Handout for Webinar II: Examples of Validity Evidence

Valid Content: Rating Tool in the Operating Room

Claim: Behaviors identified as representative of "non-technical" skills are the "right" behaviors to observe and rate when evaluating physicians, nurses, and anesthetists in the operating room.

An excellent example of content validation comes from a study of an observational behavioral rating tool called "OTAS"[1] which was designed to assess the "non-technical" skills (e.g., communication) of professionals in the operating room (OR). The authors began the process of developing this tool by sitting in on a lot of operations and observing how surgeons, anesthetists, surgical assistants, and nurses worked together. Their ethnographic field notes, combined with findings from a literature review of "best practices," and an audit of OR documents of actual procedures, and an expert review of potential items, led to the first version of the OTAS. This OTAS was then piloted in the OR with 30 surgical procedures. During the pilot, the authors found that subsets of behaviors being measured were not very relevant, or they were rarely observed, or were very difficult for trained raters to score. The authors then had experts from each profession (surgery, nursing, anesthesia) review the problematic behaviors and score them for "relevance to teamwork" and "patient safety." The experts' rank scores of the problematic items helped the authors prioritize the revisions. As a result, the authors ended up removing 21 behaviors and modifying another 23 for the final tool.

Why is this a good example? What we like about this example is the variety of methods the authors used to achieve their end, and the systematic way they went about it. It leaves us convinced that they didn't just "make up" the items. These items had a real basis in empirical observation, *and* expertise, *and* experience.

[1] Hull, L., Arora, S., Kassab, E., Kneebone, R. & Sevdalis, N. (2010). Observational teamwork assessment for surgery: Content validation and tool refinement. *J Am Coll Surg*, 212(2): 234-243.

Valid Response Processes: Simulation Training with Observer Ratings

Claim: Raters fully understand the assessment task and are able to respond consistently and accurately

To illustrate response process validation, we describe a fictitious example involving simulated case conferencing scenarios and an observational assessment rating tool measuring collaborative communication and decision-making. The simulation cases and tool were originally developed for a multi-disciplinary team of providers specializing in oncology care in a hospital setting. The scenarios are 30 minutes in length and include 6-8 team members.

Faculty members from an IPE program are interested in adopting the rating tool (but not the specific scenarios, which they already have) for training their pre-licensure students. Despite the fact that the tool was originally designed for providers (not students) and for a specific disease focus (rather than a broad acute care population), the faculty members feel the items really get at the nature of interprofessional, collegial interactions in a general way. They are aware that their own simulations run closer to 15-20 minutes in length, vs. 30 minutes. But their scenarios involve fewer participants, so they judge the time adequate for raters to observe and assess students' collaborative communication and decision-making.

To ensure appropriate response process, the faculty members conduct a focus group with five persons who will be serving as raters. Together they review the instrument: its instructions, items, scoring procedures and criteria for assigning scores. As a result, some minor word substitutions and clarifications are made to the tool. The IPE faculty then conduct formal training with all 10 raters. The raters view videotapes of simulated case conferences and practice scoring. They then discuss their scores in order to work towards more consistent scoring. The raters continue their practice by observing and rating two live simulation scenarios. Their scores are analyzed for inter-rater reliability and agreement. Results suggest that raters need some additional training around two of the behaviors being rated, as there is little agreement among their scores.

The IPE faculty then implement the simulation program with 10 groups of students and collect the assessment data. Before pooling the data for analysis, however, they conduct a quality check of the simulation exercise by reviewing the 10 videotapes. In doing so, they discover problems with how the scenarios was administered in two rooms. In one room, a fire drill cut the scenario short prematurely, giving the raters less time to assess team interactions. In another, a facilitator made some personal comments during the pre-briefing that could be viewed as prejudicial. In light of this, the IPE faculty decided to delete the scores from the two affected groups from the analysis.

