Fordham Council on Applied Psychometrics Workshop

Assessment Operationalization

with Michael Chajewski

Saturday, November 23, 2024
Fordham University Lincoln Center, Loen Lowenstein Rm. 612

Motivation

Operationalizing an assessment typically entails activities generally not covered by most psychometrics, quantitative psychology, and/or educational measurement graduate program curricula. Additionally, the nuances of legislative, client and user requirements often necessitate non-standard ("classroom") solutions. The skills required to navigate operational challenges are difficult to acquire artificially, and regularly dismay entry level psychometricians. Therefore, it is my hope that by participating in this workshop individuals who may consider a career/work in operational psychometrics will gain insight into the dynamics and workflows within which they will be expected to contribute.

Description

From supporting blueprint development to scaling decisions for score report production, many of the operational enterprise milestones--that assure the successful launch and propagation of an assessment system--rest on a priori psychometric groundwork. The workshop introduces attendees to the sequential decisions required to develop an assessment and provides an opportunity to explore a selection of standard activities that would be supported by operational psychometricians. Using assessment data, attendees will process and analyze various solutions in R, learning how psychometric deliverables are negotiated, and how assessment measurement expertise supports client and policy decisions.

Objectives

The workshop is designed to provide attendees with an opportunity to engage with a cross-section of tasks required to operationalize assessments. Specifically, an engaged participation in the prepared demonstrations will provide learners with the applied skills to…

- … identify and recommend item characteristics and scoring features to meet assessment design measurement objectives
- … establish a Rasch raw-score to theta (latent trait) conversion table (RST)
- … decide among common score transformation methodologies to produce and evaluate reportable scale scores
- … conduct criterion referenced comparisons
- … setup and execute an embedded field-testing design
- … understand where test construction form statistical targets originate

Agenda

The below provided agenda is an approximation. Depending on participant's interests, questions, and resulting discussions the timeline may require modification.

Section I – Setup	
1:00 – 1:15 PM	Introduction and materials
1:15 – 1:30 PM	Operationalization workflow
Section II – Follow-along Demonstration	
1:30 – 2:15 PM	Part 1: Blueprint, assessment, and field-test plans
2:15 – 3:00 PM	Part 2: Scaling and field-test calibration
3:00 – 3:30 PM	Break
Section III – Practical Application	
3:30 – 4:45 PM	Exercise 1: The test is too long Exercise 2: Scaling alternatives Exercise 3: Item bank sustainability
4:45 – 5:00 PM	Questions, comments and closing remarks

Software

Winsteps	The IRT calibration software, which requires a license, is not necessary but is needed if participants want to run the calibration analyses at a later time. Winsteps outputs are provided.
R/RStudio	All demonstrations and examples will be presented in RStudio.

Materials

All workshop necessary materials can be obtained in one of three ways:

1. Pull the public read-only GitHub repository, ASSESS_OP_20241123, from: https://github.com/chajewski/ASSESS_OP_20241123.git
2. Download the materials zipped archive from: https://www.chajewski.com/Teaching/Workshops/ASSESS_OP_20241123.zip
3. Copy the materials from the USB flash-drive circulating in the workshop room.

Homework

Find the operationalization workflow details for a current state-wide summative assessment. Look for technical manuals, agency references and assessment program resources (i.e. https://tea.texas.gov/student-assessment/tec-and-tac-references-for-the-texas-assessment-program).

Operationalization Workflow

The herein presented generalized common operationalization workflow, for performance evaluation systems typical of large-scale state-wide summative assessments, reflects the principal efforts for which psychometricians are consulted/provide deliverables.

Since the workflow is presented from the perspective of psychometric dependencies, non-psychometric tasks and deliverables are not included. While most of the provided workflow is generalizable to the development and implementation of high-stakes assessments, different assessment systems will address varying specific needs, making it impossible to provide an exhaustive list. As requirements are clarified and tailored, and different vendors ascribe different scope of work and ownership to their respective psychometric staff, additional incidental and ancillary workflow tasks are highly probable.

While the workflow is broadly a linear process, many decisions may require circular alignment confirmations, or may depend on other tasks' deliverables to be completed simultaneous. Usually, the client makes all final decisions on the basis of recommendations from the psychometrics, asset development, and operations/program cross-functional teams.

