Strategies for clean performance-related validation (even when the data are messy)

RCIO 2015

Christopher J. L. Cunningham, PhD
Target Audience(s)

• I-O types who could use a refresher on the reality of assessment/predictor validation in organizational settings

• Anyone who can appreciate my logic and sense of humor

• Anyone who couldn’t find another presentation they’d rather attend during this portion of our scheduled programming
Objectives

• Messiness in validation
• Why this is an issue
• To be “validated”
• Validation in principle
• Then there’s reality…
• Practical and holistic validation

NOTE: I will use “assessment” throughout this presentation to denote any predictor (including single-item indicators, observations, and other forms of measurement)
Messy

mess·y
mesē
adjective
1. untidy or dirty.
"his messy hair"
synonyms: dirty, filthy, grubby, soiled, grimy

2. (of a situation) confused and difficult to deal with.
"a messy divorce“ or “a messy validation"
synonyms: complex, intricate, tangled, confused, convoluted; “hot mess” (at least in TN...)
Messy Validation?

- Messiness in validation can result from a combination of factors, including:
  - The questions being asked
    - The stated question vs. actual question
  - The data
    - And what it represents (or doesn’t)
  - The sophistication of the analyst and client/recipient
    - Knowing “just enough to be dangerous” about methods and statistics can create problems
Messy Questions

• Client: “*Is this test valid?*”
  - **Real question:** “Can you guarantee that if I use this test, I will hire only high-performing applicants?”

• Vendor/consultant answer (hopefully): “Yes, this assessment has been validated.”

• **Real answer:** “Nothing is certain in life, but there is evidence that when this assessment is used as directed, it will increase your likelihood of making good decisions (i.e., hiring high-potentials and not hiring low-potentials)”
Messy Data – Is there any other kind?

Predictor data
- Not clearly defined
- Not linked to behaviors or attributes that can be observed or otherwise evaluated via any means other than self-report
- Oddly distributed (frequency-wise)
- Seemingly irrelevant to the work context

Outcome data
- Not clearly defined or consistently gathered
- Seriously inconsistent variability across manager or location/site
- Client “cares” about quality, but all metrics are quantity-focused
- Supervisor ratings restricted to only the high end of rating scale
- Client has decided to be a trend follower and drop performance evaluations altogether
Being “validated”

- Personally, it’s nice when it happens
- Professionally, I-O types would like all of their assessments to have this “badge”
  - Our own Good Housekeeping Seal
- **Challenge**: Normal people really do not care
  - We have to make them
- **Challenge**: I-O psychologists are not the only people developing and selling assessments
  - Imagine the implications
Meaning of Validation

• Validation ≠ accuracy or precision
• A test cannot really be valid in and of itself
• Validation is a “property of inference”
  • What the heck does that mean?
• Validation is achieved with evidence for the usefulness and relevance of an assessment as a predictor of something important in an organization
  • Could I make this any more generalizable?
Spectrum of Validation

• Face
  • A person responding to questions in the assessment can “see” the work relevance of what they are being asked to provide

• Content
  • Assessment evaluates content relevant to the actual work environment

• Criterion-related
  • Predictor → outcome

• Construct
  • Assessment evaluates what it is supposed to
Forms of Validation Evidence

**Challenge:** Validation is not achieved only one way; Not all ways are equally good or appropriate in all situations

<table>
<thead>
<tr>
<th>“Element”</th>
<th>Sub-elements</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion-related</td>
<td>Concurrent Predictive</td>
<td>Statistical</td>
</tr>
<tr>
<td>Construct</td>
<td>Convergent Discriminant</td>
<td>Statistical and not Statistical</td>
</tr>
<tr>
<td>Content</td>
<td>For different types of “users”</td>
<td>Your call</td>
</tr>
<tr>
<td>Face</td>
<td>For different types of candidates</td>
<td>Your call</td>
</tr>
</tbody>
</table>
Additional Considerations

• In theory, an unreliable assessment cannot be validated

• Practically speaking, I cannot make good decisions using assessment results when I:
  • do not believe the test consistently captures variability among candidates
  • am not sure how to interpret or make sense of the results of an assessment
Reliability Refresher

• “...the extent to which test scores are consistent or free from random error.”

