EVALUATING INTERPROFESSIONAL EDUCATION AND COLLABORATIVE PRACTICE:

What Should I Consider When Selecting a Measurement Tool?

Connie C. Schmitz, PhD
Director of Education Research and Development
Department of Surgery
University of Minnesota Medical School (Twin Cities)

Michael J. Cullen, PhD
Director of Evaluation
Office of Graduate Medical Education
University of Minnesota Medical School (Twin Cities)

University of Minnesota

National Center for Interprofessional Practice and Education

ABOUT THE AUTHORS

Connie C. Schmitz is an Associate Professor and the Director of Education Research and Development for the Department of Surgery in the University of Minnesota Medical School (Twin Cities). She holds a master's degree in Curriculum and Instructional Systems and a doctoral degree in Educational Psychology, both awarded by the University of Minnesota College of Education and Human Development. Much of her career has focused on learner assessment and program evaluation in the context of medical education, health care, and human services. In her early career, she was a research fellow in the Department of Family Medicine and Community Health. Prior to joining the Surgery Department in 2006, she ran her own consulting business for nine years, conducting program evaluation studies for major foundations, government agencies, and non-profit organizations. She served as the Associate Editor of the *American Journal of Evaluation* for three years and is an active member of the editorial board for the *Journal of Surgical Education*. In 2014, the Association of Surgical Education awarded her its highest honor, Distinguished Educator of the Year.

Michael J. Cullen is the Director of Evaluation for Graduate Medical Education at the University of Minnesota Medical School (Twin Cities). He holds a doctoral degree in Industrial-Organizational Psychology from the University of Minnesota, a juris doctor from the University of Toronto, and a bachelor of arts from Princeton University. Dr. Cullen has been conducting and managing research in personnel and related areas for 11 years and has participated in a wide variety of large-scale applied research projects with private sector and government clients. His research has included job analyses, test development and validation for selection, promotion, and certification; training program development, delivery, and evaluation; performance management, and competency modeling. His recent personnel selection research has focused on developing and validating a variety of predictors, including knowledge tests, structured interviews, situational judgment tests, biodata, personality tests, and behavioral assessments. Dr. Cullen is a past member of the editorial board of *Personnel Psychology* and has published his research widely, including in such outlets as the *Journal of Applied Psychology*, *American Psychologist*, and *Human Performance*.

This primer was commissioned by:
Barbara F. Brandt, PhD
Associate Vice President for Education
Academic Health Center Office of Education
University of Minnesota (Twin Cities)
Director, National Center for Interprofessional Practice and Education

Publication Date: March 20, 2015

ISBN-10: 0692392467
ISBN-13: 978-0-692-39246-1

With sincerest thanks, the National Center for Interprofessional Practice and Education would like to recognize the following individuals for their contributions to the publication of this primer.

ACKNOWLEDGEMENTS

Experts representing a variety of professions in education; measurement; interprofessional teamwork; and healthcare research, evaluation, and assessment reviewed this primer during its preparation to provide feedback to the authors. They vetted, edited, and enhanced its content and presentation immeasurably. These individuals reviewed the primer for clarity, scope, relevance, and consistency with best practices in applied measurement in mind.

Each provided careful content review to ensure that the most current thinking in the area of measurement is thoroughly and clearly communicated.

Although the reviewers listed below provided comments, additions, and suggestions, they were not asked to endorse this primer's recommendations and did not see the final draft prior to its release. Responsibility for the final content of this primer rests solely with the authors.

THE NATIONAL CENTER THANKS THE FOLLOWING REVIEWERS:

Judith Gedney Baggs, PhD, RN, FAAN
Professor and Special Assistant to the Dean for Research and Scholarship, School of Nursing, Oregon Health & Science University, Portland, OR

J. Michael Oakes, PhD
Professor, School of Public Health, University of Minnesota, Minneapolis, MN

Eduardo Salas, PhD
Trustee Chair and Professor of Psychology, University of Central Florida, Orlando, FL

Nick Sevdalis, PhD
Senior Lecturer, Department of Surgery and Cancer, Center for Patient Safety and Service Quality, Imperial College of London, UK

Joseph Zorek, PharmD
Assistant Professor, School of Pharmacy, University of Wisconsin, Madison, WI

Michelle L. Gensinger, MSEd
Doctoral Student, Evaluation Studies, College of Education and Human Development, University of Minnesota, Minneapolis, MN

Sponsors

UNIVERSITY OF MINNESOTA
Academic Health Center

The **University of Minnesota Academic Health Center Office of Education** provided funds to support *Evaluating Interprofessional Education and Collaborative Practice: What Should I Consider When Selecting a Measurement Tool?* This primer will contribute to the University's 1Health Curriculum and the work of the National Center.

National Center for Interprofessional Practice and Education

The **National Center for Interprofessional Practice and Education** is supported by Health Resources and Services Administration Cooperative Agreement Award Number UE5HP25067 and by the Josiah Macy Jr. Foundation, the Robert Wood Johnson Foundation, and the Gordon and Betty Moore Foundation.

TABLE OF CONTENTS

Foreword ..... p. v

Introduction ..... p. 1

Purpose of this Primer ..... p. 4

What is Validity and Why Does it Matter? ..... p. 5

Five Sources of Validity Evidence ..... p. 8

- Claim 1: Content
- Claim 2: Response Process
- Claim 3: Internal Structure
- Claim 4: Relationship to Other Variables
- Claim 5: Consequences of Testing

How Much Validity Evidence Should a Tool Have? ..... p. 12

Things to Consider When Selecting a Tool ..... p. 13

1. Relevance to Your Situation
2. Strength of Validity Evidence
   Example: The Nurse-Physician Questionnaire
3. Operational Considerations

Summary ..... p. 22

References ..... p. 24

APPENDICES

Glossary of Measurement Terms ..... p. 25

Rules of Thumb When Appraising Validity Data ..... p. 28

Common Threats to Validity ..... p. 30

Further Reading ..... p. 32

National Center Resource Exchange ..... p. 33

FOREWORD

Message from Barbara F. Brandt, PhD

THE NEED:

The National Center for Interprofessional Practice and Education receives numerous requests from educators, practitioners, and researchers in the interprofessional education and collaborative practice (IPECP) community. The number one expressed need is about measuring IPECP and whether it is effective. These are the types of questions we receive:

- **"How do I know what to use to measure IPECP?"**
- **"How can I select the right tool?"**
- **"What tool(s) do you recommend?"**

What we observe is that many are looking for a "magic bullet" to find a simple solution to all measurement needs. Unfortunately, there is no single answer because every setting has unique measurement needs and, even within a setting, these needs may change at different times with different groups for different purposes. In actuality, there are many IPECP instruments that address a variety of measurement needs, some of which are readily available. IPECP measurement tools are constantly being designed and improved by the developers and those using these tools.

THE RESPONSE:

In 2013, in response to questions about IPECP measurement, the National Center created the Measurement Instruments (https://nexusipe.org/measurement-instruments) collection on the Resource Exchange (https://nexusipe.org/resource-exchange). Today, this collection represents the most popular and viewed pages on the exchange. The National Center staff is committed to supporting measurement needs. Therefore, the number of resources continues to grow with additional archived webinars, new measurement tools and practical guides.

Advisors to the National Center suggested a need for a document that would function as a "primer" for the IPECP community as a companion to the measurement collection. Recognizing that measurement (psychometrics) is a science best left to the experts, they realized that a resource was needed to describe what goes into the process of selecting an IPECP measurement instrument. For those of us who are not measurement experts, it is good to know the questions to ask and the problems to address when we approach measurement issues and concerns. Therefore, I asked two accomplished experts in assessment and evaluation, Drs. Connie Schmitz and Michael Cullen, to write this primer on IPECP measurement. Both have substantial experience working with a variety of clients in the health professions in many sectors, including interprofessional collaboration in healthcare and education.

THE PURPOSE:

The primary purpose of this primer is to provide basic information about good practices and processes in measurement instrument development and use. Readers will find foundational information, tips for practical application and additional resources to explore. My hope is that this primer will help guide your decision making when selecting an IPECP tool, whether for research, assessment, or program evaluation. Those who are committed to measurement in IPECP, especially those who are new to the field, will find this primer helpful as a first step to lay a solid "foundation for research and clinical endeavors."

THE CONTENT:

In addition to introducing the concept of validity evidence, the authors describe the critical importance of defining the purpose of measuring IPECP, defining what questions to ask, and importantly, detailing the outcomes of interprofessional education (IPE) and interprofessional collaborative practice (IPCP) interventions. These are critical elements when selecting methods for effective measurement. Based in classic measurement theory and focused primarily on a quantitative approach to validation, the authors provide basic information to guide the IPECP community in what to look for when selecting an instrument and how to appraise the strength of its validity evidence.

> *"Science rests on the adequacy of its measurement. Poor measures provide a weak foundation for research and clinical endeavors."*
>
> Foster and Cone, 1995[1]

Barbara F. Brandt, PhD
Associate Vice President for Education
Academic Health Center Office of Education
University of Minnesota (Twin Cities)
Director, National Center for Interprofessional Practice and Education

Evaluating Interprofessional Education and Collaborative Practice: What Should I Consider When Selecting a Measurement Tool?

