

TEST CODE

TEST FORM

BOOKLET NUMBER

999999

2005

CANDIDATE'S NAME

(Please Print)

DO NOT OPEN



**Silver Instruments:
25 Years of Experience
in Developing Tests**



***PRESENTATION DEVELOPED FOR
CLEAR 25TH ANNUAL CONFERENCE
PHOENIX, ARIZONA
PRE-CONFERENCE WORKSHOP
THURSDAY, SEPTEMBER 14, 2005***

Silver Instruments: 25 Years of Experience in Developing Tests



Silver Instruments: 25 Years of Experience in Developing Tests

Thursday, September 15, 2005
8:00 – 11:00 a.m.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Topics and Presenters

- **Job Analysis:** Reed A. Castle, Director of Research and Development, SMT
- **Item Development:** Kathi Gialluca, Director of Test Development, Pearson VUE
- **Forms Assembly:** Scott Thayn, Psychometrician, Thomson Prometric

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Topics and Presenters

- **Standard Setting:** Paul Naylor, Psychometric Consultant
- **Equating and Scaling:** Steven S. Nettles, Vice President of Research and Development, AMP

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Job Analysis

Reed A. Castle, Ph.D.
Schroeder Measurement Technologies, Inc.

What is a Job Analysis?

- An investigation of the ability requirements that go with a particular job (Credentialing Exam Context)
- It is the study that helps establish a link between test scores and the content of the profession.
- *The Joint Technical Standards 14.14*

"The content domain to be covered by a credentialing test should be defined clearly and justified in terms of importance of the content for the credential-worthy performance in an occupation or profession. A rationale should be provided to support a claim that the knowledge or skills being assessed are required for credential-worthy performance in an occupation and are consistent with the purpose for which the licensing or certification program was instituted."

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Why Conduct a Job Analysis?

- Need to establish eligible content for assessment
- Need to establish a validity link
- Need to reduce the threat of legal challenges
- Need to determine what is relatively important practice
- Need to understand the profession before we assess it
- Need to do it! (CLEAR, NCCA, ANSI)

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Types of Job Analyses

- Focus Group
- Traditional Survey-Based
- Electronic Survey-Based
- Transportability

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Focus Group

- Need to identify the best group of SMEs possible
 - Areas of Practice
 - Geographic Representation
 - Demographically Balanced
- 8 to 12

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Focus Group

Prior to Meeting

- Comprehensive review of profession
 - Job Descriptions
 - Performance Appraisals
 - Curriculum
 - Other job-related documents
 - State and Federal Laws
- Create an Exhaustive Master Task List
- Send list to SMEs prior to meeting to give them chance to review

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Focus Group

At Meeting

- Review Comprehensive Task List
- Determine which tasks are important
- Determine which tasks are performed with an appropriate level of frequency
- Determine which tasks are duplicative
- Identify and add missing tasks
- Organize into coherent outline

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Focus Group

- Advantages
 - May be only solution for new/emerging professions
 - Relatively quick
 - Less expensive

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Focus Group

- Disadvantages
 - Based on one group (Results may not generalize.)
 - May be considered a weaker model when considering validation
 - May result in complaints from constituents about the content of the test

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Traditional Survey-Based

- First steps are similar to the focus group (i.e., task list is generated in same manner).
- After the task list is created, three more issues must be addressed to complete the first survey development meeting.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Traditional Survey-Based

- First, demographic questions must be developed with two goals in mind.
 - Questions should help describe the sample of respondents.
 - Some questions will be used for analyses to help generalize across groups (e.g., geographic regions).

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Traditional Survey-Based

- Second, rating scale(s) should be developed.
 - Two pieces of information are meaningful.
 - Importance or significance
 - Frequency of performance
 - Additional scales can be added but may take away from response rate.
 - Shorter is sometimes better.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Traditional Survey-Based

- Sample Scale combining Importance and Frequency
- High correlation b/w Freq and Imp Ratings (.95 and higher)
- Considering both the importance and frequency, how important is this task in relation to the safe, effective, and competent performance of a Testing Professional? If you believe the task is never performed by a Testing Professional, please select the "Not performed" rating.