• Common estimates of reliability:
  • Test-retest reliability*
  • Interrater reliability
  • Internal consistency (Cronbach’s alpha)

• Most estimates should be high (ideally > .70) for assessments used to guide employment-related decisions
Empirical vs. Practical Validity

• The elements just discussed are considered central to establishing empirical evidence for validity

• Perhaps even more important is evidence for the practical validity of an assessment
  • Extent to which decisions made with the help of data generated by an assessment are accurate

• This is the ultimate goal of validation: To ensure that assessment users make good decisions
  • Ideally, better decisions with the assessment than without
So what is a “good decision”?

**Ultimate work-related performance is excellent (ok, at least decent)**

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assessment indicates this is a “keeper”; we want this person on our team.</strong></td>
<td>✔️</td>
<td>✗</td>
</tr>
<tr>
<td>True</td>
<td>✔️</td>
<td>✗</td>
</tr>
<tr>
<td>False</td>
<td>✗</td>
<td>✔️</td>
</tr>
</tbody>
</table>

Is this overly simplistic? Probably.

Is this fairly realistic? Yep.

So, why do we complicate our validation processes beyond addressing a fairly simple need?
Validation in Reality

• **Ultimate goal:** To help decision makers make better decisions

• **Process:** Rational and statistical linking of predictor to outcome data

• **Necessary elements:**
  - Data associated with the predictor and outcome
  - Collaboration with organizations and people to be assessed (incumbents and applicants)
  - Some level of research and statistics proficiency
    - The data will not always “speak for themselves”
    - Sometimes when they do, it’s in a different language
Real Validation is Messy

• Real data gathered in real organizational settings from real candidates = a real (hot) mess
  • poor data quality
  • overly complex perceived/real client needs
  • lack of careful processing by the validator

• Messy validation data should not lead to messy validation studies

• As the people doing the validation analyses and work, we need to remember not to add to the real/perceived complexity
Cleaning up a Mess

- Sometimes we “clean” by covering it up
  - Febreze
  - Statistical corrections
- If we’re serious, we get our hands dirty and we work on gathering and organizing the pieces
  - Putting the train tracks and Legos in their boxes
  - Carefully examining the actual data and what it really means
Clean Validation Techniques
*(when the data are a mess)*

- Re-evaluate the measurement scales
- Consider parametric techniques with realistic data cleaning
  - Focusing on core density of score distributions; accounting for source-effects
- Consider non-parametric alternatives
  - Chi-square
  - Observation Oriented Modeling
What is really being measured?

• **Scenarios:**
  • 7-point Likert scale of agreement:
    - Dstrong / D / Dslight / N / Aslight / A / Astrong
  • Extended frequency scale:
    - 10-point from Not at all to All of the time

• **Mess:** The ratings gathered by these types of scales can often be seriously skewed or otherwise bizarre
  • Makes these data difficult to interpret
What is really being measured?

• **Possible cleaning solution:** What is really being measured?
  • Does a continuum evaluation make sense?
  • Perhaps a simplified Agree/Disagree or Agree/Disagree/? framework is more appropriate
Parametric Stats with Realism

• **Scenario:** You observe a negative $r_{XY}$ associated with a test that is normally quite consistently positive as a predictor of certain types of performance? *What the H$!!$?*

• **Mess:**
  • Outlying data can really screw up correlational statistics (remember the see-saw that is leverage)
  • Multiple raters do not appear to be using the same yardstick
Parametric Stats with Realism

• **Possible cleaning solution:** Study the data very closely (and keep your critical thinking hat on)
  • Maybe a small number of individuals took an inconceivably brief amount of time to complete the predictive assessment
  • Excluding these few to focus on the vast majority who took the test “for realz” returns the relationship to the direction and magnitude that would be expected
Parametric Stats with Realism

• **Possible cleaning solution:** Remember that z-scores are your friend
  • Consider standardizing performance ratings within rater or at least location
  • Can help to take into account differences in rater biases and/or local behavioral norms and expectations
  • Can often analyze these z-scores using fairly straightforward parametric techniques
Candidates scoring in the top 50% of the distribution of scores for the predictive assessment are significantly more likely than candidates in the bottom 50% of this distribution to be rated as better overall performers, $t(104) = 2.53, p < .05, r = .24$.

