What do we measure? Methodological Versus Institutional Validity in Student Surveys

Includes PowerPoint Presentation

Jeffrey Alan Johnson, Ph.D.
Senior Research Analyst, Institutional Research & Information
Lecturer, Department of History & Political Science

Paper presented at the Association for Institutional Research Annual Forum
Toronto, Ontario, May 24, 2011

Abstract : This paper examines the tension in the process of designing student surveys between the methodological requirements of good survey design and the institutional needs for survey data. Building on the commonly used argumentative approach to construct validity, I build an interpretive argument for student opinion surveys that allows assessment of the inferences and assumptions underlying the claim that an opinion survey item is a valid measure of a construct. I then evaluate the content validity of measures of program satisfaction and academic growth in Utah Valley University's 2010 Graduated Alumni Survey and 2009 Graduating Student Survey, surveys designed to maximize conformity with existing and legacy institutional priorities and demands. Using analytical assessment and empirical tests including response change across surveys, inter-item correlation, and factor analysis I show that UVU's surveys—and by implication, all surveys developed with primarily institutional validity in mind—are subject to grave challenges to the construct validity. I conclude by suggesting that effective operationalization of institutional priorities can bring together construct and institutional validity.

jeffrey.johnson@uvu.edu
http://johnsonanalytical.com

800 West University Parkway
Orem, Utah 84058
(801) 863-8993

Table of Contents

What do we measure? Methodological Versus Institutional Validity in Student Surveys

With student surveys an integral part of assessing institutional effectiveness, the methodological validity of such surveys is an exceptionally pressing issue in institutional research. But in institutional-specific student survey research, methodological validity can be overridden by the need for survey instruments to conform to established institutional priorities, that is to say, by "institutional validity." An exceptionally problematic form of face validity (the extent to which an operationalization appears intuitively connected to the concept operationalized), a survey item can be said to be institutionally valid to the extent that it is recognizable to stakeholders other than survey respondents as a representation of those stakeholders' objectives or priorities. The result is often a survey that may or may not have methodological validity either as individual items or as a coherent set of measures.

This paper examines the tension in the process of designing student surveys between the methodological requirements of good survey design and the institutional needs for survey data. Building on the commonly used argumentative approach to construct validity, I build an interpretive argument for student opinion surveys that allows assessment of the inferences and assumptions underlying the claim that an opinion survey item is a valid measure of a construct. I then evaluate the content validity of measures of program satisfaction and academic growth in Utah Valley University's 2010

Graduated Alumni Survey and 2009 Graduating Student Survey, surveys designed to maximize conformity with existing and legacy institutional priorities and demands. Using analytical assessment and empirical tests including response change across surveys, inter-item correlation, and factor analysis I show that UVU's surveys—and by implication, all surveys developed with primarily institutional validity in mind—are subject to grave challenges to the construct validity. I conclude by suggesting that effective operationalization of institutional priorities can bring together construct and institutional validity.

1 CONSTRUCT AND INSTITUTIONAL VALIDITY

1.1 Construct Validity in the Social Sciences

Along with reliability, validity is the core standard for evaluating the adequacy of variable measurement. But while reliability is generally well defined and can be tested by established statistical techniques in both education (Haertel, 2006) and the social sciences (Pollack, 2008, pp. 17-18), understandings of validity often recall United States Supreme Court Justice Potter Stewart's famed definition of pornography: "I know it when I see it." (Jacobellis v. Ohio, 1964) Mertens (2010, p. 383) initially defines validity as "the extent to which [an instrument] measures what it was intended to measure," a definition offering little help in research design but that is nonetheless considered the standard definition. (Scriven, 1991)

This standard definition is used in spite of the longstanding existence of much more effective definitions, such as that offered by Selltiz and colleagues:

> The validity of a measuring instrument may be defined as the extent to which differences in scores on it reflect true differences among the individuals, groups, or situations in the characteristic which it seeks to measure, or true differences in the same individual, group, or situation from one occasion to another.[1] (Selltiz, Jahoda, Deutsch, & Cook, 1959, p. 155)

Similarly, Mertens continues the standard definition by clarifying that "[i]n practice, however, the validity of an instrument is assessed in relation to the extent to which evidence can be generated that supports the claim that the instrument measures attributes targeted in the proposed research." (p. 383) This relationship between measure and attribute I take to be the essence of measurement validity.

Because the validity of a measure is based on a connection between a measure and a non-observable attribute (i.e., some construct that does not come already quantified by its very nature), validity cannot be assessed through simple statistical tests as reliability can. Scholars across fields agree that validity is to be evaluated holistically, as "a unitary concept that measures the degree to which all the accumulated evidence supports the intended interpretation" (Mertens, 2010, p. 383) or "by the extent to which its results are compatible with other relevant evidence . . . [which] depends on the nature and purpose of the measuring instrument." (Selltiz, Jahoda, Deutsch, & Cook, 1959, p. 156) This leads Kane, in his successor to the seminal work by Messick (1989), to argue that:

[1] Such a definition can be further simplified as "the extent to which differences in scores on [a measuring instrument] reflect true differences among units of analysis."

> The term “validation” and to a lesser extent the term ‘validity’ tend to have two distinct but closely related usages in discussions of measurement. In the first usage, ‘validation’ involves the development of evidence to support the proposed interpretations and uses . . . . In the second usage, “validation” is associated with an evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate. (2006, p. 17)

In essence, this is a demand that the validation of measures be both empirical and analytic.

It is now generally agreed that the heart of validity is construct validity, the ability to infer from a measurement to the existence or condition of a hypothesized construct. Measurement validity is thus “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.” (Kane, 2006, pp. 20-21) Validity is not the validation of measurements themselves but of the interpretations of such measurements, which requires the development of an argument that interprets such measurements (the interpretive argument) and an evaluation of both the analytical soundness and the empirical adequacy of the interpretive argument (the validity argument). The former, an example of which is shown in Table 3, must “specif[y] the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances.” (Kane, 2006, p. 23)

As such, an important consideration of construct validity is the theory underlying the relationship between the construct and the observation, (Selltiz, Jahoda, Deutsch, & Cook, 1959, pp. 159-161) as this more than anything allows one to argue for the inference from the observation to the construct. Kane (2006) holds that the evaluation of construct validity is primarily hypothetico-deductive, though the argument from measurement to construct is more properly considered an abductive inference (Johnson, 2000) based on the argument that the best explanation for the pattern of observations is the proposition that that the pattern will occur given the conditions of the construct. A sound theoretical argument in support of that proposition combined with empirical demonstration of the observations allows the inference of the construct conditions, just as a sound theory and observation of the predicted conditions allows the inference of the truth of a hypothesis.