INTRODUCTION

The problem of knowing whether or not our measurement tools are measuring the "right" things; of whether the scores are reliable and true; of whether the decisions we make based on the scores are useful and fair... is not a new problem. It is an old problem, an interesting problem, and one that is common throughout education, health care, and research. We call this problem "validity," and understanding the process of validation is central to the work we do in interprofessional education (IPE) and interprofessional collaborative practice (IPCP) or IPECP. To paraphrase Foster and Cone,[1] our capacity to understand the effects of IPECP on the Triple Aim[2,3] of improving the experience of care, population health, and reducing per capita costs depends on the adequacy of our measurement tools.

Lack of existing tools with potential relevance to IPECP is not a problem. Understanding their reliability, validity, and utility is. A growing number of surveys, questionnaires, learner or team assessment instruments, and course or curriculum evaluation tools, can be found on the Canadian Interprofessional Health Collaborative (CIHC) website (http://www.cihc.ca) and on the National Center's Resource Exchange website [https://nexusipe.org/measurement-instruments].

Additional instruments have been identified in a recent literature search of "teamwork" in IPECP settings.[4] Typically, when searching for high quality tools suitable for particular purposes, we turn to the scholarly literature to find systematic reviews of existing instruments and evidence of their validity. We are aware of only two such reviews published for the IPECP field.[5,6] Each used different databases and different search terms, and they focused on different aspects of IPECP. Collectively, they address only some of our measurement needs. Thus, there is still a lot we don't know about the validity of the data generated by all of the tools circulating in the IPECP field.

Moreover, it would be very difficult to systematically review all of the instruments with potential relevance for IPECP. This is because our work spans many professions, disciplines, and subfields, such as those shown in Table 1. A true composite search would be Herculean in nature and take years.

Table 1: Examples of Professions, Disciplines, and Subfields Involved with IPECP

EDUCATION	Curriculum, instruction, student learning, faculty development, staff training, learner assessment, program evaluation, research methods
HEALTH CARE	Nursing, medicine, pharmacy, dentistry, veterinary medicine, food dietetics, nutrition, rehabilitation/physical therapy, occupational therapy, and other
PUBLIC HEALTH	Health care administration, health promotion, disease prevention, quality improvement, patient safety, informatics
HEALTH CARE RESEARCH	Population health indices, health care workforce, health care costs, patient satisfaction, patient outcomes, biostatistics
SOCIAL CARE	Social work, spiritual care, ethics
PSYCHOLOGY	Counseling, learning theory, professional development, professionalism, organizational development, assessment and research methods

There are also many possible things we might want to assess or evaluate with IPECP. In IPE, for example, we typically focus on outcomes we wish to change in our learners, i.e.: knowledge, skills, behaviors, and affective states (see Table 2 for examples). One summary of existing tools for IPE suggests that a majority are designed to measure attitudes, and most are subjective in nature (e.g., self-report) vs. objective (e.g., tests, standardized observations).[7] If true, there is a lot we are not measuring. Similar lists of IPCP outcomes could be delineated, as well.

Table 2: Education Outcome Categories with Examples Relevant for IPECP

KNOWLEDGE	SKILLS	BEHAVIORS	AFFECTIVE STATES
Knowledge of…(e.g.) • Own profession • Other professions • Job duties • Cost-effective care • Patient centered care • IPCP care pathways • Quality measures • Teamwork • Patient safety • Health care systems • Triple Aim	**Skilled in…(e.g.)** • Pager etiquette • Hand-off transitions • EMR documentation • Patient safety protocols • Leading effective team meetings • Communication • Conflict negotiation • Collaborative practice, leadership	**Demonstrates…(e.g.)** • Professionalism • Ethical decision making • Timely consults • Collaborative decisions for care transitions • Effective end of life family conferences	**Has…** • Attitudes • Beliefs • Feelings • Perceptions • Self-confidence • Self-efficacy • Locus of control

Trying to systematically review all of the possible tools for IPECP would also be challenging, precisely because the spectrum of potential tools is so large. Some instruments may be designed to assess or solicit responses from individuals; others may focus on teams, the clinical practice site (e.g., intensive care unit), or organization (e.g., hospital or clinic). Within the framework of knowledge, skills, behaviors, and affective states, instruments may be more subjective or more objective in design, as shown by the examples in Table 3.

Table 3: Examples of Measurement Tools by Respondent Level and Outcome Category

RESPONDENT LEVEL	KNOWLEDGE		SKILLS		BEHAVIORS		AFFECTIVE
	SUBJECTIVE	OBJECTIVE	SUBJECTIVE	OBJECTIVE	SUBJECTIVE	OBJECTIVE	SUBJECTIVE
INDIVIDUAL RESPONDENT	Pre-post self-assessment	Multiple choice test	Pre-post self-assessment	Ratings of individual simulated performance	Self-reflection inventory	360 degree evaluation; Situational judgment test	Opinion survey: Pre-post confidence
TEAM	Pre-post team self-assessment	Team quiz	Pre-post team self-assessment	Ratings of team simulations	Team debriefing	Observation rating tool	Summary of team interviews
ORGANIZATION	Key leader assessment of needs	Readiness for IPE audit	Trainer feedback on course	Review of quality measures	Self-study	External site visit, review of documents	Climate survey

Subjective: e.g., self-report, self-assessment
Objective: e.g., standardized tests, observed by others using standardized methods, systematic reviews of logs

The breadth of the field, the diversity of measurement needs, and the scope of potential instruments (both known and unknown) combine to pose significant challenges for practitioners and measurement experts alike. Unfortunately, there are no easy answers to the questions, “Which are the best tools?” or “Which tool(s) should I use?”

PURPOSE OF THIS PRIMER

The primary purpose of this primer is to raise awareness among practicing IPECP professionals of the meaning of validity and the questions we need to ask ourselves in measuring desired outcomes of IPE and IPCP. We won't tell you what outcomes to measure; we can't tell you which tools to use; and we don't provide a list of favorites. We will, however, provide guidance on what to look for when selecting an instrument, and how to appraise an instrument's evidence of validity. Along the way, we will attempt to translate some key measurement principles into lay language. We will spend time deconstructing the validity evidence from one instrument in detail.

For readers who wish to dig deeper, our Appendix contains a glossary of measurement terms; guidance on how to interpret validity data; a chart showing common threats to validity; and materials for further reading. We do this in order to promote a common language and shared understanding.

Our ultimate goal is to enable you to make wise choices when selecting instruments, and to ask informed questions if you are involved in developing new instruments.

Some clarifications before we proceed. Depending on their discipline and place of training, measurement professionals use different words to describe similar things. In this primer, we use the terms "tools" and "instruments" interchangeably. With either term, we are referring to a spectrum of data collection devices as shown in Table 3. By "assessment" we mean the assessment of individuals (learners, employees) or teams. By "evaluation" we mean the evaluation of courses, curricula, programs, practice models, sites, etc. Evaluation often relies on assessment data (along with other types of data). We alternate the word "scores" with "responses:" both refer to whatever types of data result from the application of assessment/evaluation tools.

WHAT IS VALIDITY AND WHY DOES IT MATTER?

We begin with a brief review of what is meant by validity and why it matters. To put it in laypersons' terms, valid conclusions are those which can be trusted. We reach a state of trust if various arguments convince us that the conclusions are well reasoned and in some manner confirmable.

The process of validation is, in itself, a research study. It begins with a clear purpose and definition of what's being measured (e.g., subject matter content, psychological constructs, job tasks), followed by a set of claims (hypotheses) for what the scores mean. These claims are then tested empirically by administering the instrument to intended subjects and using their scores to calculate validity statistics.

During its development, an instrument typically goes through several iterations before a psychometrically sound version can be achieved. Once released into a community, researchers seek additional evidence from other settings that similar validity statistics can be obtained. Researchers who develop an instrument may collect initial evidence of validity, only to have subsequent researchers challenge the findings or contribute additional validity evidence. This builds a line of research which remains open to scrutiny and improvement. The rigor of this process is important.

As stated previously, good research on IPECP depends on having valid data from well-designed measurement tools and processes.

> *"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose."*
>
> AERA, APA, NCME. Standards for Educational and Psychological Testing, 1999[8]

> *"The validity of a proposed interpretation and use [of scores] depends on the plausibility of the claims being made, and validation involves the evaluation of these claims."*
>
> Kane, 2013[9]

COMMON MISCONCEPTIONS ABOUT VALIDITY

- **Validity is a property of the instrument**
- **If an instrument is found valid once in one setting, it is automatically valid in another**
- **Instruments are either valid or not valid (all or nothing approval rating)**

It is easy to misunderstand validity. Validity is not inherent to an instrument (although we often speak as though it is). We don't validate a tool; we validate our interpretations and intended uses of scores. This is a subtle but important distinction. Scores are collected via the tool from samples of respondents under particular conditions.

By "conditions," we mean such things as:

- Mode of administration – how the tool is introduced and presented (e.g., online, face to face)
- Timing – when the tool is administered and the amount of time people are given to respond
- Incentives and disincentives to participants for responding
- Consequences of results (e.g., low- vs. high-stakes assessment)

Measurement error = *deviation from "true" scores due to factors unrelated to what the assessment or evaluation was designed to measure.*

Reliability = *precision of measurement; consistency, reproducibility of scores.*

Whenever data from a tool are collected, a certain amount of measurement error is present. A properly designed instrument with standardized procedures and scoring rules can control some of the sources of measurement error. When that happens, the reliability of scores increases. Reliability refers to the precision with which an intended construct or outcome is measured. In theory, a precise measurement process is one that would produce the same scores if given twice to the same people under the same conditions with no intervening event. For example, if you were to step on and off a weight scale twice in a row in quick succession, the same number should come up both times. Other terms that characterize reliable scores are: stable, consistent, reproducible, or repeatable.