The figure summarizes mean supervisor ratings for these two groups, after standardizing the ratings by supervisor (ratings source).
Non-parametric Alternatives

- Most commonly taught statistical analysis techniques carry serious baggage
  - We call them assumptions
- Alternatives exist for just about every traditional, parametric analysis tool
  - May lack statistical power
  - May not be ideal for estimating population parameters
  - **Reality check**: When validating an assessment in a specific organizational context the goal is not always parameter estimation.
## Comparison of Analysis Options

<table>
<thead>
<tr>
<th>Analytical purpose</th>
<th>Parametric</th>
<th>Non-parametric*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central tendency</td>
<td>Mean</td>
<td>Median; Mode</td>
</tr>
<tr>
<td>Range</td>
<td>Standard deviation</td>
<td>SIR</td>
</tr>
<tr>
<td>Relationship</td>
<td>Pearson $r$</td>
<td>Kendall’s Tau;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spearman rho;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chi-square</td>
</tr>
<tr>
<td>2 groups (independent)</td>
<td>$t$ (independent</td>
<td>Mann-Whitney $U$ ;</td>
</tr>
<tr>
<td></td>
<td>samples)</td>
<td>Wilcoxon rank sum</td>
</tr>
<tr>
<td>&gt; 2 groups (independent)</td>
<td>ANOVA (between</td>
<td>Kruskal-Wallis</td>
</tr>
<tr>
<td></td>
<td>groups)</td>
<td></td>
</tr>
<tr>
<td>Repeated measures (2 groups)</td>
<td>$t$ (paired samples)</td>
<td>Wilcoxon signed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ranks; sign test</td>
</tr>
<tr>
<td>Repeated measures ( &gt; 2 groups)</td>
<td>ANOVA (repeated</td>
<td>Friedman’s</td>
</tr>
<tr>
<td></td>
<td>measures)</td>
<td></td>
</tr>
</tbody>
</table>

*please note that non-parametric does not mean assumption free*
Chi-square Can Be Your Friend

• Raw frequencies are often easier for recruiters and managers to understand than abstract ratings
• Demonstrating that higher scorers on a predictive assessment are substantially more likely than lower scorers to be good performers is strong validation support
• Chi-square techniques can help
# Chi-Square Example

<table>
<thead>
<tr>
<th>Predictor assessment score</th>
<th>Performance relative to peers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>48</td>
</tr>
<tr>
<td>High</td>
<td>2</td>
</tr>
</tbody>
</table>

I know, small numbers of $f_{oc}$, but remember that what matters more in chi-square are the $f_{oc}$, and for this analysis all of these were $> 5$

**Helpful interpretation aid – Odds Ratios**

Odds of high performer, if low predictor score: $47/48 = 0.979$
Odds of high performer, if high predictor score: $10/2 = 5.000$
Odds ratio (high / low): $5.00/0.978 = 5.11$

High scorers on the assessment are **5x** more likely to demonstrate higher performance than their peers, compared to low scorers on the assessment.
Alternative Data Analyses

• All preceding statistical tools are applied within traditional NHST perspective
  • Testing observations against null hypothesis

• Increasingly, NHST is criticized as extremely limiting to our science
  • Especially when the questions we want to answer are about person-level phenomenon, and not group-level effects
Main NHST Limitation for Validation

• Traditional “evidence” of validity =
  • degree to which assessment X “explains” variability in outcome Y

• The potential problem:
  • explained variance ≠ predictive accuracy

• Example:

\[ y = 0.5x + 3, \quad R^2 = 0.663 \]