RELATIVE	ABSOLUTE
0 = Not performed	Not performed
1 = Minimal importance	Of no importance
2 = Below average or low importance	Of little importance
3 = Average or medium importance	Moderately important
4 = Above average or high importance	Very important
5 = Extreme or critical importance	Extremely important

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Traditional Survey-Based

- Third, a sampling strategy must be established.
 - One of the more important considerations is the sampling model employed.
 - Surveys should be distributed to a sample that is reflective of the entire population.
 - Demographic questions help describe the sample.
 - One should anticipate a low response rate (20%) when planning for an appropriate number of responses.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Traditional Survey-Based

- Mailing surveys
- Enclose a postage paid return envelope
- Plan well in advance for international mailings (can be logistically painful with different countries)
- When bulk mailed, plan extra time
- Keep daily track of return volume

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Electronic Survey-Based

- Identical to traditional, but delivery and return are different
- Need email addresses
- Need profession with ready access to Internet

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Electronic Survey-Based

- Advantages
 - Faster response time
 - Data entry is no longer needed.
 - Reduced processing time on R & D side
 - Possibly less expense (fewer administrative costs)
 - Can modify sampling and survey on the fly if needed
 - Follow-up is easier and quicker.
 - Sample can be the population with little additional cost.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Electronic Survey-Based

- Disadvantages
 - Need email addresses
 - High rate of "bounce-back"
 - Control for ballot stuffing
 - Data compatibility

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Transportability

- Using the results of another job analysis
- Determine compatibility or transportability
- Similar to Focus Group

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Four Types Review

- Focus Group
- Traditional Survey-Based
- Electronic Survey-Based
- Transportability

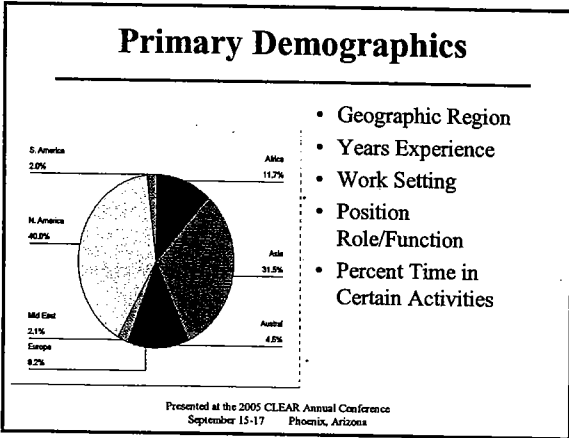
Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Data

- Demographics
- Importance Ratings
- Frequency Ratings
- Composite
- Sub group Analyses
- Decision Rules
- Reliability
 - Raters
 - Instrument
- Survey Adequacy

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests



Mean Importance Ratings

3.0 criterion

Task	Mean
Task 6	2.45
Task 4	2.97
Task 1	3.21
Task 5	3.85
Task 3	3.91
Task 7	4.25
Task 2	4.28

Out ↑
↓ In

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

% Not Performed Ratings

Criterion 25% (75% perform)

Task	% NP
Task 6	38%
Task 4	29%
Task 1	26%
Task 5	16%
Task 2	10%
Task 3	5%
Task 7	3%

Out ↑
↓ In

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Composite Ratings

- Composite ratings using rating scale Natural Logs (when multiple scales are used) can be calculated and combined based on some weighting scheme.
 - For example, if you want to weight frequency 33.33% and importance 66.66%, you can adjust for this in the composite rating equation.
- Personal opinion is that you will likely end up in a very similar place if establishing decision criteria on each scale individually.
- In addition, multiple decision rules are more conservative.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Mean Importance Sub-group Analyses

	Region						<=3.0
	Africa	Asia	Australia	Europe	N. America	S. America	
Task 1	3.22	3.12	3.01	2.96	3.21	3.18	1
Task 2	4.21	4.08	3.85	3.84	4.51	4.38	0
Task 3	3.91	3.87	3.78	3.75	3.48	3.25	0
Task 4	2.95	2.99	3.03	3.1	2.91	2.89	4
Task 5	3.88	3.82	3.84	3.89	3.78	3.48	0
Task 6	2.41	2.85	2.14	2.47	2.85	2.35	6
Task 7	4.22	4.09	3.85	3.84	4.47	4.25	0