Errors can be systematic or random. You can think of systematic errors as those reflecting problems with the instrument itself (such as ambiguous wording, unclear instructions), or the conditions surrounding its administration (e.g., poor timing, inappropriate incentives, technology failures with online submission). Random errors are everything else related to how a respondent might respond on a given day, and cause him or her to respond in ways that are not "true." Respondent motivation, age, background, personality traits and states can all affect responses unpredictably. For example, a group of health care providers who are completing a team climate scale after a stressful day and no sleep may score items differently than they would have, had they been well-rested.

The amount of measurement error present in any given situation lowers the reliability of scores, which then lowers the validity of an assessment or evaluation. Because validity is not a property of tools, but an inference we make from obtained scores, and scores may vary by setting, participants, and conditions, validity statistics often fluctuate from one administration to another. The goal is to understand and minimize unintended fluctuations. Speaking hypothetically, with a perfectly valid assessment or evaluation, all of the differences reflected in the instrument scores would be attributed to real differences between individual people (or groups, or organizations). In a completely non-valid assessment or evaluation, all of the differences would be attributed to measurement error – that is, random, extraneous factors, unrelated to what the assessment process was designed to measure.

Validity is not an all or nothing proposition. In fact, we refer to validity statistics as estimates. How do we estimate the amount of validity present in a set of data? Some validity statistics are based on correlation coefficients, signified by (*r*), as in $r = 0.65$, and are expressed on a scale where zero represents complete absence of validity.

For instance, in a criterion-related validation study, a correlation of 1.0 between an assessment score and an outcome measure it was intended to predict would indicate perfect validity; a correlation of 0.0 would indicate that scores are totally random, completely unassociated with the intended outcome. In reality, we don't expect to see either 0.0 or 1.0, but levels high enough and consistent enough over time and place to provide a reasonable level of trust.

Validity statistics (or validity data, validity estimates) = *broad term that includes different types of results from studies of an instrument's content, the internal structure of its items, and the relationships between its scores and other variables (e.g., other tests, surveys, ratings from patient satisfaction surveys, quality or cost indicators).*

Other statistics quantify the degree of validity as a percent (%) of variation in the scores that can be attributed to real differences or relationships, vs. measurement error. For example, developers may investigate the amount of variance in an outcome measure that is associated with factors that may influence scores (such as gender, age, instructor, or class). On a 100% scale, a higher percentage of explained variance (real differences/relationships + known factors) is better than a lower one.

FIVE SOURCES OF VALIDITY EVIDENCE

Earlier we said that the validation process begins with a set of claims (hypotheses) about what the scores mean. These claims are then tested through a process of collecting and analyzing data from a sample of people from the intended population. What exactly are these claims? There are five general types, each with a different set of potential sources for validity evidence.[8] We briefly explain them below. For ease of explanation, we will use the word "assessment" (rather than "assessment and/or evaluation"). Similarly, we will use "scores" (rather than "scores/responses"), and "domain" to signify an instrument's content. This content may be formatted by subject matter, job-related skill sets, psychological constructs, outcome variables (e.g., amount of ICU nurse turnover), or other frameworks.

CLAIM 1: Content

The first claim is that the items, cases, or tasks included in an assessment adequately represent or reflect the intended content domain. For example, if you are trying to measure the domain of "communication skills," the items on the tool should adequately cover all of the important communication skills in question (as defined by the authors). Since a theoretically infinite number of items, cases, or tasks could be written to represent "communication skills," the method of selecting or creating them needs to be justified. Types of evidence that support this claim might be that the content is based on accepted theory, previous research, or observed behavior. Or, an author might report that the case scenarios written for a simulation, for example, were based on a systematically drawn sample of patient complications that were found in chart audits or the literature. More common to IPECP, researchers may state that items were written by experts or practitioners in the field and mapped to core competencies. Ideally, this type of content would be approved through a structured consensus process with key experts, stakeholders, or by other methods.

CLAIM 2: Response process

The second claim is that when presented with an assessment tool or situation, participants understand what they are expected to do and attempt to respond in a manner fitting with the true intention of the assessment. Whether participants answer correctly or incorrectly is not the issue.

Rather, the issue is whether the assessment provides them with opportunities to knowingly demonstrate what they know or don't know, can or cannot do, believe or don't believe, relative to the domain. Types of evidence that support this claim include qualitative data gained from interviews during which developers ask respondents to read the tool and "think aloud" what the instructions ask them to do, what each item or task means, and how they would go about responding. Debriefing sessions following simulations may also uncover participant misunderstandings in what was expected, what certain prompts meant, or how much time they had. For tools that are scored by other people, interviews with raters can clarify whether they understood and applied scoring guidelines similarly.

How else would one know if an assessment's response process was flawed? Large amounts of missing data, indiscriminate response patterns (e.g., "fives" are checked for every item), and unaccountably large differences between groups are clues. Additionally, some types of reliability data can support or challenge a claim of correct response processes. For example, if multiple raters are used in an assessment, low agreement among their scores for the same participant suggests they aren't employing the assessment tool similarly or correctly. (This is known as inter-rater reliability). Or, if students taking the same test twice within a short amount of time without any intervening training or experience answer the same items very differently, we can infer that they misunderstood the questions, were guessing, or were answering randomly. (This is known as test-retest reliability).

CLAIM 3: Internal structure

The third general claim concerns the internal structure of the tool and the relationships among its items. If a single domain is to be measured (e.g., "collaboration"), then we seek evidence that scores to each item in the tool correlate highly with those from every other item. The best known statistic for this type of internal consistency reliability is Cronbach's alpha, which is represented by a reliability coefficient (r) and is expressed on the same 0.0 – 1.0 scale as discussed previously. If multiple domains are to be measured (e.g., "collaboration," "communication skills," and "organizational incentives for teamwork"), we seek evidence that item scores within each of these domains are inter-correlated. Other methods involve factor analysis (exploratory and confirmatory), analyzing item difficulty and discrimination, and calculating item-total correlations.

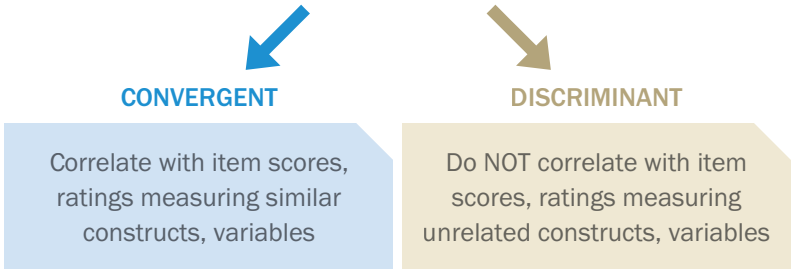
Cronbach's alpha = *degree to which items within a scale or tool are inter-correlated. High correlation suggests the items reflect an underlying construct. In general, we seek reliability coefficients around r = 0.80 or higher.*

CLAIM 4: Relationship to other variables (data)

Scores become meaningful when we understand how they relate to the world outside the instrument. Several claims can be made that involve the relationship between assessment scores and other types of information. For example, one claim might be that scores from a new tool designed to measure interprofessional communication skills will correlate with scores from similar, more established instruments of uniprofessional communication skills. If they do, our confidence in the content of the new tool increases. Another claim might be that the scores from instruments measuring attitudes will correlate with data reflecting behaviors. For example, we might look for correlations between students' beliefs about the value of IPCP and the number of IPE classes they sign up for voluntarily (assuming there are a lot of options). One might look for correlations between self-assessed skills in conflict negotiation and performance in this area as judged by trained observers during simulated team events. One might look for correlations between individuals' self-reported attitudes towards IPCP, and the results of their 360-degree evaluations from coworkers. If expected relationships between assessment scores and these other variables are found, our trust in the tool increases.

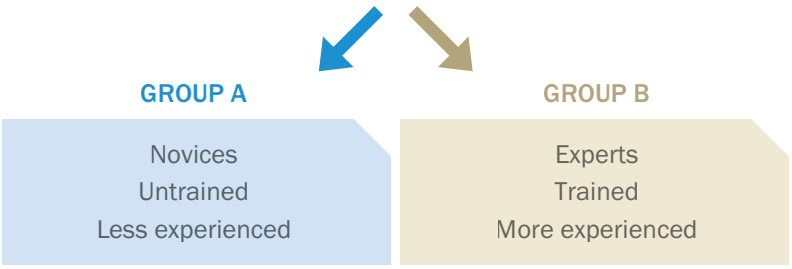
To further establish validity, we might also try to confirm that scores do not correlate with measures of knowledge, skills, or attitudes which we believe to be unrelated to the tool's domains. For example, we would not wish that ratings from a self-assessment tool on "teamwork" to correlate with a person's IQ scores, or with the number of recent nights spent on call. If they did, we would have to concede that our tool is measuring, to an uncomfortable degree, general mental ability, or work-related fatigue, rather than teamwork. A variation of this approach is shown with our deconstructed example, the Nurse-Physician Questionnaire. Together with the examples listed above, these analyses produce what is known as convergent/discriminant validity evidence.