1.2 “Institutional” Validity

As can be seen in the extended discussion of the construct validity of surveys below, construct validity is an exceptionally robust test of research effectiveness. Experience suggests, however, that the development of many measures in institutional research is not guided by such robust considerations. Indeed, outside of major national surveys such as the National Survey of Student Engagement (NSSE) it is likely that the development of many—perhaps most—survey items in institutional research is guided by what I shall term “institutional” validity: the extent to which a survey item serves as an artifact of institutional policy rather than an operational measurements of a construct.

Institutionally validated measures are supported by two dubious forms of validity. When survey items directly reflect the language of institutional policy statements such as mission statements, institutional values, or desired learning outcomes the reflect face validity. Face validity is the extent to which “the relevance of the measuring instrument to what one is trying to measure is apparent ‘on the face of it.’” (Selltiz, Jahoda, Deutsch, & Cook, 1959, p. 165) For example, a mission statement might include an institutional commitment to “promoting the virtues of multiculturalism and diversity.” A survey question that asks the respondent to “Rate your growth in multiculturalism and diversity” appears, on its face, to be a valid measure of success in that institutional commitment simply because it mirrors the language of the mission statement.

When survey measures are designed to be recognizable to stakeholders within or beyond the institution as recognition of those stakeholders’ priorities, the measures are supported by the equally dubious social validity:

> a criterion used to judge the quality of the research from the perspective of its social importance, the social significance of its goals, and the appropriateness of the procedures . . . . [T]hat is, it has to be viewed by those implementing it as applicable, feasible, and useful. (Mertens, 2010, p. 212)

Key to social validity is that it is acceptable to the stakeholders “as acceptable, feasible, and useful” because it is a direct expression of their goals or priorities. Institutionally valid questions are most likely to reflect social validity when they are the result of broad-based development processes with extensive participation by those without backgrounds in social scientific research. For example, a survey development process might include representation from a campus center encouraging classroom engagement. A question asking students to “Rate your satisfaction with engagement in the classroom” would be included as a (presumably valid) measure of classroom engagement because doing so satisfies the center's leadership that the institution considers the center important.

In both cases, the most important difference between construct and institutional validity is that the latter lacks a process of operationalization. The institutional policies and stakeholder are not translated into concrete expressions of the underlying constructs; rather the survey respondent is confronted with a construct label directly and asked to express some evaluation of the construct label. At best this requires accepting uncritically a range of assumptions that would ordinarily be evaluated formally in the process of assessing the construct validity of a measure. At worst, measures that are not operationalized reflect a total lack of awareness of those assumptions. The example of the multiculturalism rating question above assumes, among other things, that the survey item is understood by the respondents as a measure of the institutional goal and that other considerations do not influence the response. In an environment in which "multiculturalism" and "diversity" are objects of political contention neither assumption holds. Respondents, on the one hand, may not see student activities that allow groups to express their cultures as promotion of multiculturalism while at the same time responding based on their positions on the political debate rather than the actual practices of the institution. The same would be true of the classroom engagement question.

The results of this process are survey questions that are of unknown construct validity or reliability. In the absence of operational measurements of underlying constructs it is impossible to determine whether respondents are answering questions consistently given their experiences, whether the ideas that inform their answers reflect the same understanding of the construct held by the researchers and end users, or whether factors other than the condition of that construct are shaping the respondents' answers. This undermines the ability to draw sound conclusions from survey responses:

> There is little reason to assume that intuitive judgments about complex attitudes are inherently valid, even when the judgments are made by highly trained and intelligent people. If the concept of what is being measured is vague, as it is in some rating scales, it is unlikely that the ratings will be clear in meaning. When the concept of what is being measured is ambiguous, the ordering of individuals may actually be quite arbitrary, and even distinctions of greater or less become meaningless. (Selltiz, Jahoda, Deutsch, & Cook, 1959, p. 356)

The claims that survey respondents are, by and large, "highly trained and intelligent people" when it comes to parsing the language of surveys and that they offer highly trained and intelligent responses to on-the-spot survey questions is dubious to say the least. While nothing inherently prevents institutionally validated items from possessing construct validity as well, to rely solely on the former in developing an institutional survey is relying on luck to deliver the latter.

1.3 Examples of Institutionally Valid Measures: Utah Valley University's Graduating Student and Graduated Alumni Surveys

Utah Valley University is a large, public Baccalaureate/Associate's university (though it has recently added several graduate programs) located in Orem, Utah. UVU's Institutional Research & Information (IRI) office conducts two surveys of completing students annually. Both surveys consist of similar questions concerning the respondents' perceptions of UVU's educational quality, the respondents' employment at or since graduation, and their plans for further education. Educational quality data includes three general quality questions, a set of items rating personal and intellectual growth in specific areas, and a set of items rating satisfaction with specific aspects of academic programs. All are evaluated as individual items rather than as elements of a combined scale. These questions are generally identical between surveys, though methodological changes have been made occasionally, especially with regard to the response scales for the questions. The data is analyzed both by individual years and in standardized longitudinal datasets that adjust for annual changes in survey methodology to the extent possible. Data from comparable questions on corresponding surveys of an individual class is maintained in annual datasets for both surveys, though the number of overlapping respondents in a graduating class is typically about one fourth of the total respondents for either survey.

The Graduating Student Survey is conducted online. The graduation survey contacts all students who have submitted an application to graduate during a term. Students receive an email invitation to participate, and data is linked to individual students through a unique URL for each student. The most recent completed survey, covering the 2009-2010 graduating class, invited all 3,610 graduates to participate, with 756 responding to the survey. Longitudinal data is available since the 2005-2005 class. IRI also conducts an annual Graduated Alumni Survey (an awkward title that is itself an example of institutional validity). This telephone survey contacts all graduates in the second summer following their graduation (i.e., all graduates from the summer 2008, fall 2008, or spring 2009 terms were surveyed in September 2010). The most recent survey

Using the scale very satisfied, satisfied, neutral, unsatisfied, or very unsatisfied, how would you rate each of the following aspects of your academic program?