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Assessment Type

- SMEs are asked to determine which assessment type will best measure a given task:
 - Multiple choice
 - Performance
 - Essay/short answer

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Cognitive Levels

- Each task on the content outline requires some level of cognition to perform.
- Three basic types of cognition exist (from Bloom's Taxonomy).
 - Knowledge/Recall
 - Application
 - Analysis

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Weighting

- Weighting is usually done with SMEs based on some type of data.
 - For example, average importance or composite rating for a given content area
- Applied to assessment type and cognitive levels

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Test Specifications/Weights

- Standard Exclusion/Inclusion criteria
- Assessment type/Cognitive levels
- Weights based on rational approach

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Item Development

Kathi Gialluca, PhD
Pearson VUE

Item Development

- What?
 - Determining what items need to be written

- How?
 - Item writing
 - Item reviews
 - Field testing of new items

- How managed?

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Development: What?

- Job/Task analysis & test plan (specifications)
 - Proportion of items on exam from each content area

- Assessment design
 - CAT or linear? How many forms? Scores?
 - Volume and frequency of testing?
 - Retest policy?
 - Item types?

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Item Development: What?

- Inventory of current item bank
 - How many items?
 - Distribution of content, item difficulty,...
- Item Development Plan
 - How many and what kinds of items need to be written?

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Development: How?

- Determine how process will be managed
 - In-person workshops vs. distributed
 - Payment for item developers?
- Select item writers and reviewers
 - Representative of the profession
 - Gender/Ethnicity/Geography
 - Tenure/Type of practice or specialty
 - Rotated periodically

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Writing

- Training
 - What makes a “good” item?
 - What types of items should be written?
 - What is writer expected to provide?
 - Item stem, response options, key
 - Validating reference
 - Content code(s)?
 - Rationale
- References

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Item Writing

- Style guidelines
- Assignments
 - Acceptable and not-acceptable content areas
 - Conformance with test plan
- Feedback/Review

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Reviews

- Test service review
 - Review/Edit items
 - Provide additional references
 - Provide additional coding
 - Explicit links to test plan and job/task analysis
 - Cognitive level?
 - Any coding or rationale not provided by item writer

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Reviews

- Editorial review
 - Ensure items are clear, consistent, and grammatically correct
- Technical review
 - Independent set of reviewers
 - Review without the key
 - Technically correct? Appropriate?

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Item Reviews

- Bias/Sensitivity reviews
 - Item *appearance*, not item *performance*
 - Not to stereotype, offend, defame, or patronize any individual or group
 - Ensure measurement of intended construct and not “noise”

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Reviews

- Bias/Sensitivity reviews (continued)
 - Inappropriate terminology
 - Generic: *he, spokesman*
 - Specific terms: *spouse or partner vs. husband or wife*
 - Idiomatic expressions: *soda/pop; kick the bucket*
 - Stereotypes
 - Women as intuitive
 - Asian Americans as refugees
 - People with disabilities as heroic victims
 - Older people as feeble and incompetent

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Reviews

- Bias/Sensitivity reviews (continued)
 - Inappropriate tone (elitist, patronizing, inflammatory):
 - *Lady doctor, feisty senior citizen; coed*
- Test service and/or sponsor review
 - Final review before items are field-tested

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Field Testing of New Items

- To determine how items are performing
 - % of total test length
- Options
 - Embedded in test form
 - Not identified or scored for test-takers
 - Content areas balanced re blueprint
 - Evaluate/Eliminate items first, and then score exams
 - Separate test administration

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Field Testing of New Items

- Statistical evaluation of new items
 - Pre-determined criteria for accepting or rejecting new items
 - Item difficulty, item discrimination, DIF,...
 - Select “best” 100 items for exam
 - Content and statistical considerations

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Item Development: How Managed?