Figure 1: Scores from Test A



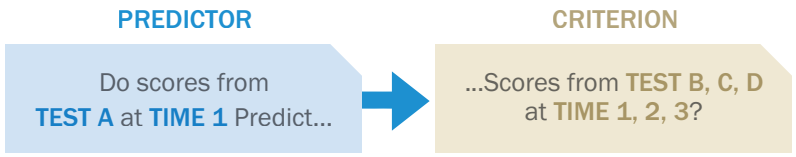
Another kind of claim is that a tool will differentiate between groups of people who would be expected to score differently based on an underlying theory or assumption. For example, if we theorize that persons develop IPCP expertise over time, we might posit that experienced professionals working in high-quality, team-based environments would score higher on a self-assessment tool measuring their IPCP competencies than persons who are new to clinical practice. If they do, we are more likely to believe that the IPCP competency tool is measuring the right things. If those differences aren't found, then we might suspect that there is something wrong with the IPCP competencies, or the response process, or with the selection of the comparison groups, or with our theory about score meaning.

Figure 2: Scores from Test A Differentiate Between



Lastly, another important kind of claim involves prediction. Do the opinions or scores from our instrument predict future attitudes, choices, or performance? Can the scores from a tool measuring IPCP knowledge, skills, or attitudes during the preclinical years predict the trainees' level of satisfaction with team-based care environments, as reported during their clinical years? Can trainees' clinical scores predict their 360-evaluation scores obtained 1-2 years later from coworkers in the practice setting? Through methods of correlation and statistical regression, we can examine the strength of these associations. If the associations are strong, we gain further confidence that baseline measures are examining durable and important qualities. Such information can help inform IPECP research, as well as our curriculum development and program evaluation processes.

Figure 3: Predictive Validity



CLAIM 5: Consequences of assessment

The last type of claim concerns the ultimate benefit of assessment, and the extent to which the use of scores leads to more accurate decision-making, improved planning, or better outcomes. For example, if scores of team performance are associated with the Triple Aim, this justifies our use of the tool for team evaluations, for developing more team-based care models, for aligning incentives that promote teamwork. Conversely, a claim for consequential validity might involve evidence that the assessment did not lead to unintended outcomes, such as marginalization of particular groups, heightened conflict, or deeper stereotyping across the professions.

HOW MUCH VALIDITY EVIDENCE SHOULD A TOOL HAVE?

Given the five potential sources of validity evidence, and the multiple ways in which claims could be investigated, practitioners often want to know, "How many types of validity data should be gathered or cited?" before using the instrument.

In general, every instrument should provide validity evidence for Claims 1-3 (content, response process, and internal structure). Beyond that, the amount and type of validity data depends on the instrument, its purpose, and the proposed validity argument.[9] Speaking practically, tools being used for high-stakes assessments, evaluations or for research require more validity evidence than instruments being used for low-stakes purposes in local settings. Many learner assessment and most course evaluation tools, for example, fall into this category. Even here, however, validity evidence may be important if the assessment or evaluation tool is to be used across IPE offerings; if it is testing a theory of IPCP development (for example) over time; or if it seeks to build on existing research or contribute to new (generalizable) knowledge.

The main justifications for collecting validity data for Claim 4 (relationship between scores and other variables) and Claim 5 (the consequences of assessment) are that these types of evidence give meaning to the scores; they help us understand how to interpret and use the scores. It is possible for an instrument to have a well-defined domain, but what if that domain is unrelated to effective performance in the clinical setting? It is possible for an instrument to generate very reliable scores that, sadly, don't correlate with or predict anything worth knowing. Until we understand how scores from an instrument relate to external variables in the real world and examine their consequences of use, we should continue to question what they mean and use them with caution.

THINGS TO CONSIDER WHEN SELECTING A TOOL

Now that we've clarified the definition of validity and potential sources of validity evidence, let's turn to the process of selecting an instrument. There are three main areas to think about:

1. **RELEVANCE OF THE TOOL FOR YOUR SITUATION**
2. **THE STRENGTH OF THE TOOL'S VALIDITY EVIDENCE**
3. **OPERATIONAL CONSIDERATIONS**

We will discuss each of these areas separately, spending the majority of our time with the strength of validity evidence. To do this, we will illustrate the process of appraising validity evidence presented for a Nurse-Physician Questionnaire. One caveat: although we strive to use the best tools possible, we do recognize that few instruments have gone through extensive development and research. Tools for which the authors have provided information and satisfactory evidence for a majority of the questions listed on the following pages are worth considering.

1. RELEVANCE TO YOUR SITUATION

You need to think deeply about the ideas and questions that matter to you. Knowing some of the published literature in a given area is very helpful for clarifying what you're trying to measure.

Table 4: Relevance to Your Situation

CHARACTERISTICS	SAMPLE QUESTIONS TO ASK YOURSELF
CONTENT	• Does the tool cover the content domain that I am interested in? • For example, does it: (a) Cover what I teach? (b) Reflect the goals of our program? (c) Address the research questions I am asking?
PURPOSE	• What is the purpose of this tool, and to what extent does it match my reasons for assessment/evaluation/research?
POPULATION	• For whom is this tool intended? • What types of individuals (teams, programs, sites, etc.) were included in the sample during tool development, or during its subsequent use? • To what degree does this sample resemble my population? • How large was the sample? • Was the tool used successfully across different professions and programs? • Is there information on how these different populations responded?

Take the example of "teamwork." There are a variety of teamwork constructs an educator, researcher, or clinician may be interested in. These range from (a) the knowledge, skills, and attitudes required of team members, to (b) the processes that characterize teams (such as team mental models[10, 11]), to (c) team climate[12] and (d) the processes that influence team effectiveness – such as coordination, cooperation, and communication. The definition of the very word, "teams," is complex, as found in the review article cited earlier.[6] Are you interested in assessing stable, "bonded" teams comprised of people who routinely work with one another; fluid teams of people who change with every shift; or people responding to a mass casualty emergency? Which professions are typically represented on your teams? Do these teams work in community clinics; independent dental offices; hospital floors, or operating rooms? Certain items in a tool may transcend these circumstances, but others may not. Additionally, other themes that are critical to your situation may be left out of the tool altogether.

In terms of purpose, some tools may be developed for purposes of selection (e.g., hiring, assignment to a team, determining which clinical practice sites are most ready for IPE). Others may be crafted for development (e.g., training, coaching). Others are designed to evaluate educational interventions. The length of your IPE intervention (e.g., seminar, workshop, course, or curriculum) will determine to a significant extent the types of outcomes that can be expected of participants. This, in turn, may influence the appropriateness of the items on a given tool for your needs. Lastly, some instruments are intended for research on processes, underlying traits, or constructs, such as professional identity formation. While you can learn something from looking at a variety of tools, it makes sense to choose one that was designed primarily for the same purpose as yours.

It is valuable to look at the characteristics of the respondent sample on which the tool was piloted, as well as the types of groups that were involved in subsequent applications. Most tools are developed with a specific population in mind. Given the interdisciplinary nature of IPECP, however, the constructs you are interested in probably cross disciplinary boundaries. It is certainly possible to use (or adapt) a tool that was developed for another setting (e.g., business vs. health care), but you need to be careful. The key question is whether you can expect the validity data gathered in the original setting to generalize to your population. Remember, just because a tool produced strong validity data in one setting does not mean it will do so in yours. The more you change an instrument to fit your population, the greater the need to collect local validity evidence.

Adopt with Care: *Validity data collected in the original setting may not generalize to your population if the two populations are different.*

2. STRENGTH OF VALIDITY EVIDENCE

Most of the information you need to answer questions about the validity of tools published in reputable journals can be found in their methods and results sections. As you work through the questions below, feel free to refer to the Appendix for help: Glossary of Measurement Terms, Rules of Thumb When Appraising Validity Data, and Common Threats to Validity.

Table 5: Validity Evidence

VALIDITY CLAIM	QUESTIONS TO ASK YOURSELF
CONTENT	• What provided the theoretical or practical framework of the tool? • How were the items/tasks/cases selected? • What evidence suggests that the tool neither under-represents the construct (content domain) I am interested in, nor introduces construct-irrelevant variance?
RESPONSE PROCESS	• Are the instructions to respondents clear? • Are the items clear? Do they ask only one thing at a time? • Are the response options clearly labeled, and congruent with what the question is prompting the respondent to do? • How was the tool administered? What were the conditions of its administration? • What did the developers learn during pilot testing or debriefings about the response process? • Are there any reliability data that suggest the scoring guidelines were applied consistently from rater to rater (e.g., inter-rater agreement) • Are there any reliability data that suggest the scores are stable across different forms of the tool, or repeatable (e.g., "test-retest" correlations)?
INTERNAL STRUCTURE	• Is there evidence that the items in the test are highly correlated with each other (e.g., internal consistency reliability)? • Is there evidence that the items which form subscales within a test are also highly correlated? • Does factor analysis confirm the same grouping of items as theorized? • Are there other statistics that support the relevance and functionality of the items (e.g., item-total correlations; item discrimination; item difficulty)? • Is there information on the standard error of measurement? • Are the confidence intervals around the scores reasonably small?
RELATIONSHIP TO OTHER VARIABLES	• Do scores from this tool correlate with similar tests, surveys, or other measures? • Are the scores from this tool unrelated to characteristics or measures that have nothing to do with what I'm trying to measure? • Do the scores differentiate logically between groups that I would expect to answer or perform differently on this tool? • Have the scores from this tool been shown to predict future attitudes, skills, behaviors, or other outcomes?
CONSEQUENCES OF TESTING	• Is there information on the extent to which the instrument meets technical standards of fairness across subgroups? • Is there evidence that use of this tool improves decision accuracy, program planning or policies, or leads to better outcomes?

