Catching Up: How Principled Assessment Design Can Welcome Assessments to the 21st Century

Rob Brucia, American Board of Pediatrics – rbrucia@abpeds.org
Karen Hoeve, American Board of Pediatrics – khoeve@abpeds.org
Topics

• Validity and validation
• Constructs
• Content outlines/blueprints
• Item development
• Psychometrics
What is Principled Assessment Design (PAD)?

• Incorporates theories of learning and cognition in assessment design

• Three common characteristics
  • Construct-centered approach
    • Define the construct first; all other assessment design is driven from this definition
  • Engineering towards intended interpretations and uses
  • Explicit design decisions and rationales

• Assessments should address the three elements of the assessment triangle
  • Cognition
  • Observation
  • Interpretation

(Nichols, Kobrin, Lai, & Koepfler, 2017)
The assessment triangle

**Observation**
Specification of content features that will elicit learner performance identified as evidence of status with regard to targets of inference

**Validity Argument**
Documentation of design decisions and rationales Regarding targets of inference, content features and performance features

**Interpretation**
Examination of the features of learners’ performance considered evidence of status with regard to targets of inference
Use of coding and aggregation of learner performances to make inferences about learners’ status with regards to targets of inference

**Cognition**
Construct-Centered approach beginning with careful and comprehensive examination of the constructs to be assessed to specify targets of inference

(Nichols, Kobrin, Lai, & Koepfler, 2017, p. 21)
PAD has five foundation elements and an organizing element

Organizing element
• Ongoing accumulation of evidence to support validity arguments

Foundation elements
• Clearly defined assessment targets (Cognition)
• Statement of intended test score interpretations and uses (Cognition)
• Model of cognition, learning, or performance (Cognition)
• Aligned measurement models and reporting scales (Interpretation)
• Manipulation of assessment activities to align with assessment (Observation)
Validity argument – the organizing element

Throughout the design, development and implementation process, PAD approaches:

• Accumulate validity evidence and build validity arguments to support intended score interpretations and uses
• Builds in the ongoing collection and synthesis of evidence to support validity arguments
Kane’s interpretive use argument

Trait Interpretation

Universe Score Expected Value over TD

Universe Score Expected Value over UG

Generalization

Extrapolation

Implication

Other Traits

Context

Observation Methods

Target Domain (TD)

Universe of Generalization (UG)

Sample of Observations

Observed Score

Scoring

Trait

Other Traits

Context

Observation Methods
The four main approaches in the PAD family

• Cognitive design systems
• Evidence-centered design (ECD)
• Principled design for efficacy (PDE)
• Assessment engineering (AE) – (mainly drawn on for this presentation)
PAD challenges in certification

• Mostly operationalized in the K-12 setting with minimal experimentation in high-stakes certification
• Construct maps identify the progression of knowledge, skills, and abilities, and PAD item development focuses on task modeling around these progressions
• Certification only takes into account examinee ability at the cut point, which ignores and nullifies much of the construct progression
• Implementing new strategies is difficult in a constantly moving and high-stakes environment
  • When is there time to try different things?
So….Why (PAD)?

IF IT AIN'T BROKE DON'T FIX IT.
FALSE.
IMPROVEMENT IS ALWAYS POSSIBLE.
Defining the construct

- **Construct** - the knowledge, skill, or ability that is being tested but that cannot be directly observed or measured
- “While medical licensure sets the minimum competency requirements to diagnose and treat patients, it is not specialty specific. *Board certification demonstrates a physician's exceptional expertise in a particular specialty and/or subspecialty of medical practice.*” – ABMS website
Construct – Exceptional Expertise – How do we define it?