- Biggest challenges:
 - Knowing which items are in which stages of development
 - Tracking item revisions/version control
- Formal Item Banking System

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Item Development: How Managed?

- Contents
 - Item text (+key)
 - Item status
 - Test plan codes, other content codes
 - Validating references, rationale, or comments
 - Item statistics
 - Item writer information
- Reporting capability

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Forms Assembly

Scott Thayn
Thomson Prometric

What are Forms?

- Measure the same construct
- Meet the same blueprint requirements
- Different versions of an exam
- Comprised of both unique and shared items

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Why have Multiple Forms?

- Security
- Retake
- Reduce item exposure

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Equating

- Post Administration Equating
 - Forms are assembled to blueprint specification.
 - After the forms are administered, statistical analysis is performed and the scores equated to compensate for differences in the forms.
- Pre-Equating
 - Data gathered by a beta (try-out) test
 - Data then used to build parallel (balanced) fixed forms or other assembly algorithms

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Parallel Forms

Balanced in:

- Content (based on the blueprint)
- Difficulty
- Discrimination
- Reliability (for classical test theory)
- Standard Error of Measurement (for IRT)
- Variance
- Time

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Forms Assembly

- Classical
 - Fixed Forms
- Item Response Theory (IRT) and Rasch
 - Linear On the Fly Testing (LOFT)
 - Computer Adaptive Testing (CAT)
 - Computer Mastery Testing (CMT)
 - Fixed Forms
- Hybrid
 - Combining Rasch and Classical

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Setting the Passing Standard

Paul D. Naylor, Ph.D.
Consultant

Standard Setting

- The process used to arrive at a passing score
- Lowest score that permits entry to the field
- *Recommended* standard

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Standards

- ◆ Mandated
- ◆ Norm-referenced
- ◆ Criterion-referenced

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Mandated Standards

- Often used in licensing
- Difficult to defend
- Not related to minimum qualification

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Norm-referenced Standards

- Popular in schools
- Limits entry
- Inconsistent results

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Criterion-referenced Standards

- Wide acceptance in professional testing
- Determines minimum qualification
- Not test population dependent
- Exam or item centered

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Minimally Competent Performance

- Minimum acceptable performance
- Minimal qualification
- Borderline
- It's all relative.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Procedures

- Angoff (modified)
- Nedelsky
- Ebell
- Others

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Angoff Method

- Judges
 - Selection
 - Training
- Probabilities
- Would vs. Should
- Rater agreement
- Tabulation

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Ratings

AV	1	2	3	4	5	6
68.3	60	75	70	50	80	75
52.5	60	50	45	45	55	60
72.5	85	80	75	70	70	55
89.2	95	95	90	80	85	90
76.7	75	70	90	70	75	80
77.5	75	70	85	80	85	70
72.78	75.0	73.33	75.83	65.83	75.0	71.66

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Application

- Adjustment
- Angoff Values
- Alternative Forms of Exam
- Passing Score

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Validity Evidence

- How was the passing score set?
- Who set the passing score?

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Minimum Requirements

- Description of method
- Results of method used
- List of participants and qualifications
 - Demographics
 - Credentials
 - Affiliations

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Additional Evidence

- Data from method used
- Description of training given to participants

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

References

- Livingston, S.A. & Zieky, M.J.(1982). *Passing Scores*. ETS.
- Cizek, C.J. Standard setting guidelines. *Educational Measurement: Issues and Practices*, 15 (1), 13-21, 12.
- *CLEAR Exam Review* (Winter 2001, Summer 2001 and others)

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Test Score Equating and Scaling

Steven Nettles, EdD
Applied Measurement Professionals, Inc.