Explained variance ≠ Accuracy

Observation Oriented Modeling

• Very different option, worth considering
• Moves us past aggregate summaries to analysis of patterns at person level
  • Which is a level at which validation has not typically focused
• Also provides alternative to NHST and parameter-based statistical methods
  • No meaningless aggregation of data
  • Person-centered
  • No assumption-laden $p$-values
Statistically Significant ≠ Meaningful

**ANOVA**

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>7.311</td>
<td>2</td>
<td>3.655</td>
<td>5.688</td>
<td>.004^a</td>
</tr>
<tr>
<td>Residual</td>
<td>122.731</td>
<td>191</td>
<td>.643</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.041</td>
<td>193</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), OSU GPA, COMPOSITE ACT  
b. Dependent Variable: CONSENSUS

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.237^a</td>
<td>.056</td>
<td>.046</td>
<td>.802</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), OSU GPA, COMPOSITE ACT

**Coefficients**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Zero-order</td>
</tr>
<tr>
<td>(Constant)</td>
<td>1.568</td>
<td>.391</td>
<td></td>
<td>4.006</td>
<td>.000</td>
</tr>
<tr>
<td>COMPOSITE ACT</td>
<td>.046</td>
<td>.017</td>
<td>.224</td>
<td>2.742</td>
<td>.007</td>
</tr>
<tr>
<td>OSU GPA</td>
<td>.032</td>
<td>.111</td>
<td>.024</td>
<td>.293</td>
<td>.770</td>
</tr>
</tbody>
</table>

a. Dependent Variable: CONSENSUS

Significant ≠ Meaningful
No clear pattern linking these data

Alt Example

SO HOW CAN I BE A CLEANER VALIDATOR?

So glad you asked…
Data Quality x Methodological Simplicity

• Focus on the highest-quality data available
  • Consciously avoid variables that capture no variability
  • Filter out data that are likely poor quality
    • Severely skewed, limited variability
    • Questionable ratings source or meaning

• Use appropriate and simple analytical methods
  • Check basic assumptions and use appropriate, simple statistical tests
General Starting Points

• Carefully consider client criteria for success
• Consider restrictions on predictors and outcomes (as operationalized)
• Review quality of available data
  • Representativeness of sample (data source)
  • Understanding why these indicators
    • Because they exist (or are easy to access)
    • Because they rationally seem optimal
    • Because they are theoretically/empirically supported for the targeted purpose
Clear and Complete Reporting

• Ultimate goal is to tell a story
• No PhD should be required to understand validation evidence and the utility of an assessment
  • Make the “so what” points clear
• Also need to summarize and provide sufficient details to enable replication and clear following of logic, process
Real Validation…

• …should provide clarity and reduce complexity

• …should be targeted at “worst-case” scenario

  • Need solid justification/rationale for statistical corrections and other magical forms of interpreting reality
Validation is a Process

- Review and recalibrate
  - Validity studies provide a snapshot, but changes in hiring needs and recruiting practices may require re-validation/re-calibration
- Validation studies often shine light on other challenges
  - Is a 70% pass-rate through the assessment really ideal?
    - Does the organization have problems with its recruitment/attraction processes or other aspects of its selection screening funnel?
  - Wouldn’t limiting interviews and resume reviews be more efficient?
Consider the Assessment in Context

- Statistically demonstrating a predictor-outcome link does not mean that an assessment will actually help a company make better decisions about applicants
  - Maybe the assessment takes 2+ hours (!) or requires way more than it should from the applicant (blood, sweat, tears, etc.)
  - Maybe recruiters and managers can’t wrap their minds around why applicants are asked about experiences stealing paper clips or where they would sit during a baseball game
  - Maybe company technology hasn’t been updated since 1995 and systems can’t manage new data
Validate the Assessment + Process

• How is the assessment going to be used to guide decisions?
  • If top-down, rank-ordered, then do that (but only if the assessment and outcome actually work for this type of linkage)
  • Remember that good/bad decision making is a dichotomy – sometimes thinking along a continuum only complicates things
    • If pass/fail, then validate for pass/fail decisions
THANK YOU for PARTICIPATING ANY QUESTIONS?

chris-cunningham@utc.edu