Why is this a good example? In this example we tried to show that valid data depend not just on an instrument, but on the interpretation and actions of those responding to it. It is especially important to check on response process validity when adapting a tool for a slightly different purpose and population than the one for which it was developed. Valid data also depend on the conditions in which the data are collected. In our fictitious example, the faculty did the right thing to preserve the integrity of the data by eliminating suspect ratings, even if the potential impact on scoring from the fire drill and faculty comments was small.

Validity based on Internal Structure: Team Member Survey

Claim: Items on a self-report survey will accurately measure five constructs that are key to a theory of goal-driven teamwork.

Here we turn again to a fictitious example. This example involves a 25-item survey being developed in which individual team members are asked to rate their perceptions of each other team member's commitment to work towards team goals. The items are being written to measure five constructs believed important based on a literature review: situational leadership, patient-centeredness, communication, collaboration, and coordination. During a pilot test of the instrument, however, a factor analysis finds a high degree of overlap between communication, collaboration, and coordination. This leads to low item factor loadings and an unclear pattern of association between items and these supposedly distinct constructs. The researchers delete 5 items with low factor loadings and re-run the analysis. This reveals the instrument appears to be successfully measuring three distinct factors: leadership, patient-centeredness, and a broader interpretation of collaboration. Each of these factor scales has an internal consistency reliability of about 0.59. The Cronbach alpha for the all 20 items combined, however, is 0.89. As a result, the authors recommend that people using the tool should only calculate and use the total survey score, and not the sub-scores for leadership, patient-centeredness, and collaboration.

Why is this a good example? In this example we tried to show that by analyzing data from pilot test subjects, we can learn a lot about what an instrument is actually measuring. Although theoretically the constructs of communication, collaboration, and coordination are different, the differences are subtle.

In this example, participants may have interpreted an item written to represent communication as evidence of collaboration. Regardless of how the participants interpreted the items, what the data tell us is that they tended to rate their peers very similarly on these items. Thus, a colleague who was deemed a good communicator was also deemed a good collaborator. To establish theory, it would be important for researchers to keep writing and testing items so the three overlapping constructs are more clearly delineated. In practice, however, the distinction may not matter quite so much. It depends on the purpose and use of scores.

The example shows that even after deleting redundant items, the tool still seemed to lack reliability for the remaining constructs (leadership, patient-centeredness, and collaboration): each factor's scale reliability was only r = 0.59. The researchers' solution was to qualify how the instrument should be used. They recommended that only the total score be used, since the sub-scores for the three remaining constructs were unstable. Using the instrument this way would limit, however, the ability of trainers to score trainees in leadership, patient-centeredness, and collaboration. Trainers could, however, provide trainees with their results for feedback and learning.

Validity based on Relationships with Other Variables: Three Different Examples

Several different hypothetical examples are used to illustrate this source of validity evidence.

Claim 1: A situational judgment test designed to identify candidates with leadership ability predicts effective leadership in practice.

A situational judgment test (SJT) requires candidates seeking a position to read or view simulated scenarios in which they have to choose how they would respond to certain leadership challenges. The value of the SJT lies in its ability to predict which candidates who would, or would not be effective in a leadership role in practice. In this fictitious example, a health system wants to hire providers who will be effective leaders in an IPECP setting. They find a promising instrument with significant research behind it, including some persuasive convergent and divergent validity data. In developing the SJT, the researchers piloted it with multiple professionals. They collected over 300 forms from professionals who had been stratified into two groups: those deemed more effective, and those deemed less effective in collaborative leadership, as rated by their supervisors, peers, and subordinants. The researchers reported that the scores from the SJT correlated highly with the external rater groups. To ensure test fairness, the researchers also collected race and gender data on the participants. They reported that the scores on the SJT did NOT correlate with either race or gender. (If they had, the researchers would have had to re-examine the tool for latent racism and sexism.)

Why is this a good example? We wrote it to show the value not just of convergent data (i.e., the correlation between data measuring leadership ability from two different sources), but divergent data (i.e., ruling out error from extraneous factors, such as race and gender, affecting the results). (Note: divergent validity is also called discriminant validity.)