• Traditional approach
  • A job analysis is conducted to determine the knowledge, skills, and abilities (KSAs) that are demonstrated in the profession (allows for generalizability)
  • Test items are written to assess these KSAs
  • An exam sampling the entire scope of KSAs is constructed
  • A standard is set for the exam that determines the cut point that a minimally qualified examinee, who demonstrates exceptional expertise, should score at or above

• In the traditional approach, is the construct actually defined during standard setting?
PAD Approach – Layered construct mapping

- Construct map – a map representing the knowledge, skills, and abilities (that are linked to the construct and are to be measured) and their progression along a single scale
  - Developed prior to the content outline/blueprint
- Content outline/blueprint areas fall on the construct map and items written to those areas should fit both content and ability (low/high)

<table>
<thead>
<tr>
<th>Pediatric &quot;Exceptional Expertise&quot; Construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
</tr>
<tr>
<td>Time Independent</td>
</tr>
<tr>
<td>Low Patient Risk</td>
</tr>
<tr>
<td>Knowledge of Basic Principles</td>
</tr>
<tr>
<td>Diagnosis</td>
</tr>
<tr>
<td>High Patient Risk</td>
</tr>
<tr>
<td>Time Dependent</td>
</tr>
</tbody>
</table>

PAD approach to KSAs and score interpretations and uses

Assessment targets and intended test score interpretations and uses are specified simultaneously

- These are the KSAs we want to assess
- This is what we want to be able to say about examinees, based on test scores
- These are the decisions we want to make about examinees, based on test scores

(Ferrara, Lai, Reilly, and Nichols, 2017, p. 56)
Alternative ideas?

- Is generalizability worth the payoff for sacrificing test items/data points on content areas that are known by all (qualified and non-qualified) examinees?
  - Classical test theory and Rasch models are summative so easy items or content area that lends itself to easy items do not provide much information
  - Do we trust licensure and training to cover some basic areas that may show up on a job analysis, but don’t necessarily need to be covered by a board-certification exam?
- What if to start the process, you asked a group of doctors for only the KSAs that you’d expect a board-certified doctor to possess AND a expect a non-board certified doctor to not possess?
  - These KSAs could be the basis for the content outline/blueprint
  - If items assessed the KSAs, the test would make up all relevant data points
  - Would 100% (or close to it) be expected?
Cut Scores

• Should be conceptualized concurrently with KSAs and the statement of what we want to be able to say about examinees during development and revision stages

• Should be set in such a way that the decisions they guide produce the fewest possible serious classification errors; for example, the physician who is incorrectly denied specialty recertification (false negative) or the certification of a physician who does not have exceptional expertise in a particular specialty (false positive).

• Should not be relegated to the end of the test development process; cut scores inform, and are informed by, the entire process.
Item development – Traditional approach

• Content outline/blueprint developed
• Subject matter experts (SMEs) trained on item writing basics
• SMEs assigned content areas for which to write items
• SMEs given leeway in developing items
  • Cognitive levels may vary over the same content areas
  • Item difficulty may vary over the same content areas
• Often considered an ‘art’ to craft a good performing item
Item Development – PAD approach

- PAD approach to item development is much more restrictive (science over art) in what is or is not included in an item
- Contents are restricted to concretely link the item, what it is measuring, and the item difficulty level to the construct map
- Approach is easily used or adapted to automated item generation

Task Models
- Map to a location and subsequent layers on the construct map
- Use Task Model Grammars (TGMs)
  - Coding that identifies the actions, objections, and relations of task model templates
  - Restrict the item writer or automated item generator (AIG) in what can/can’t be in an item
Item Development – Task Models

\[ \text{action}_2 \left[ \text{action}_1 \left( \text{is.related} \left( \text{object}_1, \text{object}_2 \right), \text{object}_3 \left| \text{context}, \text{aux.tools} \right. \right) \right] \]

Skill = identify
Objects = one, simple concept
Relations = none
Context = match word \rightarrow definition
Tools = none

Skills = identify, compare, evaluate
Objects = 3-4 complex properties
Relations = hierarchical (3 levels)
Context = complex text, dense info.
Tools = facilitative if used correctly
A [Age]-[Age_Type]-old [Gender] is evaluated because of [Presentation1] and [Presentation2] of [Accuity_of_Onset] duration. Initial physical examination shows that the patient has [Physical_Findings1], [Physical_Findings2], [Physical_Findings3], [Physical_Findings4], and [Physical_Findings5]. The patient’s vital signs are temperature [Temperature]°C, pulse [PulseRate] beats per minute, blood pressure [BloodPressure], and oxygen saturation [OxygenSaturation]. Which of the following is most appropriate treatment for this patient?