1. **Construct**
    a. Identify purpose (i.e. evaluating skills and knowledge, meeting legislative requirements, etc.)
    b. Establish content, curriculum and performance standards
    c. Define performance, proficiency, or achievement levels
    d. Confirm assessed curriculum (not to be confused with scope and sequence)
    e. Derive student (specific skills and knowledge) expectations
    f. Decide on synchronicity of general and alternate student population testing
    g. Collect assessment policies (i.e. data retention, AI usage, etc.)
2. **Test Design**
    a. Adopt assessment (i.e. evidence-centered design, etc.) and measurement (i.e. CTT, IRT, etc.) frameworks
    b. Identify usage (i.e. on-track progress measures, graduation requirements, school and district accountability, etc.)
    c. Collect relevant timely input and feedback (i.e. from educator focus groups, etc.)
    d. Conduct cognitive laboratory studies (i.e. think-aloud research for novel item types, cognitive load, and comprehension evaluations, etc.)
    e. Select test delivery (i.e. online linear form, multi-stage adaptive, single-item computerized adaptive testing, etc.)
    f. Determine performance outcomes and consequential decision dissemination process (i.e. on-demand score reports, online portals, paper reports, etc.)
3. **Blueprint Development**
    a. Review standards, content and task alignments
    b. Find number of items and item types (i.e. MCQs, TEIs, CRs, etc.)

c. Determine reportable outcomes (i.e. test-level, reporting categories, etc.)
d. Evaluate testing time
e. Determine scoring processes (i.e. machine scored, automated engine scoring, human scoring, double scorers, etc.)
f. Develop scoring rubrics (i.e. single- and multi-points, trait scoring, weighted, etc.)
g. Conduct exploratory simulations (i.e. item exposure rates, routing decisions, etc.)
h. Pilot
   (Ideally conducted on representative sample prior to launch; at-risk pilot evaluations are sometimes completed on launch data)
   i. Check score attainability and rubric functionality
   ii. Preliminary item and test level analyses
   iii. Dimensionality confirmation
   iv. Speededness checks (i.e. time-on-task, total testing time, etc.)
   v. Explore feature usage (i.e. tutorial items, highlighting, process mining, etc.)
   vi. Reliability and measurement error confirmation
   vii. Collect validity evidence (i.e. MTMM, etc.)

4. Implementation tasks

a. Outline standards adoption vs. assessed timeline
b. Plan for field notification and sampler requirements
c. Understand the need for a bridge-year / transition period
d. Support development of engagement definitions (i.e. attemptedness rules, minimum reporting requirements, etc.)
e. Identify launch year waivers
   i. Delayed score reporting
   ii. Suspension of progress measures and accountability
   iii. Volume of alternate student population tested (<= 1%)
   iv. Etc.
f. Integrate eligibility definitions and processes
   i. Accessibility features (i.e. synchronous classroom supports, etc.)
   ii. Locally designated supports (i.e. content and language supports, braille/refreshable braille, etc.)
   iii. State designated supports (i.e. extra time, mathematics scribe, etc.)
g. Establish scale propagation plan
   i. Equating design (i.e. randomly equivalent, NEAT aka CINEG, etc.)
   ii. Equating method (i.e. pre-smoothed equipercentile, IRT based, etc.)
   iii. Scale monitoring
      1. Equating error
      2. Anchor representativeness and stability definitions
      3. Drift evaluation plan (i.e. form chaining, etc.)
h. Field-testing design
   i. Review the item development plan
   ii. Propose testing/exposure design (i.e. fully embedded, stand-alone field-test studies, etc.)

iii. Select calibration design and method
iv. Create an item bank sustainability analysis plan

5. (Annual) **Operational processes**

a. Create assessment plan
    i. Set/confirm test administrations (i.e. make-up test dates, etc.)
    ii. Confirm accommodations / accessibility solutions (i.e. paper forms, braille, ASL translation)
    iii. Outline non-English form requirements
        1. Transadapted (i.e. grade 4 math, grade 5 science, etc.)
        2. Native development (i.e. grade 2-5 reading language arts, etc.)
    iv. Determine required number of unique forms per administration
    v. Set/confirm item and form reuse schedule
    vi. Set/confirm item and form release schedule
    vii. Set/confirm form vaulting / archiving schedule
b. Pre-administration requirements
    i. Review registration and eligibility confirmation volumes
    ii. Conduct form spiraling
    iii. Develop sampling designs
    iv. Develop scoring and calibration specifications
    v. Conduct mock/dry-runs and user acceptance testing (UAT)
c. In-window (during administration) requirements
    i. Monitor testing integrity (i.e. testing irregularity investigation, etc.)
    ii. Conduct on-site audit expectations
    iii. Support key-check and adjudication
    iv. Performance scoring (i.e. machine scoring, assigning score codes, etc.)
    v. *(Where applicable)* Conduct client scoring progress reviews (i.e. rater agreement, etc.)
d. On-time (just following the administration) requirements
    i. Score certification
    ii. *(Where applicable)* Conduct post-equating with quality assurance plan (i.e. third party replicator, etc.)
    iii. Deliver conversion tables, impact information, and approval forms
    iv. Release score / report production