Overall education and training experience
Quality of instruction
Course content
Class size
Engagement in the classroom (Added in 2009-10)
Engagement in the community (Added in 2009-10)
Accessibility of instructors
Faculty interest and caring for students
Professional and vocational advising
Academic advising (Added in 2009-10)

Table 1: Program satisfaction items from UVU Graduating Student and Graduated Alumni Surveys

interviewed graduates from the 2008-09 academic year in 2010. IRI attempted to contact all graduates, including those who had continued to a further degree at UVU; 1,353 responded. Longitudinal data is available since the 2003-2004 graduating class.

The educational quality questions used in the two surveys are supported primarily by institutional validity, primarily in the social validity form. The program satisfaction questions are shown in Table 1. These items were first used in an *ad hoc* alumni survey in 2001. At the time, the items were developed in consultation with faculty and administrators to reflect the perceived priorities for service delivery at the academic program level. They have since been adjusted to reflect additional institutional stakeholders and changing institutional priorities, most notably through the inclusion of an academic advising option following the institutional separation of career and academic advising and the classroom and community engagement items following the designation of engagement as a core theme and objective over the course of the 2008-09 and 2009-10 academic years. A similar though more formal process was used to develop the list of personal and intellectual growth items shown in Table 2 beginning with the 2005 graduate and alumni surveys.

How much did your education at UVU contribute to your growth in the following areas?

Knowledge in major field of study
Critical thinking and problem solving skills
Communication skills
Mathematical skills
Health and wellness knowledge
Understanding and use of computers and technology
Interpersonal skills
Ethics
Preparation for real-world problems
Job seeking skills
Leadership and team management
Art and cultural knowledge
Community involvement and citizenship
Global perspective
Understanding diversity, different races and cultures

Response Scales

2010: Great contribution, moderate contribution, minor contribution, no contribution
2009: Major contribution, minor contribution, no contribution
2005-2008: Very great, great, average, little, none

Table 2: Personal and intellectual growth items from UVU Graduating Student and Graduated Alumni Surveys

Both of these items also display one of the expected hallmarks of questions developed through the social validity form of institutional validation. The overriding goal of the social validity process is inclusion: having an item included constitutes institutional recognition of stakeholders' priorities. The major constraint on survey length, the respondents' willingness to continue answering questions, is generally not a concern of the stakeholders unless it presents a major problem in collecting meaningful data. Even then, experience suggests that many stakeholders without technical backgrounds in survey research and a commitment to evidence-based decision making use data to justify rather than evaluate performance. The unreliability of data is thus not likely to be a concern as unreliable data that portrays the stakeholder in a positive light will be used anyway that the unreliability of data that portrays the stakeholder in a negative light will justify disregarding the data. The result is that socially-driven processes of institutional validation will lead to lengthy “laundry lists” of items added without regard to their methodological value.

The items concerning satisfaction with engagement in the classroom and the community are especially good examples of institutional validity in its face validity form as well. In a process finalized in June 2010, UVU adopted four core themes and objectives: Student success, serious, inclusive, and engaged. The last theme is defined by the statement, “UVU engages its communities in mutually beneficial collaboration and emphasizes engaged learning.” (Utah Valley University, 2010) In the spring of 2009, two corresponding items were added to the question, asking students to evaluate “Opportunities to engage in the community” and “Your engagement in the educational process.” In the 2010 alumni survey these were shortened to their current form. Nowhere in “engagement” operationalized; the reflection of the institutional policy in the language of the survey item is taken as sufficient warrant for its validity without regard to how respondents are likely to interpret—or misinterpret—engagement.

As items developed through a process of institutional rather than methodological validation, the educational quality measures in UVU’s Graduated Student and Graduating Alumni Surveys are inherently suspect (though, to be sure, no more so than other institutions to the extent that institutional validation is the norm in institutional research). But since institutionally valid measures can also have construct validity—indeed, the ideal may well be that measures would be both institutionally and methodologically valid, as discussed below—formal evaluation of their construct validity is necessary.

2 THE CONSTRUCT VALIDITY OF OPINION SURVEYS

2.1 A Theory of Opinion Survey Response

Evaluating the construct validity of a survey measure requires both a theory of survey response and an interpretive argument informed by that theory. Scholars in public opinion research have developed a robust theory of survey response that lends itself well to student opinion surveys in institutional research. In the past quarter century survey researchers have confronted an increasing number of identified error sources in survey responses both random (such as the instability of responses over even short periods of time) and systematic (such as response and question wording effects). As a consequence models of survey response that understand the response as a statement of an underlying "true" opinion or attitude have been largely abandoned in favor of models that see the survey response as an *ad hoc* answer to a question posed by the interviewer. (Zaller, 1992) The idea that respondents have stable, well thought-out attitudes on most matters has given way to approaches that see responses as improvised answers based on mentally averaging across large bodies of not always consistent information. Most notable among these approaches is Zaller's (1992) *Receive-Accept-Sample* (RAS) model of public opinion. Though the model was developed specifically in the context of public opinion about political issues, it can be generalized to opinion surveys generally.

Zaller begins by arguing that expressions of opinion are not statements about a fundamentally stable state of mind but rather functions of the considerations at hand when a question is asked. Considerations are "any reason[s] that might induce an individual to decide a political issue one way or another" and reflect both cognition and affect. (Zaller, 1992, pp. 40-44) In Zaller's theory of political opinion, considerations come primarily from elite discourse (especially as presented in the media), but there is no reason to limit the idea of considerations to reasons shaping opinion on political issues or to those factors that come from elite discourse.[2] Broadening the universe of opinions studied broadens both the range of considerations and of sources but does not fundamentally change the conditions from which opinions should form. With regard to student opinions on the kinds of issues often addressed in institutional research, the students' own experiences should be a major source of considerations. But other sources can play into this as well, such as interpersonal communication or media coverage of the students' own university and that of others to which students might compare their institution.