Introduction

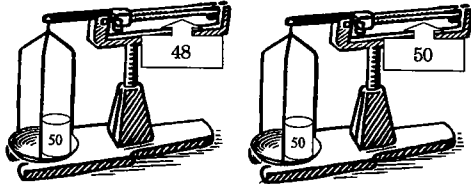
- Lord, F.M. (1977)
 - It should not matter to candidates which version of a test they take.
- Kolen and Brennan, 1995
 - Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably.
 - Equating adjusts for differences in difficulty in test forms that are built to be similar in difficulty and content.
- Primary motivations to equate
 - Treat candidates fairly
 - Maintain the meaning of a credential

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Equating Rationale

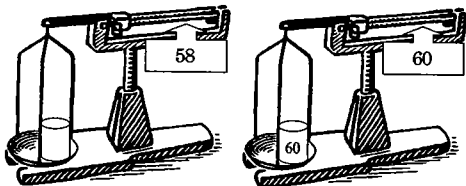
What would you do if these two scales showed two different values for the same object?



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Equating Rationale

What would you do if these two scales showed two different values for two different objects?



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Common Equating Models

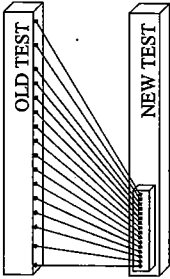
- Common item
 - Necessary when a test includes new items
 - Typically uses post-administration methods
- Pre-equating
 - Necessary when an organization wants to release results instantly to candidates
 - Requires that every item selected for the test was previously used

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Common Item Equating

Concept

- Items from an old test become part of the new test.
- Comparing values earned by different groups on the common item set permits a comparison of these groups.



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Comment

- A common item equating plan should directly affect test development.
 - The plan is more than post-test statistical manipulations.
 - Equating success requires carefully followed test development procedures.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Test Development

- Common items are selected to mirror content and statistical properties of the whole test - the original scale.
 - Content domains are represented in equal proportions as the whole test.
 - Difficulty and discrimination properties of items are matched to the whole old test and the whole new test.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Test Development

Equator Block / Whole Test Specifications				
Content	Cognitive Levels			Total
	Recall	Application	Analysis	
1	3 / 11	3 / 14	1 / 4	7 / 29
2	2 / 9	2 / 9	0 / 0	4 / 18
3	2 / 6	4 / 14	2 / 10	8 / 30
4	2 / 7	3 / 13	1 / 3	6 / 23
TOTAL	9 / 33	12 / 50	4 / 17	25 / 100
Target Means: $p=.75$, $r_{pb}=.30$				

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Inferences

- By comparing responses from two groups to the same items, one can confidently compare abilities.
- Once candidate and test properties are known, the appropriate equating adjustment for new test scores can be made.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

The Adjustment

- Assuming values on both scales are linearly related to one another and candidate groups are the same, the appropriate adjustment can be derived: $y = ax + b$

where

- $a = SD_Y / SD_X$
- x = a scale point from the old test
- b = mean $y - a(\text{mean } x)$
- y = a scale point for the new test



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

An Example

- An old test (x) had a mean=80, $SD=10$, and a cut score=75.
- A new test (y) shows a mean=70 and $SD=9$.
- What is the equitable cut score for the new examination?

$$y = (9/10) * (75 - 80) + 70$$
$$y = 65.5$$

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Group Differences

- When the test administration plan permits evaluation of group differences, the magnitude of difference is measured.
- Equating formulas accounting for group differences are substantially more complex than the example illustrates.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Equating Formulas and Assumptions

- The Tucker formula has strong similarity assumptions for candidate groups.
- The Levine formula and IRT based equating only assume both sets of test scores express the same construct.
 - These two are more appropriate when candidate groups are expected or known to be somewhat different.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Assessing Group Differences

- Mean scores for new and old test groups are compared for the common item set.
 - If the difference is less than 0.25 SD, then groups are similar enough to use the Tucker result; otherwise, a Levine or IRT result is recommended.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Assessing Group Differences

- The variance of new and old test group scores for the common item set is compared.
 - If the ratio of equator score variances is between 0.8 and 1.25, then groups are similar enough to use the Tucker result; otherwise, a Levine or IRT result is recommended.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Equating Precision

- The common-item block should be at least 20% of the new test or 20 items, whichever is larger.
 - Assumes total test length is at least 40 items
- The example specifications show 5 extra items were included among common items.
 - Some items may need to be dropped, but we still have sufficient equating precision.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Equating Precision