Claim 2: A team performance simulation and rating tool will accurately identify who needs training.

In this example, a provider system wants to know which clinical teams in their network might benefit from team training. They develop a teamwork performance simulation and a rating tool and administer it to three groups: (1) experienced teams with prior training in teamwork, (2) experienced teams without prior training in teamwork, and (3) newly formed teams with little experience working together and unknown prior training in teamwork. As expected, the two experienced teams performed better than the inexperienced ones, but the experienced teams without prior team training performed less well than experienced teams with prior training. This result confirms that the training is helpful for newly formed teams and for experienced teams without prior training. Had the teams scored similarly, the trainers would have concluded that the test was too easy (if everyone passed), or too difficult (if everyone failed), or simply not useful (if every group scored in very mixed ways).

Why is this a good example? We wrote it to show how piloting an instrument to test one's assumptions about its scores and utility can directly inform programming decisions.

Claim 3: The quality and quantity of communication between providers during interprofessional, collaborative care conferences will predict patient satisfaction with care.

In our third example, researchers are studying a new model of collaborative care conferencing for a particular patient population with chronic disease. They developed an observational tool measuring the amount and quality of interprofessional communication during care conferences. They collected observational score data from two groups: the treatment group which had been training under the new care conference model, and a comparison group that continued "business as usual." Researchers simultaneously collected satisfaction data from patients being seen by both provider groups. Scores from both groups predicted patient satisfaction with care, but satisfaction was higher with teams using the new collaborative care model.

Why is this a good example? We wrote it show how predictive validity data can get at the heart of IPECP's mission and influence changes in the provision of care.

Validity based on Consequences of Use

Here are some (positive) examples of consequential validity for the three fictitious examples shown above:

- In the case of the SJT, better hiring of candidates with collaborative leadership potential was found to lead to higher satisfaction among clinical teams and more efficient patient care. An additional positive consequence was that more women and people of color were hired. The SJT proponents attributed this to the objective, behaviorally-based nature of the SJT. Latent bias is a known problem when hiring decisions are based on more fluid impressions by those doing the interviewing.

- In the case of team training, the teamwork simulation rating tool was so reliable, accurate, and predictive of performance in practice, that it enabled the health care system to save $10,000 by only training the teams that needed it, rather than mandating it for everyone.

- In the last example, given the direct correlation between an IPCP care conference model and patient satisfaction, the health system replicated the model across clinics. Further study showed that this model led to a 5% reduction in hospital readmissions, as well as positive patient satisfaction.

Conversely, here are some (negative) examples of consequential validity for the same three examples:

- The SJT was not used very often, given the low number of new positions needing to be filled in the hospital. Over three years, it led to the hiring of only three new clinical directors, two of whom “didn’t work out” for reasons that could not be predicted by the SJT. The SJT was eventually phased out because the stakeholders lost confidence in it. It also took a lot of time to administer, and it also discouraged some candidates from applying.

- Over time, teams being directed to simulated training started to “game the system.” That is, they knew ahead of time what the scenarios covered and how they were expected to “behave.” They started to take the training less seriously. Additionally, some administrators started to grumble about the cost of replacing providers when they were in mandated training. Eventually, trainers introduced an abbreviated, “just-in-time” training for new and un-trained teams at the start of shifts.

- The positive correlation between the innovative case conference model and patient satisfaction reported in the research study was not actually replicated in the practice setting. After several attempts to implement the model, team members kept resorting to “previous behavior.” There was no discernible impact on patient satisfaction levels, which had actually been pretty high to begin with. The health system continued to promote the case conference model, but stopped monitoring the teams with observational data.

Why are these good examples? In these examples we tried to show that ultimately, assessment and evaluation tools must serve stakeholder needs and purposes, and benefit those involved. It is important to continue to research the accuracy and utility of instruments after they have been introduced. Health care environments are dynamic by nature, and this can impact the instrumental value of the tools and the larger assessment processes.