Respondents' stores of considerations, however, are not infinite, perfect records of every consideration every encountered. Two filters operate to select the considerations on which a respondent can draw when asked a question from the general body of information about the issue. The first is reception. Only a relatively small amount of information from the environment will be noticed, or "received," by an individual. This reception of information is determined by a person's level of cognitive engagement with an issue. A second filter is resistance. Noticing information does not in itself make the respondent more likely to use it as the basis for opinions. Information that is inconsistent with a person's predispositions is likely to be rejected from the process of forming

[2] Indeed, Zaller explicitly states that his claims about the sources of information are auxiliary assumptions that are not formally part of the model that he develops.

opinion. Zaller's predispositions (presumably reflecting the types of opinions in which he is interested) are primarily broad political attitudes, but there is no reason that the predispositions that allow resistance should be inherently limited to attitudes; predispositions about sources and types of arguments or evidence seem equally reasonable in the absence of empirical evidence to the contrary. However, resistance is possible only to the extent that respondents are capable of identifying information as inconsistent with their predispositions based on the respondents' awareness of appropriate contextual information connecting the information and the predisposition. Information that has been received and accepted becomes a consideration on which respondents can draw in answering a question. (Zaller, 1992, pp. 40-48)

Respondents answer questions, Zaller argues, by averaging across considerations, a process usually undertaken only at the moment a respondent answers a question. The considerations used, however, are only a sample of all of in the respondent's store of considerations. "Persons who have been asked a survey question do not normally canvass their minds for all considerations relevant to the issue; rather, they answer the question on the basis of whatever considerations are accessible 'at the top of the head.'" (Zaller, 1992, p. 49) The availability of considerations, which Zaller terms the "salience" of the condition is thus the key factor in the respondent's answer. Asking a survey question initiates a process in which respondents build a sample from among the store of considerations cued by the question consisting of those that are most salient. Most likely to be salient in answering a question are those considerations that have been recently used in some other context; often this will be a single dominant consideration. The answer given by a respondent then reflects a rough average of those considerations brought to the top of the head in the sampling process.

Survey items thus cannot be seen as reflections of a stable, "true" attitude or opinion on the part of the respondent. The RAS model implies, first, that very different sets of considerations may be available to a respondent answering the same question in the same survey at different moments, shaped by a plethora of external influences that may have brought different sets of considerations to the top of the head. Moreover, it implies that the survey itself can shape those considerations by cueing certain considerations over others, for example through question wording and the order of previous questions. To the extent that these sets of considerations are not consistent with each other, the varying sets of salient considerations across which respondents average will result in different responses given the same supposedly "true" attitude (i.e., the same total store of considerations).

Though not widely used in institutional research, the RAS model is as informative in the interpretation of student surveys as it is in public opinion surveys. Given a model of survey response it is possible to make a conceptual argument that would show why a given construct should manifest itself in a particular survey response: the question cues a set of considerations that, when respondents average across them, are indicative of the respondents' underlying stores of considerations relevant to the question and the likelihood of salience for each consideration in that store. Such a theory is an integral part of the interpretive argument for the construct validity of any opinion survey measure.

I1: **Scoring:** from observed performance to an observed score

A1.1: The scoring rule is appropriate.
A1.2: The scoring rule is applied as specified.
A1.3: The scoring is free of bias.
A1.4: The data fit any scaling model employed in scoring.

I2: **Generalization:** from observed score to universe score

A2.1: The sample of observations is representative of the universe of generalization.
A2.2: The sample of observations is large enough to control random error.

I3: **Extrapolation:** from universe score to target score

A3.1: The universe score is related to the target score.
A3.2: There are no systematic errors that are likely to undermine the extrapolation.

I4: **Implication:** from target score to verbal description

A4.1: The implications associated with a trait are appropriate.
A4.2: The properties of the observed scores support the implications associated with the trait label.

Table 3: "Interpretive Argument for a Trait Interpretation" (Kane, 2006, p. 34 Table 2.2) showing inferences (I) and assumptions (A).

2.2 The Scoring Inference

As yet there is no generally recognized model interpretive argument for student opinion surveys. Kane (2006), however, presents interpretive arguments for individual tests, traits, and theoretical constructs. Kane's examples serve here as models on which I build an interpretive argument for survey items measuring respondents' opinions. A preliminary[3] statement of the interpretive argument for surveys of student opinions is similar to Kane's argument for trait interpretations (see Table 3), using identical inferences but, in the cases of the extrapolation and implication inferences, supporting them with different assumptions.

[3] Like Kane's other examples, interpretive argument stated here is an *ad hoc* argument that applies solely to measurement of student opinions through surveys. My ongoing research into this question (which is far from complete) suggests the possibility that all three of Kane's argument and the one that I present here can be put into a general form in which the inferences are identical for all combinations of measurement and construct; only the assumptions supporting the interpretive arguments vary. Should this line of inquiry prove successful, it is likely that the argument here will change somewhat.

The initial inference is the scoring inference, which infers from the observation to the score assigned to it. It is satisfied if the score assigned to a survey response allows adequate categorization and differentiation of responses for the intended interpretive purposes. The assumptions supporting the scoring inference are identical in both opinion surveys and trait implication, though with slight differences in application. The scoring inference assumes that the scoring rule is appropriate to the item being scored, that the rule is applied as specified in the research protocols, and that the scoring rule lacks biases. The great virtue of survey research is that it simplifies meeting these assumptions. In closed-ended questions it is almost always the case that these three premises are sound, as the observation and the act of scoring are identical: the selection of a response from a pre-defined set of allowed responses. Matters beyond the relation of observation to score such as the relation between sample and population, the adequacy of the response options in capturing the full range of variation in the construct, or the attribution of the observed behavior to the construct are not within the scope of the scoring inference (though they are within the scope of subsequent inferences). This becomes far more problematic in the case of open-response items, where reliability and accuracy of the coding process must be demonstrated. (Haertel, 2006)

More generally problematic is the assumption that the scoring process uses a scaling method that is appropriate to the data used. In survey research this assumption is violated routinely through the construction of multi-item scales with interval or even ratio values from individual survey items measured ordinally. This problem is notable, for example, on NSSE. NSSE uses individual questions with a four-point ordinal response scale (such as “very often,” “often,” “sometimes,” and “never”) to create benchmark scales measuring various aspects of engagement. These benchmarks are then reported as the mean scaled values of the contributing items. (National Survey of Student Engagement, 2011) The problem here is that means cannot be calculated unless the difference between values is constant and there is a true zero value, a condition satisfied only by ratio variables. The numerical representations of ordinal NSSE items in the benchmark process (whether raw or converted to the 100-point scale used in the benchmark) do not correspond to points on a continuous number line. If the difference between “often” and “sometimes” is not one-fourth of the difference between “very often” and “never” then one cannot say that the mean responses are equal for the two pairs of items. In some cases one might be able to say that the responses are approximately equally distributed if there is empirical evidence to support this or the scale explicitly reflects this (e.g., options for individual items use quantified labels such as “100% of the time,” “66% of the time,” etc.). But scales that utilize median response value or mean response percentile rather than mean response value are much more likely to satisfy this assumption. This is, of course, not a problem where items are used as individual measures, though that can become a problem for the extrapolation inference.