EXAMPLE: THE NURSE-PHYSICIAN QUESTIONNAIRE

CLAIM 1: Content

To best understand how validation works, let's turn to an example. In 1991, S. M. Shortell et al. reported results from a national study of 42 intensive care units involving over 1,700 respondents.[13] These authors had developed an Intensive Care Unit (ICU) Nurse-Physician Questionnaire based on a model of organizational and managerial factors affecting ICU performance.

The organizational component of the model had two areas:

- ***Leadership*** (ability of nurse and physician leaders to set high standards, communicate goals, respond to changing needs and to unit members' needs and perspectives) and
- ***Culture*** (team satisfaction orientation vs. people security orientation vs. task security orientation).

The managerial component contained three areas:

- ***Communication*** (openness, accuracy of information, timeliness, understanding, effectiveness of between-shift communication, and satisfaction);
- ***Coordination*** (written plans and schedules, treatment protocols, policies and procedures, efforts and interactions, between unit coordination and relationships); and four different modes of
- ***Problem Solving*** (open/collaborative, arbitration, avoidance, and forcing).

These components were theorized to impact two types of performance outcomes: ***team cohesion*** and ***unit effectiveness***, which was defined as achieving technical quality of care, meeting family needs, and low nurse turnover. This model, which was based on previous health services research, supports Claim 1, Content Validity.

CLAIM 2: Response Process

The article then described the process of developing Likert-scale questionnaire items to measure all of the constructs (e.g., team-satisfaction orientation) representing the organizational and managerial components of this model and the two performance outcomes. The authors conducted a series of pilot studies to test item wording and content. In so doing, they recognized the need to have separate (but parallel) instruments for nurses and physicians, because questions had to be asked somewhat differently in order to achieve the same meaning. The researchers also learned the importance of testing *within*-group (i.e., nurse-nurse) as well as *between*-group (i.e., nurse-physician) relations. Thus, they elaborated the instrument to measure, for example, openness of communication between physicians, as well as communication between physicians and nurses. Similar dyad distinctions were made for other constructs (e.g., accuracy of information received), and for all four problem-solving approaches. After collecting and reviewing the data, the authors also revised a few items with low reliability. Although it was not clear from the article how they discovered problems with item meaning (e.g., interviews with respondents?), or exactly how they investigated reliability, it was clear they had attended to Claim 2, Response Process.

CLAIM 3: Internal Structure

The revised Nurse Physician Questionnaire was then administered to a national sample of 42 medical/surgical ICUs in 40 non-governmental hospitals. Respondents included nurses and physicians (full and part-time) associated with the ICU, residents, and the physicians who accounted for the greatest number of ICU admissions. The article reported the Cronbach's alpha reliability coefficient for each of the constructs in the model. For example, the construct of nurse leadership was measured with eight items. The results for nurse leadership (r = 0.87) meant that these eight items were inter-correlated and, taken altogether, represented a good measure of nurse leadership. Across the 28 constructs, reliabilities varied from a low of r = 0.68 to a high of r = 0.88.

Additionally, the authors used a method called factor analysis, which groups items together based on how the answers correlate with one another statistically. Item responses that statistically correlate with each other are said to "load" on to a central idea (factor). The hope is that the factor loadings substantiate the way the items are grouped on the instrument, and thereby confirm the constructs being theorized by the model. Factor loadings are also expressed as a coefficient. Loadings that are less than 0.30 or 0.40 are generally not considered very meaningful.[14] In this study, the factor loadings were all above 0.40 and generally ranged much higher. Together with the reliability coefficients, the factor loadings provided strong evidence for Claim 3, Internal Structure.

CLAIM 4: Relationship to Other Variables

The authors also collected perceptions of the organizational culture through a survey of ward clerks and members of the hospital's top management team (e.g., CEO, VP for Nursing, and VP for Strategic Planning, Marketing, and Human Resources). They did this in order to examine the correlation between scores for organizational characteristics, managerial processes, and outcomes. Essentially they were looking for evidence of convergent and discriminant validity. According to their theory, ICUs with high scores in organizational characteristics, and specifically a culture that was oriented toward team satisfaction, would also score higher in managerial processes (i.e., coordination, communication and conflict resolution), as well as the desired performance outcomes of team cohesiveness and perceived unit effectiveness (convergent validity). According to their theory, ICUs with negative organizational cultures were those in which superficially smooth relationships and unquestioned obedience to authority were the norm ("person-security"), and those in which rigid conformity to tasks ("task security") cultivated perfectionistic, competitive, and mistrustful behavior. The authors theorized that units high in either person-security or task-security scores would score lower on managerial processes and outcomes (discriminant validity).

This turned out to be largely the case, as shown by a pattern of correlations among respondents' answers. Responses for items measuring positive nurse and physician leadership, for example, were positively correlated with those measuring high team satisfaction (r = 0.49), better ICU coordination (r = 0.52), better communication (r = 0.40), more open problem solving (r = 0.47), higher team cohesion (r = 0.49), higher perceived technical quality of care (r = 0.48), higher perceived family satisfaction (r = 0.32), and lower nurse turnover (r = -0.29).

The converse pattern was also true: low leadership scores and ICUs with strong person-security and task-security cultures were negatively correlated with desired processes and outcomes. What makes these correlations important is not just their magnitude, but their overall pattern which corresponded to the theoretical model.

To further confirm these quantitative findings, the authors completed in-depth follow-up interviews with nine ICUs, randomly selected based on their overall ratings (high, medium, low). In addition to speaking with nurse and physician leaders and a sample of nurses (all shifts), other staff (e.g., respiratory therapists), and administrators, the authors spent time observing team interactions during patient rounds, patient care, and shift changes. These interview and observational data, together with the quantitative results substantiating convergent and discriminant validity, constitute excellent examples of evidence for Claim 4, Relationship to Other Variables.

CLAIM 5: Consequences of Testing

What of Claim 5? Were there any data in support of the Consequences of Testing? As with many studies, this article fell somewhat short of analyzing the impacts of using the Nurse-Physician Questionnaire as an assessment tool. Their primary aim was to establish the content, response process, and internal structure validity of the questionnaire, and to test the hypothesis that people working in ICUs with a team-oriented culture exhibited greater cohesion and higher perceived performance outcomes. The authors did, however, report on the utility of the tool. For example, they discussed the time burden for respondents to complete the questionnaire, the ease of its administration and analysis, and the potential for its scores to be used for benchmarking with other ICUs and for organizational development purposes. It was left to "further research" to complete the full validity quotient.

Other Examples

For readers who wish to see additional examples of how to appraise validity data from existing instruments, we recommend two articles from our reference section. The first, by D. A. Cook and T. J. Beckman, analyzes an instrument used by urologists to assess symptoms of benign prostatic hypertrophy in men.[15] Although this is a clinical assessment tool and not relevant to IPECP in terms of content, its validation process remains the same. Their concrete and familiar examples may help readers extrapolate to their own fields. A second primer, authored by T.A. O'Neill and colleagues, analyzes the evidence for the "Teamwork Knowledge, Skills, and Ability (KSA) test."[16] This in-depth review identifies measurement weaknesses, as well as strengths, from a validity standpoint

3. OPERATIONAL CONSIDERATIONS

We return now to our third and last set of considerations for tool adaption: Operational Considerations.

Table 6: Operational Considerations

CONSIDERATIONS	QUESTIONS TO ASK YOURSELF
USER ACCEPTABILITY	• Is there anything about the tool – its appearance, conditions of administration, scoring or reporting – that would detract from its plausibility or acceptability to my stakeholders (e.g., persons sponsoring the tool, faculty teaching the curriculum, respondents completing the survey)? • Do the instructions or items reflect any kind of cultural bias? • Would any wording need to be translated or changed for my population?
TOOL DURABILITY	• Would it be possible for a respondent to "fake" the answers to this tool? • How likely is it that respondents will answer in socially accepted ways? • In situations where the test questions and answers are to be secure, would it be possible for respondents to improve their scores by guessing, practicing ahead of time, or being coached?
TRAINING IMPLICATIONS	• How hard would it be to train participants, teachers, assessors, or other personnel for this assessment to produce valid score data?
ADMINISTRATION AND SCORING	• Are the instructions for administering the assessment and scoring the tool clearly provided so it can be standardized across administrators and sites, and implemented consistently over time?
COSTS	• Is the tool copyrighted? What are the costs of permission? • What start-up costs are associated with conducting this assessment (e.g., equipment, IT, programming, training, statistical analysis)? • What are the maintenance costs (e.g., ongoing training, the frequency of creating alternate test forms or new items, item bank maintenance)?

If there is a good match between a given tool and your needs, purpose and population; if you are persuaded there is enough validity evidence to pilot the tool in your setting; then it's time to consider the operational aspects of tool adoption. A first step is to ensure that the stakeholders you are working with and likely to be affected by the assessment find the tool non-offensive and potentially useful, based on its face value. Sometimes, making small changes in wording can make a big difference for your audience. Language can be a key to unconscious cultural or intellectual biases.

"Test durability" concerns the ability of a tool and the conditions surrounding its use to resist non-authentic responses. By non-authentic we mean: (1) faking a performance or inflating one's estimate of one's abilities; (2) succumbing to social pressures to act, keep quiet, or answer in specific ways; and (3) cheating.