6. (Only after the launch administration) **Scaling**

a. Calibrate the base (scaling) form
b. Confirm measurement scale properties
c. Performance standard setting
    i. Standard setting plan (viz. method, schedule, approval process, etc.)
    ii. *(Where applicable)* Map historic cut scores
    iii. Hold standard setting meetings
    iv. Obtain client approval for panelist cut score recommendations
    v. Disseminate findings (i.e. executive summary, full report, etc.)
d. Derive cut scores
e. *(Where applicable)* Conduct vertical scaling

f. Explore transformations for score reporting
g. Derive scaling constants
h. Set scale truncations (i.e. LOSS/HOSS, etc.)
i. Develop test construction specifications
    i. Review scaling form
    ii. Confirm blueprint ranges
    iii. Determine item difficulty distribution target ranges
    iv. Check form specifications (and targets) against field-test plan

7. Post-administration support

a. Rescoring / score arbitration support
b. Item data review
    i. Operational and field-test item analysis
    ii. Set item flags (i.e. low rbis, DIF, etc.)
    iii. Update item banks (i.e. post-equated IRT parameters, admin/cohort CTT statistics, etc.)
c. Developing technical manual/report/digest
d. Finalize state-wide cohort level data
e. Provide federal peer review critical element evidence
f. Support technical advisory committee (TAC) meetings
g. Conduct product research and ad-hoc studies

Part 1:
Blueprint, assessment, and field-test plans

Blueprint

In Table 1 below the core structural components of the operational base form are provided. Forms are allowed to vary year-over-year in difficulty and composition but need to fundamentally measure the same construct via demonstrable item alignments.

The scaling form is simply the first operational (OP) base form with which the assessment program launches its administration. Because many subsequent measurement components are derived based on the student performances on the scaling form, it is in the program's best interest to construct a representative and best-example test form administered to the target population of interest for whom the assessment is intended.

Assessment Plan

The assessment plan typically consists of a repository / database that tracks *specific* forms and their item composition year-over-year. The single record in Table 2 shows the reporting category (RC) item breakdown as well as the student assessment experience with respect to field-test (FT) items. The assessment plan is intended to communicate the administration specific student testing experience. In the grade 8 social studies demonstration test, each 8th grader is expected to answer 39-40 items (33 OP + 6-7 FT items), with a total testing time of approximately 40-60 minutes.

Table 1. Blueprint item and point distributions

	Items				Points							
					MCQ		TEI		SCR		Total Possible	
	MCQ	TEI	SCR	Max	Min	Max	Min	Max	Min	Max	Min	Max
RC1	8-12	0-1	--	13	0	12	0	2	--	--	0	14
RC2	8-12	0-1	--	13	0	12	0	2	--	--	0	14
RC3	6-10	0-1	--	11	0	10	0	2	--	--	0	12
RC4	--	--	1	1	--	--	--	--	2	8	2	8
OP	30	2	1	33	0	30	0	4	2	8	2	42

Note. RC = Reporting category; MCQ = Multiple-choice question; TEI = Technology enhanced item; SCR = Short constructed response; Item ranges are compensatory, with the OP row reporting the operational base form total number of items and points.

Table 2. Assessment plan for the (scaling) base form administration

	Items				Points								FT Items*		
					MCQ		TEI		SCR		Total				
	MCQ	TEI	SCR	Total	Min	Max	Min	Max	Min	Max	Min	Max	MCQ	TEI	SCR
RC1	11	--	--	11	0	11	--	--	--	--	0	11	--	--	--
RC2	9	1	--	10	0	9	0	2	--	--	0	11	--	--	--
RC3	10	1	--	11	0	10	0	2	--	--	0	12	--	--	--
RC4	--	--	1	1	--	--	--	--	2	8	2	8	--	--	--
Total	30	2	1	33	0	30	0	4	2	8	2	42	5	0 or 2	0 or 1

*Note. RC = Reporting category; MCQ = Multiple-choice question; TEI = Technology enhanced item; SCR = Short constructed response; -- indicate the blueprint value is not applicable for the specific form; * Every field test form consists of 5 MCQ items in addition to either 2 TEIs or 1 SCR. No student will receive both TEI and SCR field-test items.*

Field-test Design

Field testing enables test construction of future test forms. Its design is operationalized through the delivery mechanism by which student will be exposed to items. In the demonstration, the field-test design consists of a fixed field-test form, whereby the static operational base form every student will receive is paired with a specific selection of field-test items. Thusly, the field-test form is the unique combination of operational items and specific field test items.