- When candidate groups are non-random and more likely to slightly differ, content representation is critical.
 - Non-random groups with slight differences in ability describe a typical credentialing testing scenario, so content representation during item selection is a requirement.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Equating Precision

- Difficulty differences between the common-item block and the rest of the test vary directly with equating error.
 - To minimize equating error, closely match statistical properties of an equator block to the whole test
 - Match mean item properties
 - Match item statistic distributions
 - Categorize test statistics by ranges of p-value and r_{pb}

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Equating Precision

- Larger candidate samples are better.
 - Minimum published values for samples are about 100-200 in each group.
 - First time candidates' responses are best.
 - Larger samples typically result in less error.
- General Rule
 - Common items should appear in approximately the same location on new and old tests to control for an order effect.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Sample Standard Error of Equating

- Given the following:
 - Old test cut-score = 56; Old test mean = 60
 - Old test and new test have SDs = 8
- If sample sizes = 50 each, SEE = 1.7
- If sample sizes = 150 each, SEE = .98
- If sample sizes = 1500 each, SEE = .31
- If sample sizes = 15000 each, SEE = .1

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Linkage Plans

- Chain
- Base
- Hybrid
- Double

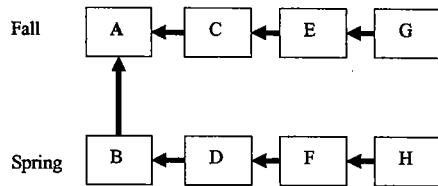
Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Linkage Plan Guidelines

- Minimize equating lineages
 - Link to a form in the same season
 - Minimize the number of links back to the initial form
 - Minimize links to the same form
- But - no one plan can satisfy each of these rules, so compromises must be chosen.**

From: Kolen MJ and Brennan RL, p. 260
Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Chain Linkage Plan



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Chain Linkage Plan

- Best ensures similar groups are the object of each equating
 - Intervals between administrations are short.
 - Groups from the same season are compared.
- However, accumulated equating error is maximized toward the end of each chain.
 - To diminish: frequently re-standardize and the restart the linkage plan
- Unwanted equating lineages also are most likely under this scenario.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Base Linkage Plan

```
graph TD; B --> A; C --> A; D --> A; E --> A; F --> A;
```

- Equating lineages are eliminated by this plan.
- But several linkages are made to the same form further and further removed across time.
- It may become difficult to continue to find common items with Form A that persist over time.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Hybrid Linkage Plans

- Many potential hybrid plans could be developed to
 - minimize
 - equating lineages
 - accumulated equating errors
 - encourage
 - shorter intervals between administrations
 - similar seasons

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Hybrid Plan Example

```
graph TD; subgraph Fall; A; C; E; G; end; subgraph Spring; B; D; F; H; end; A --> B; C --> D; E --> F; G --> H; C --> F; E --> B;
```

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

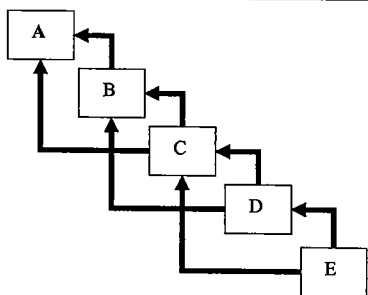
Silver Instruments: 25 Years of Experience in Developing Tests

Double Linkage Plans

- May be useful to mitigate effects of equating strains/lineages
 - Two acceptable equating results can be averaged.
- A contingency is available should an equating link break.
 - Two candidate groups may be too different.
 - A form could be compromised.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Double Linkage Example



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Pre-Equating

Silver Instruments: 25 Years of Experience in Developing Tests

Concept

- If measurement components are known and consistent, then consistent, similar measurement instruments can be developed.
- If one has confidence in item characteristics, then consistent, similar scales are likely.
- Scores may be released instantly.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Methods

- In this context, *similar* implies test scores project equivalent information about candidates' abilities at the predetermined passing point.
- Item statistics are used to build tests with similar properties.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Test Development