Cheating is more of a concern with high-stakes tests, which are uncommon in IPE, but could be present, as with any other test situation, if results lead to recognition or rewards. Faking "good" is a problem particularly in tests of selection, but it can surface in any self-report measure, in simulation exercises, and in site visits conducted by external evaluators. The possibility of faking is a predominant concern when using self-report teamwork measures. The research evidence is clear that, when instructed to do so (for research purposes), subjects can significantly increase their scores on self-report personality instruments through faking.[17] Faked answers are a problem because they compromise the integrity not only of the results and any follow-up actions based on the results, but also the validity data cited to support the instrument.

Fortunately, there are a number of potential remedies for both detecting and reducing "faking" on instruments. For example, readers might look for information about: (1) "lie scales" on multiple-choice tests; (2) "verification" questions for which the answer is already known from other sources; (3) "decoy constructs," so that a participant cannot easily determine which constructs being measured are the constructs of interest; and (4) instructions that warn participants about possible negative consequences of intentional faking. The other main durability concern in IPECP is the pressure exerted by the wording of the tool itself, or by the instructors/leaders conducting the assessment, or by the peer group to answer in socially desirable ways. Examples of these challenges are listed below, along with suggestions for minimizing them.

Content/Item bias. In an IPE setting, items that show a latent bias towards IPE have the potential to produce responses consistent with the bias, especially with learners who are early in their training. The bias can be seen if a tool assessing the value of IPE as a training mode, for instance, does not give respondents equal opportunity to rate *alternatives* to IPE for achieving the desired outcomes. Bias can also be seen if non-IPE approaches are described in negatively loaded terms, such as: "the *medical* model," "the *traditional* way," or "the *physician*-centric approach." A qualitative pre-pilot in your own setting with a small sample of respondents can help you discern whether the tool's scope of content restricts the respondents' ability to give their full opinion in an authentic way. It can also help you to see if any wording makes some subgroups feel denigrated or marginalized by virtue of their specialty, age, preferences, or philosophical viewpoint.

Administrator bias. The persons asking participants to complete a tool have the power to both encourage and discourage socially desirable responses. In an IPE setting, instructor influence can be seen in both subtle and not-so-subtle ways. Popular professors can unwittingly incite biased responses from students because students will want to please them. The same goes for any situation in which there is a power differential (e.g., bosses vs. employees, senior vs. juniors in a practice setting). Persons in these roles should never make personal pleas to participants (even in jest) to answer questions in a certain direction. They should never allude to the importance of getting "good ratings" from the tool for purposes of faculty promotion, course continuation, future funding, or the like.

Ways to avert administrator bias in high-stake assessments begin with separating the person who sponsors, teaches, or controls the intervention from the person administering the tool. Ensuring anonymity for certain attitudinal surveys often is a prerequisite. Additionally, in written and verbal introductions to the instrument, administrators need to clarify how the data will be used. In opinion surveys, administrators should emphasize the importance of candid and honest opinions; they should clarify there are no repercussions to so-called negative comments. With self-assessment instruments, reinforce the importance of accurate and honest ratings. The instructions could explain, for example, that: "every individual learns or progresses at his/her own rate;" "being at the beginning of a learning curve is not 'bad;'" or "saying 'I don't know' is an entirely valid answer and perhaps the only honest answer for you at this time."

Peer bias. The literature on peer pressure in any kind of small group self-reflection or activity debriefing session is sizable. Good facilitators know how to minimize, if not avoid peer pressure through the use of ground rules and non-biased, open-ended questions; by equalizing conversation time between dominant and recessive members; by asking persons to rephrase demeaning or dogmatic comments; and by redirecting questions. To minimize peer influences on individual surveys or tests, administrators can ensure there is enough privacy (physical space) between participants who are completing the tool in a real-time group setting. Another alternative is to put their tools online. Again, instructions to participants can help if they include statements such as, "please answer what's true for you – whether or not you think that opinion is shared by others."

The remaining operational considerations listed on p. 25 (i.e., training, consistency of administration and scoring, and operation costs) are fairly straightforward. Before adopting a tool, consider your operational assets and deficits. Meet with persons who would be expected to implement the tool. Some tools, such as paper and pencil opinion surveys or self-assessments instruments, may require only minimal training of data managers to implement, and their data can be easily analyzed by commercial survey software. Other assessments, such as simulated teamwork exercises, require significant planning, a variety of trained personnel, greater attention to standardization, and potentially more complex scoring. Any instrument to be used across sites requires additional layers of planning, communications, and coordination. Last but not least: estimate the costs involved. Some tools are proprietary and require a royalty fee, which is often based on the number of examinees and length of use. Even when tools are in the public domain, permissions may be required (with or without a fee).

SUMMARY

There is an old adage that says, "We measure that which we value." Based on our 10,000-foot view of assessment and evaluation needs and existing tools, the IPECP community clearly places great value on promoting a set of underlying attitudes and beliefs about the nature of interprofessional collaborative practice and the promise of interprofessional education. It has also embraced the behavioral components of "teamwork," as illustrated by the Valentine, et al. (2014) review[6] and a recent presentation by Thistlethwaite.[4] But there is work to be done. The field is vibrant with questions about the meaning of scores resulting from these instruments; it appears eager to expand its reach with newer instruments that could address the broader spectrum of measurement needs in IPECP.

In this primer, we've addressed the need for guidance on selecting measurement tools by talking first about validity and what it means. We hope that in doing so, we've increased your awareness of the amount of time, expertise, and teamwork that is required to build the high quality instruments we need if we are to successfully research the impact of IPE and IPCP on the Triple Aim. It is so tempting to simplify the validity problem by just anointing one tool over another. Science is slow, however, and it challenges us to think deeper about the problem of verifying our interpretation of results. Certainly, some tools are better constructed than others. Some have more complete documentation of their strengths and limitations; some have been replicated or cited more often than others. But as we hopefully have made clear, the validation of our interpretation and uses of scores requires us to be knowledgeable consumers of assessment and evaluation tools. And as users of these tools, our actions matter. A poorly implemented tool can produce non-valid data, regardless of how well it was constructed.

In this large and quickly moving field, it would be presumptuous of us to claim we know all that's being done to identify or develop high quality instruments. From what we can tell, however, despite the enormous number of existing tools in circulation, a large majority are lacking in even minimum validity evidence. This leaves many of us in a quandary when it comes to tool selection. It also does not appear that many measurement professionals have been engaged to support the educators and clinicians who are leading the field. Perhaps that is why we lack systematic literature reviews that critique existing instruments, and why we lack validity studies.

To move the field forward, we urge readers to seek out and consult with measurement professionals whenever possible. Persons trained in measurement carry different job titles, but they include people like us (educational and organizational psychologists), clinical and behavioral psychologists, sociologists, research scientists of various persuasions, epidemiologists, and persons specializing in psychometrics and statistics. In the U.S., they may be found in schools of public health, education, and psychology; research centers; and via professional organizations, such as the American Education Research Association (Health Professions Division), the American Evaluation Association, and the Association of Psychologists in Academic Health Centers.[18]

We also urge IPECP leaders to move towards a unified agenda in researching the validity of data from the field's most promising tools. Coordinating efforts in validity research under a consensus agenda is important given the scope of the field. To make progress, we also need to think strategically about what we've learned so far and what needs to be done. As noted by Archibald, et al. (2014), for example, few studies assessing IPE interventions have reported significant change in attitudes towards IPE.[19] This may be due to flawed content items; or to our limited understanding of the "true" trajectory of attitudinal development; to poor response processes for the respondents being surveyed; or to the inherent bias of pre-post self-report. Lack of attitudinal change could also be due to ineffective interventions! When a line of research runs into dead ends, it's time to step back and regroup.

In conclusion, we hope this primer helps support IPECP professionals as they search for useful tools. We hope that readers use it to build a common language around assessment and evaluation. We hope it supports your efforts, as members of the IPECP community, to work together to achieve the Triple Aim.