The field-test design usually contributes substantively to core aspects of an assessment program such as: (1) the item bank sustainability, (2) sample size requirements, (3) the student testing experience (i.e. interaction type sequencing, total testing time, etc.), (4) test construction and (5) the pre-equating / propagation of the measurement scale. Table 4 shows how the total FT item pool of 41 items could be distributed across 10 FT forms.

Tables 3 and 4 demonstrate how FT forms are spiraled within the annual administration in order to meet minimum item calibration sample size requirements. In this assessment example, a minimum 500 students are required per MCQ, 750 students per TEI and 1,000 students for the SCR items. The increasing sample size per item type is designed to assure that ability distributions are observed across all possible item type obtainable raw score points—without which an item cannot be properly IRT calibrated to meet blueprint requirements.

Table 3. Example of a field-test form item delivery (or item assignment) matrix (1 = administered to student, 0 = not applicable).

	Operational Base Form						Field-test Items											
	MCQ			TEI		SCR	MCQ					TEI				SCR		
Student	OP1	[…]	OP30	OP31	OP32	OP33	FT1	FT2	[…]	FT34	FT35	FT36	FT37	FT38	FT39	FT40	FT41	
0001	1		1	1	1	1	1	1		0	0	1	1	0	0	0	0	
0002	1		1	1	1	1	1	1		0	0	1	0	1	0	0	0	
0003	1		1	1	1	1	1	1		0	0	1	0	0	1	0	0	
0004	1		1	1	1	1	1	1		0	0	0	1	1	0	0	0	
0005	1		1	1	1	1	1	1		0	0	0	1	0	1	0	0	
[…]																		
3496	1		1	1	1	1	0	0		1	1	0	0	0	0	1	0	
3497	1		1	1	1	1	0	0		1	1	0	0	0	0	1	0	
3498	1		1	1	1	1	0	0		1	1	0	0	0	0	1	0	
3499	1		1	1	1	1	0	0		1	1	0	0	0	0	0	1	
3500	1		1	1	1	1	0	0		1	1	0	0	0	0	0	1	

Note. MCQ = Multiple-choice question; TEI = Technology enhanced item; SCR = Short constructed response; OP# = Operational base form item; FT# = Field-test item.

Table 4 .Field-test form item membership matrix (1 = assigned to form, 0 = not on form).

		MCQ																				TEI				SCR	
Form	n	FT1	[…]	FT5	FT6	[…]	FT10	FT11	[…]	FT15	FT16	[…]	FT20	FT21	[…]	FT25	FT26	[…]	FT30	FT31	[…]	FT35	FT36	FT37	FT38	FT39	FT40	FT41
1	250	1		1	0		0	0		0	0		0	0		0	0		0	0		0	1	1	0	0	0	0
2	250	0		0	1		1	0		0	0		0	0		0	0		0	0		0	1	0	1	0	0	0
3	250	0		0	0		0	1		1	0		0	0		0	0		0	0		0	1	0	0	1	0	0
4	250	1		1	0		0	0		0	0		0	0		0	0		0	0		0	0	1	1	0	0	0
5	250	0		0	1		1	0		0	0		0	0		0	0		0	0		0	0	1	0	1	0	0
6	250	0		0	0		0	1		1	0		0	0		0	0		0	0		0	0	0	1	1	0	0
7	500	0		0	0		0	0		0	1		1	0		0	0		0	0		0	0	0	0	0	1	0
8	500	0		0	0		0	0		0	0		0	1		1	0		0	0		0	0	0	0	0	1	0
9	500	0		0	0		0	0		0	0		0	0		0	1		1	0		0	0	0	0	0	0	1
10	500	0		0	0		0	0		0	0		0	0		0	0		0	1		1	0	0	0	0	0	1
Total	3,500	500		500	500		500	500		500	500		500	500		500	500		500	500		500	750	750	750	750	1,000	1,000

Note. Form = Field-test form; n = expected target number of examinees per field-test form; MCQ = Multiple-choice questions; TEI = Technology enhanced items; SCR = Short constructed response items; Total = Total expected minimum number of students per specific item in field-test pool.