- A base form serves as the model when selecting items for the new test forms.
 - The new form should project scores:
 - centered at the same point as the model
 - that vary to the same degree as the model
 - that are as precise as the model

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Test Development

- Because items selected for forms must all be previously used, pre-equating plans typically incorporate some version of pre-testing.
 - Items that are exposed to candidates:
 - can be embedded in with scored items
 - or can use proxies in a field-testing environment

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Scaling

Concept

- Two scales may be used to express the same quantity (e.g., °F, °C).
- Credentialing tests typically hold characteristics, including the cut point, of one (standard) scale constant while characteristics of a raw scale vary.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

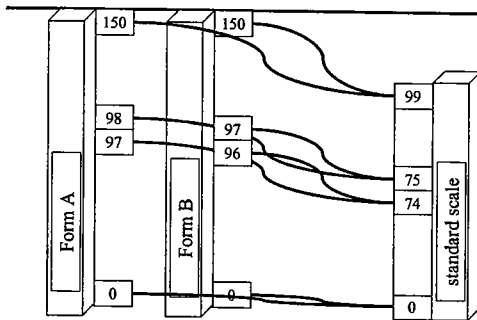
Silver Instruments: 25 Years of Experience in Developing Tests

Scaled Scores

- Variable raw cut scores may be the product of an equating plan.
- The challenge is to communicate to stakeholders that a different raw cut was applied to hold the passing standard constant.
 - AERA, APA, NCME Standard 4.3 encourages anticipation of misinterpretations.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Raw to Scaled Conversions



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

When are equating and scaling inappropriate?

- Candidates differ too much
 - Too much time between administrations
 - Candidate preparation changes
- Tests differ too much
 - AERA, APA, NCME Standard 4.16 states test scores should be rescaled when test specifications change.
 - This implies a new passing point study should be conducted with every new job (practice) analysis.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Silver Instruments: 25 Years of Experience in Developing Tests

Resources

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for education and psychological testing. Washington DC. AERA.
- Kolen, MJ and Brennan RL. (1995). Test Equating Methods and Practices. New York, Springer.
- Shea, JA and Norcini JJ. (1995). Equating in Licensure Testing: Purposes, Procedures, and Practices. Editor Impara JC. Lincoln, NE. Buros Institute of Mental Measurements

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona



**Silver Instruments:
25 Years of Experience in Developing Tests**

QUESTIONS & ANSWERS



**Silver Instruments:
25 Years of Experience in Developing Tests**

***THANK YOU
FOR YOUR ATTENTION***

ENJOY THE CONFERENCE

Silver Instruments: 25 Years of Experience in Developing Tests

Presentation Follow-up

- Please pick up a handout from this presentation -AND/OR-
- Please leave your business card to receive an email of the presentation materials -OR-
- Presentation materials will be posted on CLEAR's website.

Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Speaker Contact Information

Reed A. Castle, PhD
Director, Research and Development
Schroeder Measurement Technologies, Inc.
2494 Bayshore Blvd, Suite 201
Ph: 727-738-8727
Fax: 727-734-9397
rcastle@smtest.com
www.smtest.com

Kathleen A. Gialluca, PhD
Pearson VUE
Ph: 952-681-3856
kathleen.gialluca@pearson.com
www.pearsonvue.com



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona

Speaker Contact Information

Scott Thayer
Thomson Prometric
350 South 400 West
Lindon, Utah 84042
Ph: 801-852-4307
Fax: 801-852-4142
scott.thayer@thomson.com
www.prometric.com

Paul D. Naylor, Ph.D.
Psychometric Consultant
3508 Manford Drive
Durham, NC 27707
Ph: 919-730-1593
Fax: 919-489-8224 (fax)
naylorpaul@msa.com

Steven Nettles, EdD
Vice-President, Research and Development
Applied Measurement Professionals, Inc.
8310 Nieman Road,
Lenexa, KS, 66214
Ph: 913-495-4405
snettle@goamp.com
http://www.goamp.com



Presented at the 2005 CLEAR Annual Conference
September 15-17 Phoenix, Arizona