2.3 The Generalization Inference

The generalization inference infers that the observed score is true of the “universe of generalization,” which is the set of all observed and unobserved instances of the type of behavior observed. It is satisfied if the measurement of actually observed responses is representative of a (hypothetical) measurement of the universe of behavior that could be observed given a survey question. Kane offers two assumptions that support the inference: that the sample of observations is representative, and that is it large enough to control for random error. The latter is supported rather straightforwardly by statistical sampling theory. The RAS model of survey response, however, complicates satisfaction of the former assumption somewhat. Given the traditional "true attitude" model of survey response it is sufficient to ask a question once in order to get a sufficiently representative sample of a respondent's attitudes. While question wording or order could change the observed response, it would also change the universe of generalization, maintaining the ability to generalize from observed score to universe score. The only threats to the representativeness of the sample of observations compared to the universe of observations (as opposed to the target score, which is addressed below) are those from inconsistent survey procedures—for example, changes in the interviewer that could result in changes in individual response.

But there is much greater variation in response given the RAS model, in which respondents themselves bring different considerations to the response at different times in spite of uniform survey protocols. As such, the assumption needs to be more precisely specified by defining the universe of generalization for survey responses:

> A2.1: The sample of observations is representative of the set of considerations cued by the question weighted by the frequency with which considerations are at the top of respondents' heads among the entire population of respondents.

The generalization inference thus requires both generalization across respondents and generalization across instances of response for individual respondents. Fully satisfying this assumption would require contacting respondents on multiple occasions with the same questions. This is likely to be impractical, but it exposes a significant weakness in the validity surveys as measures of student opinion.

2.4 The Extrapolation Inference

The RAS model of survey response has the most substantial effect on the extrapolation inference. It is insufficient to simply specify that measurement of the universe of generalization (which was approximated by the observed measurement in the generalization inference) is "related to the target score" (the score indicating the value of the construct across its entire domain of manifestations). The nature of the relation is in general specified by the theory relating the value of construct to the observation that is measured, in this case the RAS model. The assumption requires the ability to explain or predict the relationship between the observation (and thus, assuming the satisfaction of the scoring and generalization inferences, the observed and universe scores) from the value of the construct itself (only hypothetically known in practice). If one can show that a given construct value either logically entails or is empirically associated with a given observation, one has satisfied this assumption. Hence the first assumption of the extrapolation inference can be restated as:

A3.1: The universe score can be predicted or explained based on the target score. This, in practice, assumes that observed score is a valid estimate of the universe score, which will be true if the scoring and generalization inferences are sound.

It is also necessary to more precisely specify the assumption that the sample is free of systematic error. In the RAS model, systematic error can be explained primarily as conditions that distort the sampling and averaging process. This can itself be reduced to distortions in the sampling process by the proposition that an apparent distortion in the averaging process is in fact the introduction of a non-substantive consideration (e.g., the race of the interviewer who asks a question about racism) that outweighs all substantive considerations. Thus a measure is free of systematic error when:

> A3.2: The considerations underlying the universe score are representative of those throughout the target domain.

This is a straightforward consequence of the RAS model. The question cues certain considerations at the time of the survey, which the respondent samples and averages across to give an answer that is used to estimate the universe score. If the considerations that were at the top of the respondents' heads at the time of the survey are representative of those salient across the entire construct domain taking into account the relative frequency of the salience of each consideration, then the sampling and averaging process should indicate the same construct value for the universe score and the target domain itself.

Evaluation of the more technically defined forms of validity can be useful here. Assessment of content validity, construct underrepresentation and construct irrelevant variance (Kane, 2006) supports this assumption by showing that the universe of generalization is coextensive with the target domain. Assessment of and convergent and discriminant validity (LaNasa, Cabrera, & Trangsrud, 2009) and criterion validity allows one satisfy this assumption based on the claim that interpretation of the construct using the measure in question will have similar results to using other measures of the construct.

Though an extrapolation inference supported by any one test from the latter set of approaches is non-circular only where the validity of the other measures or criteria is known, (Donovan, Kendall, Young, & Rosenbak, 2008) substantial consistency of results across multiple measures of unknown validity can support this assumption abductively (i.e., the assumption is likely to be true because it is the best explanation for the observed consistency across multiple tests).

This is not, however, the only source of systematic error that is derived from the RAS model. Self-report bias is acknowledged as a major source of systematic error in survey research. It is most often characterized as a desire by respondents “to respond in a way that makes them look as good as possible.” (Donaldson & Grant-Vallone, 2002, p. 247) In the “true attitude” model of opinion this is the major source of self-report bias, and in the RAS model this desire can be understood, as described above, as a consideration in itself. But the RAS model also introduces a second source of self-report bias: a predisposition against information that would reflect poorly on the respondent that biases acceptance of the information. In this case the problem is not that the sample of considerations is distorted in favor of the respondent but that the information never becomes a consideration in the first place. The response reflects, thus, not the respondents’ willing distortion of what they actually believe but a good faith statement of the respondents’ “true” beliefs about themselves (i.e., the average of all considerations weighted by salience), albeit one filtered through the rose-colored glasses of the acceptance stage of the RAS model.

This suggests an additional assumption, derivative of assumption A3.2 in Table 3, supporting the extrapolation inference:

A3.3: The acceptance of considerations relevant to the target domain is not biased by respondents' predispositions against information that portrays them negatively.

There are three ways in which this assumption can be satisfied. The most straightforward is when the target domain does not involve respondents' views of themselves. This would be common in, for instance, evaluations of an institution to which the respondent has no connection personally such as a survey of employers. A second path toward satisfying this assumption is showing that the construct is a self-assessment by the respondent. In this case, the acceptance of information as a consideration is itself part of the construct. This assumption is also satisfied if there is no such predisposition on the part of the respondents; one would expect, however, that the circumstances under which this is the case are exceptionally limited and would thus require substantial empirical evidence in support.

2.5 The Implication Inference

The final inference needed to validate a survey response as a measure of student opinion is the implication inference, which infers from the target score to the "verbal description" of the construct (Kane, 2006, p. 34). This is to say, essentially, that the observed score (which, assuming satisfaction of the scoring, generalization, and extrapolation inferences is taken as an estimate of the target score) is representative of the conceptual claims invoked by the construct. Kane's first assumption here, that the implications associated with the construct are appropriate, is exceptionally vague, likely reflecting the difficulty of stating in general form tests for the claims of specific traits. This is somewhat true of surveys as well, suggesting the merit of retaining the assumption with minor adjustments to its terms:

> A4.1: The characteristics associated with the construct are logically and empirically consistent with the target score.