REFERENCES

1. Foster SL, Cone JD. Validity issues in clinical assessment. *Psych Assess*, 1995;7:248-60.
2. Berwick DM, Nolan TW, Wittington J. The triple aim: Care, health, and cost. *Health Affairs*, 2008;27:759-69.
3. Brandt B, Lutfiyya MN, King JA, Chioresco C. A scoping review of interprofessional collaborative practice and education using the lens of the Triple Aim. *J Interprof Care*, 2014;28(5):393-9.
4. Thistlethwaite J. The development of work-based assessment (WBA) of teamwork instruments – an interprofessional approach. All Together Better Health VII conference, Pittsburg PA, June 6, 2014.
5. Thannhauser J, Russell-Mayhew S, Scott C. Measures of interprofessional education and collaboration. *J Interprof Care*, 2010;23(4):336-49.
6. Valentine MA, Nembhard IM, Edmondson AC. Measuring teamwork in health care settings: A review of survey instruments. *Med Care*, 2014 (April); epub ahead of print (PMID: 24189550).
7. Blue A, Chesluk B, Conforti L. Assessment and Evaluation in IPE: Lessons Learned from a Multi-methods Study. Webinar posted to the NEXUS Community and Resources, Resource Exchange, accessed 6/19/14 http://nexusipe.org.
8. American Education Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington DC: American Education Research Association, 1999.
9. Kane MT. The argument-based approach to validation. *School Psych Rev*, 2013;42(4):448-57.
10. Cannon-Bowers JA, Salas E, Converse SA. Shared mental models in expert team decision-making. In NJ Castellan (ed.), *Individual and Group Decision Making* (pp.221-246). Hillsdale (NJ): Erlbaum, 1993.
11. TeamSTEPPS: *Strategies & Tools to Enhance Performance and Patient Safety*. Agency for Healthcare Research and Quality and the U.S. Department of Defense, www.ahrq.gov.
12. Anderson N, West MA. Measuring climate for work group innovation: Development and validation of the team climate inventory. *J Org Behav*, 1998;19:235-58.
13. Shortell SM, Rousseau DM, Gillies RR, Devers MA, Simons TL. Organizational assessment in intensive care units (ICUs): Construct development, reliability and validity of the ICU Nurse-Physician Questionnaire. *Med Care*, 1991;29(8):709-26.
14. Vogt WP. *Dictionary of Statistics & Methodology*, 3rd ed. Thousand Oaks (CA): Sage Publications, 2005, p. 119.
15. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med*, 2006;119:166.e7-166.e16.
16. O'Neill TA, Goffin RD, Gellatly IR. The knowledge, skill, and ability requirements for teamwork: Revisiting the Teamwork-KSA test's validity. *International J of Selection and Assessment*, 2012;20(1):36-52.
17. Ones DS, Viswesvaran C, Reiss AD. Role of social desirability in personality testing for personnel selection: The red herring. *J Applied Psych*,1996;81(6):660-79.
18. Robiner WN, Dixson KE, Miner JL, Hong BA. Psychologists in medical schools and academic medical centers. *Am Psych*, 2014;69(3):230-48.
19. Archibald D, Trumpower D, MacDonald CJ. Validation of the interprofessional collaborative competency attainment survey (ICCAS). *J Interprof Care*, 2014;28(6):553-8.

APPENDIX 1

GLOSSARY OF MEASUREMENT TERMS

WHAT WE MEASURE

Construct = an abstract idea or concept, such as "collaboration," that exists in theory and can't be directly observed (heard, felt, touched, or sensed). Constructs can be inferred from behavior (e.g., the manner in which persons speak to one another), or intuited from other signs (e.g., the number of courses co-taught by persons). An "operationalized" construct is one that has been defined in measurable terms (see "variable").

Scale = a collection of items purported to be measuring slightly different aspects of the same thing. For example, "physician leadership" (a construct) may be measured by eight items on a survey, covering eight slightly different characteristics, traits, or actions believed to represent leadership. Collectively, these eight items are said to represent a "scale."

Content Domain = all of the knowledge, skills, behaviors, or attitudes that could theoretically pertain to a defined area of interest being measured by a measurement instrument, such as "interprofessional collaboration core competencies." This term originated in education with subject matter tests. In this primer, we use the term more broadly to mean whatever content is covered in a measurement tool, or in materials (such as simulation cases, scenarios, tasks) that are necessary for an assessment or evaluation.

Outcomes = the results of any kind of process or condition, natural or planned. In education and health care, we typically consider outcomes as desired end goals of a planned intervention (e.g., course, curriculum, clinical model, etc.).

Variables = phenomena of interest to researchers that can vary within or across individuals, (groups, locations, trials, years, etc.) and can be measured (i.e., assigned a number). Variables may be operationalized constructs, simple counts (e.g., the number of participants in a class), or other types of numeric indicators (e.g., faculty-student ratio). In research studies, "independent" or "predictor" variables represent inputs to a theoretical model or system, such as individual motivation or prior experience. Examples of "process" variables include such things as the number of classes held, or the amount of time students spend viewing education content online. "Dependent" variables represent outcomes, such as post-test gain scores, hospital readmissions, or costs per visit.

MEASUREMENT ERROR

True Score = according to measurement theory, the true score is the score that would be obtained if no measurement error existed. We can never absolutely know a person's true score, because there is always measurement error.

Error = the difference between a true score and an obtained or observed score.

Standard Error of Measurement (SEM) = one of the most common estimates of error, the SEM represents how much a person's obtained score on a single test would vary from his or her true score, if s/he took the test over, and over, and over, and over (to infinity).

Confidence Intervals (CI) = the range of scores within which we would say the person's true score most probably lies. For example, we might say the person's score is 50, plus or minus 10 points. In that case, the total confidence interval is 20 points. The calculation of confidence intervals depends on the standard error of measurement. A larger SEM leads to larger CIs.

TERMS RELATED TO RELIABILITY

***Internal Consistency Reliability (ICR)* =** the extent to which responses to items grouped together in a scale or test are correlated with each other. One may infer from a high internal consistency reliability coefficient that all of the items in the scale, or in the test, are measuring the same underlying construct (competency, or outcome, etc.).

***Inter-rater Reliability (IRR)* =** overall consistency in scoring among raters for the same subjects (groups, sites, etc.). Based on correlations and rank orders, the IRR represents the extent to which raters assign high scores to the same examinees, and low scores to the same examinees.

***Inter-rater Agreement (IRA)* =** how often two or more raters agree on how a subject should be scored. It is based on the proportion (%) of examinees for which the assigned scores for each item in the test are the same across raters. If examinees are scored on a 5-point scale, for example, the IRA represents the percentage of examinees for which the same value (i.e., a 1, 2, 3, 4, or 5) were given by all raters. The IRA is a more stringent test of reliability than IRR for items that are scored on a scale (rather than yes/no).

***Test-Retest Reliability* =** the correlation between two sets of scores on a test given twice to the same individuals with little or no intervening time or intervention.

***Factor Analysis* =** an approach to understanding the internal structure of a test by examining the extent to which item responses correlate or group together around similar constructs ("factors"). This general approach can be applied in order to discover underlying factors ("exploratory factor analysis"), or to test the coherence of theory in which factors have already been proposed ("confirmatory factor analysis").

TERMS RELATED TO VALIDITY

***Validity* =** the degree to which data from a measurement tool can be trusted. The validation process involves stating a set of claims ("hypotheses") about what we intend the scores mean or represent, and then collecting empirical evidence that supports these claims.

***Validity Data (or, validity statistics, validity estimates)* =** a variety of statistics generated during the development or testing of a measurement tool to support validity claims. Validity data result from the analysis of scores / responses collected from sample populations.

***Item-total Correlations* =** the degree to which the responses from a single item within a test correlates with the total test score. If the correlation is low, the test's internal consistency reliability can be improved if that item is deleted from the test. This process can also be applied at the level of a scale. In this case, each item is correlated with the total scale score, and those with low correlations are removed. This process improves the precision of measurement.

***Concurrent Validity* =** the positive correlation between scores from one measurement tool with another tool that measures similar constructs (or content domains, processes, outcomes, etc.).

***Convergent Validity* =** similar to Concurrent Validity, the positive correlation between scores representing constructs (etc.) that are believed to co-vary (go together).

Discriminant Validity = the opposite of Convergent Validity, the negative correlation between scores representing opposite constructs or unrelated variables.

Predictive or Criterion-Related Validity = the degree to which an outcome variable (also often called the "criterion") can be predicted from an independent variable (e.g., scores).

Face Validity = this term is discouraged within the measurement community. It refers to the overall impression or face value that an instrument may have, given its general appearance, choice of terms, and content. While user acceptance and political credibility are both important to the successful implementation of a tool, these qualities do not mean the data are valid.

OTHER TERMS USED IN THIS PRIMER

Psychometrics = a field of research that focuses on the process of turning constructs into measureable variables, and the measurement properties (metrics) of instruments collecting data on those variables.

Correlation Coefficient = the degree to which two variables are related, based on their correlation of scores. The range of a positive correlation is 0.0 to 1.0. The range of a negative correlation is 0.0 to -1.0.

Item Difficulty = the proportion of examinees who answer a test question correctly.

Item Discrimination Index = the extent to which an item discriminates between test takers. That is, the question is answered correctly by people who might be expected to perform well on the test, and incorrectly by people who might be expected to perform poorly.

Likert Scale = not to be confused with a "scale" of similar items, a Likert scale is the name of a particular type of response option associated with an item. A typical Likert scale has 5 or 7 levels, or intervals, which could be marked. Likert scales are used when responses can't be answered yes/no (dichotomously), and are presumed to vary along a range from high to low, strong to weak, etc. For data to be averaged across respondents, the distances between each interval on the scale must be the same.

APPENDIX 2

RULES OF THUMBS WHEN APPRAISING VALIDITY DATA

SAMPLE SIZE: HOW BIG SHOULD IT BE?

- Ultimately it depends on the study design and type of analysis.
- The greater the diversity of people within the population of interest, the larger the sample size.
- Scores taken from samples of at least 30 people are more likely to assume a "normal curve" distribution; and the statistics (like averages) are less sensitive to "outliers" (i.e., people who score at the extreme ends of the scale).
- Samples of over 70 (per group) are usually sufficient for statistical power calculations involving comparison groups, unless the amount of difference you are trying to detect is very small.
- For multiple regression analysis, at least 50 people plus an additional 8 people per independent variable are needed.

CAN SAMPLE SIZES BE TOO BIG?

- The main deterrent to big sample sizes is the cost of collecting the data.
- With samples over 100, small differences between individuals or groups will be statistically "significant," but not necessarily very meaningful in the practical sense.

RELIABILITY (CORRELATION) COEFFICIENTS: HOW HIGH SHOULD THEY BE?