A: [Key]
B: [Distractor1]
C: [Distractor2]
D: [Distractor3]
A 5 year old male is evaluated because of fever and abdominal pain of 5 days duration. Initial physical examination shows that the patient has cracked lips, scarlatiniform rash, erythema of the hands and feet, and unilateral cervical lymphadenopathy (1.8 cm in diameter). The patient’s vital signs are temperature 38°C, pulse 90 beats per minute, blood pressure 80/60 mmHg, and oxygen saturation 96. Which of the following is most appropriate treatment for this patient?

A. Administer NSAIDs and begin a blood transfusion
B. Initiate IV doxycycline and monitor for 24 hours
C. Obtain an x-ray of the chest
D. Refer to the dermatologist
A 5 year old male is evaluated because of fever and abdominal pain of 5 days duration. Initial physical examination shows that the patient has cracked lips, scarlatiniform rash, erythema of the hands and feet, and unilateral cervical lymphadenopathy (1.8 cm in diameter). The patient’s vital signs are temperature 38°C, pulse 90 beats per minute, blood pressure 80/60 mmHg, and oxygen saturation 96. Which of the following is most appropriate treatment for this patient?

A. Administer NSAIDs and begin a blood transfusion
B. Initiate IV doxycycline and monitor for 24 hours
C. Obtain an x-ray of the chest
D. Refer to the dermatologist

Merging the two approaches

Can we analyze current items written with traditional methods to discover trends and teach SMEs?

• Start by looking at current used items with the same content specifications and same cognitive levels and determine if there are large differences in difficulty.

• If yes…
  • Does one item contain more variables than the other?
  • Are there differences in the distractors?
Example: Plan appropriate management of the different types of pneumonia

A previously healthy 18-month-old boy is hospitalized because of cough, nasal congestion, tachypnea, poor oral intake, and fever of 6 days’ duration. Physical examination reveals a well-developed, well-nourished but pale infant with dyspnea, grunting, nasal flaring, and intercostal retractions. Temperature is 39.4°C, pulse rate is 110/min, and respiratory rate is 60/min. Tympanic membranes are bulging and immobile; the pharynx is inflamed without exudate. The neck is supple. Chest auscultation reveals crackles in the right base. Remaining physical findings are normal. Laboratory studies show a hemoglobin concentration of 12.2 g/dL and a leukocyte count of 30,000/mm3.

Which of the following is the most appropriate next step in the evaluation of this patient?

A previously healthy 13-year-old girl has a 1-week history of low-grade fever to 38.1°C, headache, malaise, and a nonproductive cough. X-ray study of the chest reveals bilateral patchy infiltrates.

Which of the following is the most appropriate antibiotic therapy for this patient?
A previously healthy 18-month-old boy is hospitalized because of cough, nasal congestion, tachypnea, poor oral intake, and fever of 6 days duration. Physical examination reveals a well-developed, well-nourished but pale infant with dyspnea, grunting, nasal flaring, and intercostal retractions. Temperature is 39.4°C, pulse rate is 110/min, and respiratory rate is 60/min. Tympanic membranes are bulging and immobile; the pharynx is inflamed without exudate. The neck is supple. Chest auscultation reveals crackles in the right base. Remaining physical findings are normal. Laboratory studies show a hemoglobin concentration of 12.2 g/dL and a leukocyte count of 30,000/mm3.

Which of the following is the most appropriate next step in the evaluation of this patient?

P-value: .7698
PBIs: .2103
IRTb: .0101

A previously healthy 13-year-old girl has a 1-week history of low-grade fever to 38.1°C, headache, malaise, and a nonproductive cough. X-ray study of the chest reveals bilateral patchy infiltrates.

Which of the following is the most appropriate antibiotic therapy for this patient?