This is simply the specification of “appropriate” in terms of consistency and in reference to the target score (which is in principle the basis of the implication inference), and parallels similar assumptions supporting the previous inferences.

But there are some commonalities to all opinion surveys that allow a more specific assumption as well. Opinion surveys produce statements of the respondents’ opinions on an issue, but are often taken as representing more-or-less objective statements of more-or-less fact. For example, the NSSE asks students to report how much their institutions emphasize various aspects of educational practice, such as “Spending significant amounts of time studying and on academic work” or “Encouraging contact among students from different economic, social, and racial or ethnic backgrounds.” (National Survey of Student Engagement, 2011, p. 3) Taken as a measure of students’ perceptions of the university’s priorities this is appropriate, but not as a measure of the university’s priorities themselves. The latter is not a matter of students’ opinions but of policies set by institutional leaders. To the extent that these NSSE items are used to identify institutional priorities, students are not the appropriate sources nor are opinions the appropriate measures. This is generalized as an additional assumption supporting the implication inference:

A4.2: The target domain consists of evaluative claims that the respondents are competent to make.

Note that this is specific to opinion surveys; a survey designed to gather non-evaluative data regarding the respondents will have a different interpretive argument generally and is thus not under consideration here.

The final assumption, that of fit between the properties of the measurement and the implications of the construct, remains as important to opinion surveys as it is to trait implications. One of the most important aspects of this is that the level of measurement for the observed score be consistent with the conceptualization of the construct character. A binary measurement is not useful in applying a construct for which the condition is understood as varying continuously, for example. One might also consider whether the expected stability of the measurement and construct over time are consistent; a construct with high expected stability might require a scale based on longitudinal observations rather than a single survey. Where longitudinal comparisons of the construct are anticipated, evidence that there have been no fundamental changes in the salience of considerations resulting from construct-irrelevant factors is helpful.

The entire interpretive argument for opinion surveys in general is shown in Table 4; to be sure, specific constructs and measures may present unique assumptions that would need to be made in support of one or more inferences. For the interpretive argument to be sound, each of the inferences must itself be sound. For an inference to be sound, all of the assumptions supporting it must be sound, as must all of the preceding inferences. Where the chain of reasoning allows one to conclude, on the basis of a sound

I1: **Scoring:** from observed performance to an observed score

A1.1: The scoring rule is appropriate.
A1.2: The scoring rule is applied as specified.
A1.3: The scoring is free of bias.
A1.4: The data fit any scaling model employed in scoring.

I2: **Generalization:** from observed score to universe score

A2.1: The sample of observations is representative of the set of considerations cued by the question weighted by the frequency with which considerations are at the top of respondents' heads among the entire population of respondents.
A2.2: The sample of observations is large enough to control random error.

I3: **Extrapolation:** from universe score to target score

A3.1: The universe score can be predicted or explained based on the target score.
A3.2: The considerations on which the universe score is based are representative of those throughout the target domain.
A3.3: The acceptance of considerations relevant to the target domain is not biased by respondents' predispositions against information that portrays them negatively.

I4: **Implication:** from target score to verbal description

A4.1: The characteristics associated with the construct are logically and empirically consistent with the target score.
A4.2: The target domain consists of evaluative claims that the respondents are competent to make.
A4.3: The properties of the observed scores support the implications associated with the verbal description.

Table 4: Interpretive argument for an opinion survey showing inferences (I) and assumptions (A).

interpretive argument, that the observed score assigned by a measurement process is representative of the implications of the construct, the measurement can considered valid.

3 EVALUATING THE VALIDITY OF THE UVU SURVEYS

3.1 Scoring and Generalization

This evaluation focuses primarily on the opinions of the 2008-09 graduating class as measured in the 2009 Graduating Student Survey and the 2010 Graduated Alumni Survey in order to maximizes the consistency of survey items. Evaluation of the scoring inference and, for the most part, generalization inferences is accomplished by the standard techniques used to measure statistical reliability and the adequacy of the sampling process. UVU's surveys are not particularly unique in these respects, thus for the present purposes there is thus no need to evaluate them here.

One concern does arise with regard to generalization. As described in section 2.3, generalization of survey responses requires that one be able to generalize across all sets of considerations that are likely to be at the top of a respondent's head at a given time, requiring multiple responses to the same questions at different times. Except for surveys specifically testing response variation this is exceptionally uncommon, and probably not feasible in institutional research. To the extent that one can assume that the considerations at the tops of respondents' heads are representative on aggregate the variation can be treated as random and controlled using similar processes. Under such circumstances assumption 2.1 would be satisfied. But so treating the variation of consideration salience within individual respondents is probably better understood as a necessary simplification of reality rather than a reasonable reflection of it.

Response instability across the two should be an effective indicator of whether one can generalize from an individual response to the universe score for a given respondent. If the considerations are consistent across response instances, it is likely that the considerations sampled in the first instance (the graduation survey) are either similar to or at least consistent with those sampled in the second instance (the alumni survey). Substantial variation between the two instances indicates divergent sets of considerations. This is, of course, not an ideal test as changes in response may reflect real change in opinion over time, but stability of responses should be a good indicator of consistent cueing.

	Major Change (> 1 category)	Minor Change (1 Category)	No Change	N	*p*
Program Satisfaction		**Percentage of respondents reporting change**			
Overall education and training experience	9.7	38.6	51.7	381	0.033
Quality of instruction	10.7	35.6	53.7	382	0.001
Course content	11.3	37.7	51	382	0.539
Class size	5.6	32.6	61.9	378	0.002
Accessibility of instructors	11.1	42.5	46.4	379	0.009
Faculty interest and caring for students	13.2	37.8	48.9	376	0.652
Professional and vocational advising	26.6	42.5	30.8	360	0.038
Personal and Intellectual Growth					
Knowledge in major field of study	11.5	28.1	60.4	384	0.000
Critical thinking and problem solving skills	18.9	33.4	47.6	380	0.000
Communication skills	18.3	34.1	47.7	384	0.000
Mathematical skills	27.6	25.3	47.1	380	0.000
Health and wellness knowledge	33	26.1	41.1	380	0.042
Understanding and use of computers & technology	33.4	24.9	41.6	382	0.002
Interpersonal skills	29.2	29.2	41.6	380	0.008
Ethics	34.4	26	39.6	384	0.968
Preparation for real-world problems	30.1	28.8	41.2	379	0.000
Leadership and team management	29.1	24.1	46.7	381	0.000
Art and cultural knowledge	36.6	25.8	37.6	380	0.061
Community involvement and citizenship	38.1	22.4	39.5	380	0.643
Global perspective	32.4	23.1	44.4	381	0.000
Understanding different races and cultures (diversity)	34.5	24.6	40.9	379	0.170
					0.000

