tasting chocolate?

   A. red wine
   B. tart apple
   C. lettuce leaves
   D. salted corn chips

2. The process of removing the outer shell from cocoa beans is called

   A. clicking.
   B. cloaking.
   C. guppying.
   D. winnowing.

3. Who is credited with creating the first chocolate bar?

   A. Joseph Fry
   B. Rudolf Lind
   C. Henri Nestle
   D. Milton Snavely Hershey

4. The process that is used to grade the quality of cocoa beans is called

   A. a cut test.
   B. a dry test.
   C. blanketing.
   D. kibbling.

Establishing the cut score

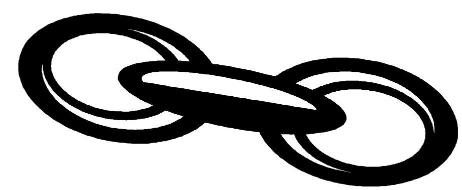
- Please announce your ratings
- Difficulty
- Small judge variability
- Suggested cut = 74%

  (or 3 out of 4)

	Items				Judge
Judge	1	2	3	4	Means
1	80	75	65	70	72.50
2	80	65	70	65	70.00
3	85	85	65	75	77.50
4	90	80	70	65	76.25
5	85	85	65	70	76.25
6	85	75	75	75	77.50
7	85	75	75	60	73.75
8	85	65	65	75	72.50
9	95	70	55	75	73.75
10	80	75	60	65	70.00
Mean	**85.00**	**75.00**	**66.50**	**69.50**	**74.00**

Summary of Validity as Applied to Credentialing Examinations

- Needed for any assessment method
- "Links in the Chain of Evidence Used to Support the Validity of Examination Results"
- Documentation



The End

- Questions?
- Follow up

Lawrence J. Fabrey, PhD
lfabrey@psionline.com (until 6/28)
After June 30:
LawrenceFabrey@gmail.com
913.980.7136