P-value: .978
PBIs: .131
IRTb: -2.3636

Item #1 has far preferable psychometric properties despite both items assessing the same content specification and requiring the examinee to diagnose the patient in the item stem and then choose the next action.

What are the differences?

- Patient age (easier to treat an older patient?)
- Additional signs/symptoms/labs in item 1 (in this case, do these distract, making it more difficult?)
- Query asking for next step vs antibiotic therapy (are medications easier to determine than next actions?)
Example: Plan appropriate management of the different types of pneumonia

A previously healthy 18-month-old boy is hospitalized because of cough, nasal congestion, tachypnea, poor oral intake, and fever of 6 days' duration. Physical examination reveals a well-developed, well-nourished but pale infant with dyspnea, grunting, nasal flaring, and intercostal retractions. Temperature is 39.4°C, pulse rate is 110/min, and respiratory rate is 60/min. Tympanic membranes are bulging and immobile; the pharynx is inflamed without exudate. The neck is supple. Chest auscultation reveals crackles in the right base. Remaining physical findings are normal. Laboratory studies show a hemoglobin concentration of 12.2 g/dL and a leukocyte count of 30,000/mm3.

Which of the following is the most appropriate next step in the evaluation of this patient?

A previously healthy 13-year-old girl has a 1-week history of low-grade fever to 38.1°C, headache, malaise, and a nonproductive cough. X-ray study of the chest reveals bilateral patchy infiltrates.

Which of the following is the most appropriate antibiotic therapy for this patient?
Using content outline specifications to focus item writing

- Process begins with the development of the content outline
- Use specification statements if both action and content area are not provided (matrix)
  - Avoid vague action verbs
    - Specifications that start with ‘Know the…’ can lead to a wide range of items
  - Avoid broad content area specifications (split into multiple parts)
  - Avoid specifications with multiple tasks unless a single test item can assess both tasks (ex: diagnose and treat)
  - Avoid lower level verbs such as ‘know’ and ‘understand’ unless you truly want to assess lower level knowledge
    - Demonstration of this knowledge does not help inform the cut score decision
  - Consider creating an internal content outline for item and exam development
  - Use higher level verbs (diagnose, treat, manage, formulate) to tap into KSAs that demonstrate ability at the cut score
- More direction should lead to higher item approval rates (offsetting the time spent up-front creating the statements)
  - More efficient than the ‘shotgun’ approach that vague content specifications produce
  - Less pressure on item writers and staff to produce large quantities of items with standard 30% discard rates
Back to our earlier example!

- Plan appropriate management of the different types of pneumonia
  - Is the desire to assess the ability to plan management or the ability to manage? If manage, eliminate ‘plan’ in the statement
  - Should the specific type of pneumonia be stated?
    - Item bank may be healthy in items on the management of **bacterial pneumonia**, but items are needed on **viral pneumonia**
    - Without specifying, the ‘shotgun’ approach may produce a handful of items, none of which touch on viral pneumonia
  - Revision: *Manage a patient with viral pneumonia*
- Higher cognitive level example
  - *Diagnose and treat a patient with viral pneumonia*
    - Item scenario would provide a patient presenting with the signs and symptoms of pneumonia and ask how to treat the patient
    - Examinee would first have to formulate the diagnosis, based on the patient description, and then choose the appropriate treatment, thus assessing ability to both *diagnose and treat*
    - While more difficult to write, these items provide far more information on medical specialty certification examinees than recall or application items
Comparing traditional and principled approaches to assessment

Principled approaches:
1. Define assessment targets;
2. Define intended score interpretations and uses;
3. Select/develop a model of cognition, learning or performance;
4. Select aligned measurement models;
5. Manipulate assessment items and tasks.

Identify constraints; design the test.
Develop a test blueprint.
Generate test items.
Generate test forms.
Conduct field test, conduct scaling and psychometric analyses.
Conduct field test, conduct scaling and psychometric analyses.
Generate test forms.
Implement the test, operate the testing program.
Principled approaches: (5) Manipulate assessment items and tasks.
References


Questions?