Table 5: Response instability across Graduating Student and Graduated Alumni Surveys

Analysis of response instability suggests that poorly operationalized items are indeed subject to greater instability, as can be seen in Table 5. On the program satisfaction question, between 5.6% and 13.2% of respondents showed major change (greater than one response category) in all areas except professional and vocational advising (26.6%). One might expect a higher rate of substantive change in career advising as many graduates do not begin to search for work in earnest (and thus discover the true effectiveness of career advising) until after graduation. This suggests that the program satisfaction questions are generally cueing the same considerations consistently. The same cannot be said for personal and intellectual growth, however. Instability is higher generally for these items, with only one (knowledge in the major field of study) within the range of most program satisfaction questions. The more directly stated items show generally more stability than the more abstract ones. This suggests that personal and intellectual growth items either probe an area of substantial change over the first two years following graduation or do not provide sufficiently reliable cues to generate a representative set of considerations.

3.2 Extrapolation

Evaluation of institutionally validated survey items hinges primarily on the extrapolation inference and is largely a function of effective operationalization of the construct in the items. Whether the observed score (assuming it is an accurate estimate of the universe score) can be predicted or explained based on the target score depends in large part on whether the item is unambiguously operationalized; the same is true of whether the considerations cued by the question (as an estimate of those underlying the universe score) are representative of those throughout the target domain. A clearly operationalized item would be expected to cue a relatively specific and therefore predictable set of considerations, leading to relatively consistent and predictable averaging across sampled considerations. This would be true, for example, of the personal and intellectual growth item relating to knowledge in the major field, especially given recent moves to make the survey specific to a particular degree and major where students had multiple majors or received more than one degree. The item specifies knowledge in a domain that is reasonably well defined not just by disciplinary lines but by the students' experiences of degree requirements. As a result it is reasonable to conclude that the considerations coming to mind when responding to the question should reflect the scope of the discipline generally (the target domain) and that responses to the item would be predictable solely given (hypothetical) knowledge of the students' assessments of how well they know the content of that discipline.

That will not be true of poorly operationalized items. The more abstract the item is, the less reliably it will cue considerations. It becomes more likely that the considerations cued by the item will exclude some considerations within the target domain and include considerations that are external to it. The resulting construct underrepresentation and construct-irrelevant variance undermines the assumption of representativeness. At the same time, since the considerations cued are not necessarily consistent across respondents, they cannot be predicted based on the target score alone; prediction requires understanding how a respondent interprets the item as well. Both are likely true of the personal and intellectual growth item "Understanding diversity, different races and cultures" [*sic*]. This item may cue a wide range of considerations that are neither part of the target domain nor predictable based on the target score itself. The unclear grammar of the item prevents students from consistently interpreting whether diversity, different races, and cultures are distinct categories or whether "diversity" is a synonym for "different races and cultures," while "understanding" might be taken as either intellectual comprehension or as internalization of a value judgment. The possibility of the item being interpreted in the latter sense also raises the likelihood that the item will cue considerations related to the political controversies discussed in section 1.2. One would thus need to know not only the target score for the measure but also the respondents' interpretations of the item and the political views on the issue of multiculturalism to predict their observed scores. One might expect similar problems with the engagement items in the program satisfaction question.

Convergent and divergent validity can be effective in assessing the correspondence between universe score and target domain. The UVU surveys do not provide much opportunity to assess convergent validity, as there are no alternative measures of the constructs at issue. However, there is opportunity to assess divergent validity. In principle, each item is understood as measuring a distinct construct. The independence of these measurements is a function of the extent to which this is true, and can be tested by inter-item correlation. For both questions, inter-item correlation of responses to the 2010 alumni survey suggests modest independence of most items.

There are no program satisfaction items[4] (see Table 6) that strongly correlate with each other (Pearson's $r \geq 0.7$) but a substantial number of items that are moderately correlated with each other ($r \geq 0.4$). The professional and academic advising items are of most concern ($r = 0.608$), suggesting that organizational independence does not necessarily translate into methodological independence. Quality of instruction correlates modestly with most classroom practice (class size, course content, engagement in the classroom) and faculty-student interaction (accessibility of instructors and faculty interest in and caring for students) measures. Students also seem to take accessibility as an aspect of faculty interest in students ($r = 0.537$). Overall, 11 of 36 bivariate relationships are moderately correlated. Inter-item correlation is somewhat stronger among the

[4] The overall satisfaction item was omitted from this analysis because it would not be expected to be fully independent of the other items.

	Quality of instruction	Content	Class size	Class Engaged	Comm. Engaged	Access	Faculty interest	Prof. advising
Course content	0.547							
Class size	0.322	0.345						
Engagement in the classroom	0.446	0.412	0.464					
Engagement in the community	0.329	0.368	0.237	0.398				
Accessibility of instructors	0.408	0.311	0.339	0.438	0.333			
Faculty interest and caring for students	0.489	0.385	0.349	0.461	0.345	0.537		
Professional and vocational advising	0.335	0.339	0.264	0.327	0.368	0.307	0.412	
Academic advising	0.285	0.305	0.197	0.233	0.274	0.289	0.296	0.608

All correlations are significant at $p \leq 0.001$. Column headings are abbreviated; correspondence to row headings is self-explanatory

Table 6: Inter-item correlation of Graduated Alumni Survey program satisfaction items

personal and intellectual growth items (see Table 7). There are still no strong relationships but nearly half (49 of 105) of bivariate relationships are moderately correlated. The most notable relationships show obvious connections: interpersonal skills correlate with communication skills (r = 0.620) and ethics (r = 0.574), preparation for real-world problems with job seeking skills (r = 0.570), and global perspective with understanding diversity (r = 0.596).