- It depends on the purpose of the assessment, but higher is always better.
- For research purposes, the reliability coefficient (r) should be 0.80 or above.
- For high stakes assessments (such as licensure, certification), the reliability should be 0.90 or above.
- For moderate stake assessments (such as end of course evaluations, year-end exams), the reliability should be in the 0.80 range and above.
- For low-stakes assessments (e.g., classroom tests created by instructors), acceptable reliabilities could be in the 0.70 range and above.
- The lower the reliability, the greater the error in your scores.

VALIDITY (CORRELATION) COEFFICIENTS WITH OTHER VARIABLES: HOW HIGH SHOULD THEY BE?

- It depends on the variables involved, but higher is always better.
- A correlation between two variables (such as a test of IPCP attitudes and a test of IPCP teamwork knowledge or skills) of 0.60 and above would be important and meaningful. A correlation less than 0.30 is usually considered to be somewhat to fairly minor.
- The more distant in time and place between a predictor variable and its outcome, the lower the expected correlation. A pattern of correlations in the expected direction, even if they are in the 0.20 – 0.40 range, might be important and useful.

- To understand the magnitude of a validity correlation, square it mathematically. A coefficient of 0.60, squared, is 36 (6 x 6 = 36). That means, 36% of the score variance in your set of data can be "explained" or "accounted for" by your test. If the validity coefficient is only 0.30, then only (3 x 3) 9% of the score variance can be explained by your test – and the rest of the variance is due to unknown factors and random error.

FACTOR LOADINGS: HOW HIGH SHOULD THEY BE?

- The higher the loading, the closer the association between that item with the group of items that make up that category / factor.
- Loadings of less than 0.30 or 0.40 are generally not considered meaningful.

MEASUREMENT ERROR

- Smaller is always better.
- The size of the SEM can be judged by looking at the range of the scale. An SEM of 10 on a test scale of 1 to 50 is quite large; an SEM of 10 on a test scale of 1 to 500 isn't so bad.

CONFIDENCE INTERVALS

- The size of the confidence interval depends on how confident or certain about the score you want to be. If you want to say you want to be "68% confident" that the person's scores lie within a particular range, the band will be smaller. If you want to be "95% confident," then the band will be much larger.
- The calculation of confidence intervals depends on the standard error of measurement, which depends on the reliability of the test.

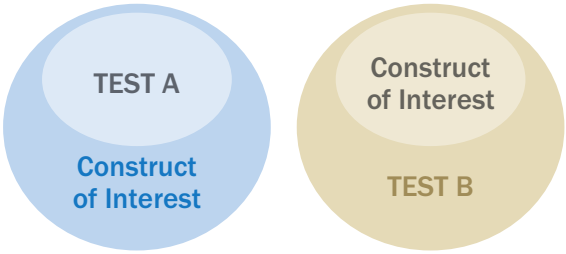
APPENDIX 3

COMMON THREATS TO VALIDITY

CLAIM 1: Content

Two of the biggest threats to content validity are (1) insufficient coverage of the domain ("construct underrepresentation"), and (2) the inclusion of items that either lie outside of the domain, or are not unique to the domain ("construct-irrelevant variance"). As shown in the diagram below, Test A fails to adequately cover all of the elements of the construct of interest. Conversely, Test B unintentionally measures things that lie outside of the construct of interest. One example of construct-irrelevant variance is when an assessment is found to have a systematic bias towards one group over another (e.g., native language speakers vs. non-native language speakers). In that instance, a large amount of error is introduced into the scores, and the tool morphs into a measure of language literacy.

Figure 4: Two Threats to Content Validity



CLAIM 2: Response Processes

Problems with response process can arise from a poorly constructed instrument (unclear instructions, ambiguous items, inadequate options for responding), or one that is not well-suited, well-timed, or administered appropriately for the participants who are completing it. Perhaps the best example of this is with certain attitudinal surveys as they have been used with early learners in IPE. This is because when surveys are given to students prior to clinical experience, (1) they don't have enough context before the intervention begins to really understand the terms being used; (2) they don't know what they don't know about IPECP; and (3) the pressures of social conformity inflate their levels of agreement with positively phrased statements of belief or opinion. Another type of threat can occur in simulated performance exams. The simulated environment, by virtue of its artificial nature, may unintentionally require the participants to perform in ways they would or could not do in an actual patient care environment. The content of the simulation may be "right," but because of such things as distorted timeframes, unavailability of equipment or equipment failure, poor attendance by team members, or poorly trained "patients," participants may not be able to respond as they normally would, or as optimally as they might.

CLAIM 3: Internal Structure

The major threat to internal structure validity in IPECP is that many of the constructs (also called factors, concepts, or traits) we are attempting to measure are difficult to define, and they overlap. Although we can conceptually differentiate between communication skills and collaboration skills, for example, individuals tend to score very much alike in both constructs. Thus, the theoretical structure of an instrument as defined by separate factors may not be substantiated empirically, which means that individuals either have "it" (i.e., a constellation of skills, values, etc. that are highly inter-correlated) or they don't have "it." Additionally, factor structures emerging from one population in one setting may not be confirmed when applied to a different population or setting. One interpretation of this irregularity is that the terms we are using – the factors we are trying to elucidate and study – are slippery and understood differently by different professionals working across disparate contexts.

CLAIM 4: Relationships to Other Variables

The major threat to Claim 4 is the unreliability or lack of validity of the other variables involved. This is especially true in IPECP, as we don't have many high quality instruments that measure the variables we want to relate to our primary instrument. It can also be quite time consuming to collect scores with outcome variables that are distant in time or place. Whatever affects the reliability / validity of the concurrent or dependent variables being measured will compromise its relationship to the scores from the primary instrument in question.

CLAIM 5: Consequences of Testing (Assessment/Evaluation)

Claim 5 represents one of the more controversial areas of validation, as it may involve social policy beyond the strict purview of the instrument and its intended purpose. One of the threats with any assessment or evaluation process is that scores will be extrapolated from one purpose, applied to another, and attached to policies which may or may not be warranted based on the data.

APPENDIX 4

FURTHER READING

American Educational Research Association (AERA), American Psychological Association (APA), & National Council for Measurement in Education (NCME). 2014. *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Brown JD. Standard error vs. standard error of measurement. Shiken: *JALT Testing & Evaluation* SIG Newsletter, 3(1), April 1999 (p.20-25). http://www.jalt.org/test/PDF/Brown4.pdf accessed 10-22-14.

Bureau of Exceptional Education and Student Services. Standard error of measurement (SEM). Technical Assistance Paper, 301959. Available from Denise Bishop (bishop@tempest.coedu.usf.edu).

Downing SM. Face validity of assessments: Faith-based interpretations or evidence-based science? *Med Ed*, 2006;40(1):7-8.

Downing SM. Reliability: On the reproducibility of assessment data. *Med Educ*, 2004;38:1006-12.

Downing SM. Validity: On meaningful interpretation of assessment data. *Med Educ*, 2003;37(9): 830-37.

Kane RL, Radosevich DM. *Conducting Health Outcomes Research*. Sudbury (MA). Jones & Barnett Learning, 2011.

Litwin MS. How to Assess and Interpret Survey Psychometrics, 2nd ed. From: *The Survey Kit*. Thousand Oakes (CA): Sage Publications, 2003.

Kasten SJ, Korndorffer J, Downing S. Validity: Giving meaning to assessments. Ch. 2 in: *Guide for Researchers in Surgical Education* (J. Capella, SJ Kasten, S. Steinemann, L. Torbeck, eds.) Woodbury (CT): Cine-Med Publishing, 2010, pp. 99-114.

Vogt WP. *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences* (3rd ed.). Thousand Oaks (CA): Sage Publications, 2005.

APPENDIX 5

THE NATIONAL CENTER RESOURCE EXCHANGE

Beyond supporting original research, the National Center is building a community of individuals and organizations interested in IPECP. At the center of this effort is the Resource Exchange, an interactive, user-supported digital library.

The Resource Exchange isn't just a place for emerging research. It also chronicles the 40-year history of IPECP, giving a unique perspective to trends through access to seminal works that have never been digitally available.

The Resource Exchange includes two special collections prepared by the National Center team:

Measurement Instruments: A curated collection of 26 measurement tools and supporting literature used for IPECP, many of which are available online. The instruments can be filtered by outcomes, subscales, and subject.

Scoping review of IPECP literature: A fully searchable literature compendium with 496 interprofessional collaborative practice and interprofessional education literature citations from 2008 through 2013. The citations focused on the relationship between IPECP and the Triple Aim as a precursor to mapping the relationships for the NCDR. The findings from this work were published in the *Journal of Interprofessional Care*.

Since it is open access, anyone can browse information in the Resource Exchange, but individuals who create user accounts on www.nexusipe.org can also contribute to and comment on materials, share ideas, or answer questions in an online forum.

JOIN THE GROWING IPECP COMMUNITY

The Resource Exchange is just one way to participate in the National Center's growing online community. Create an account at www.nexusipe.org to start conversations with more than 1,000 individuals interested in IPECP from around the world.

National Center for Interprofessional Practice and Education

nexusipe.org

The National Center for Interprofessional Practice and Education is supported by a Health Resources and Services Administration Cooperative Agreement Award No. UE5HP25067. The National Center is also funded in part by the Josiah Macy Jr. Foundation, the Robert Wood Johnson Foundation, the Gordon and Betty Moore Foundation and the University of Minnesota.