A final problem related to the predictability of the universe score in the growth items is the interpretation of how much "UVU contribute[d] to your growth" in all items. The specificity of this question may tax the extent to which some survey respondents are willing to parse language precisely. As a result, some respondents may be rating their

	Knowledge in major	Critical thinking	Comm.	Math	Health	Tech.	Personal	Ethics
Critical thinking and problem solving skills	0.394							
Communication skills	0.29	0.523						
Mathematical skills	0.179	0.293	0.24					
Health and wellness knowledge	0.235	0.342	0.368	0.33				
Understanding and use of computers and technology	0.29	0.324	0.391	0.328	0.286			
Interpersonal skills	0.293	0.435	0.62	0.23	0.392	0.448		
Ethics	0.273	0.413	0.488	0.26	0.423	0.369	0.574	
Preparation for real-world problems	0.429	0.45	0.436	0.229	0.394	0.381	0.479	0.467
Job seeking skills	0.361	0.358	0.409	0.262	0.355	0.395	0.438	0.4
Leadership and team management	0.336	0.419	0.484	0.222	0.324	0.381	0.495	0.43
Art and cultural knowledge	0.216	0.332	0.326	0.23	0.364	0.334	0.382	0.411
Community involvement and citizenship	0.304	0.386	0.441	0.29	0.497	0.365	0.482	0.457
Global perspective	0.326	0.37	0.401	0.283	0.352	0.35	0.382	0.423
Understanding different races and cultures (diversity)	0.263	0.386	0.405	0.266	0.418	0.341	0.455	0.460

	Real-world	Job seeking	Leadership	Art	Community	Global
Job seeking skills	0.57					
Leadership and team management	0.497	0.508				
Art and cultural knowledge	0.338	0.33	0.322			
Community involvement and citizenship	0.501	0.474	0.485	0.49		
Global perspective	0.452	0.431	0.417	0.414	0.552	
Understanding different races and cultures (diversity)	0.44	0.415	0.429	0.468	0.512	0.596

All correlations are significant at $p \leq 0.001$. Column headings are abbreviated; correspondence to row headings is self-explanatory

Table 7: Inter-item correlation of Graduated Alumni Survey personal and intellectual growth items

overall skills in the area while others are comparing their skills before beginning and after completing their programs. Here a convergent validity test could be useful. Among the former set of respondents one would expect higher evaluations in areas most used in their programs, while the latter respondents would typically show highest growth in areas for which they would be least prepared. This test would be most effective if data was available at the program level regarding the highest required course in an area of growth and the first course typically taken by students in the program (based on placement rather than simply curriculum). Where growth is the dominant factor the distance between the two courses should correlate with the response; where absolute level of skill is the factor the highest course regardless of placement should show stronger correlation. Unfortunately, such data is not presently available at UVU. In the absence of such data, satisfaction of assumption 3.1 remains suspect.

Assumption 3.3 is also problematic. One can expect that acceptance of considerations relevant to one's personal and intellectual growth is partially biased by respondents' predispositions against information that portrays them negatively. Low growth could be taken as suggesting intellectual inferiority, and so information suggesting low growth would face more substantial challenges to acceptance. To the extent that such information is associated with a context blaming the university rather than the respondent, however, such information could still successfully negotiate the acceptance process. But one might also see another form of predisposition bias unique to the personal and intellectual growth questions: that in favor of information consistent with the respondents' overall perceptions of UVU. In this case, respondents are likely to

	Rotated Factor Matrix		
	All Items		Selected Items
	Factor 1	Factor 2	Factor 1
Engagement in the classroom	**0.698**	0.143	**0.652**
Faculty interest and caring for students	**0.644**	0.267	**0.713**
Quality of instruction	**0.642**	0.231	**0.723**
Accessibility of instructors	**0.596**	0.213	**0.631**
Course content	**0.566**	0.258	**0.61**
Class size	**0.529**	0.129	**Not included**
Engagement in the community	**0.445**	0.291	
Professional and vocational advising	0.284	**0.76**	
Academic advising	0.191	**0.725**	
Rotation Sums of Squared Loadings	**2.585**	**1.461**	**2.226**
% of Variance	**28.718**	**16.228**	**44.517**
Correlation with Overall Evaluation (p < 0.001)	**0.615**		**0.650**

Table 8: Factor analysis of Graduated Alumni Survey program satisfaction items (principal axis factoring with varimax rotation)

incorporate as considerations for individual items only that information that would lead them to evaluate the individual item consistently with their overall opinion.

Factor analyses of the two questions suggest that this is the case. Principle Axis Factoring with Varimax rotation across the nine specific aspects of program satisfaction (excluding the overall rating) shows two factors onto which the variables load (see Table 8). The primary factor loads onto the seven items not concerned with advising with moderate loadings between 0.445 for engagement in the community and 0.698 for engagement in the classroom. This factor loads very weakly onto the advising items. A secondary factor that loads weakly on other items but exceptionally strongly on the academic issues: loading is 0.725 for academic advising and 0.760 for professional and vocational advising. A more refined model aiming to better specify the primary factor excluded the advising issues and two variables with weaker correlations with other items

	Rotated Factor Matrix(a)	
	Factor 1	Factor 2
Communication skills	**0.666**	0.284
Interpersonal skills	**0.658**	0.341
Preparation for real-world problems	**0.597**	0.388
Leadership and team management	**0.594**	0.342
Critical thinking and problem solving skills	**0.571**	0.295
Job seeking skills	**0.536**	0.384
Ethics	**0.516**	0.435
Understanding and use of computers and technology	**0.467**	0.323
Knowledge in major field of study	**0.451**	0.209
Understanding different races and cultures (diversity)	0.315	**0.669**
Community involvement and citizenship	0.387	**0.651**
Global perspective	0.323	**0.637**
Art and cultural knowledge	0.252	**0.567**
Health and wellness knowledge	0.33	**0.49**
Mathematical skills	0.247	0.323
Rotation Sums of Squared Loadings	**3.471**	**2.979**
% of Variance	**23.143**	**19.857**
Correlation with Overall Evaluation (p < 0.001)	**0.418**	

Table 9: Factor analysis of Graduated Alumni Survey personal and intellectual growth items (principal axis factoring with varimax rotation)

(class size, which had least variation in general, and engagement in the community). This five-variable model produced a single factor with loadings between 0.610 and 0.723; the sum of squared loadings was a quite robust 44.5%. The most likely candidate for the primary factor is the overall rating, which correlated strongly with both the general model (Pearson's $r = 0.615$, $p \leq 0.001$) and the primary factor model ($r = 0.650$, $p \leq 0.001$).

This effect was weaker but still notable among the 15 growth items asked in 2010. Again, two factors emerged. A primary factor loaded strongly on items concerned with intellectual, professional, and interpersonal skills with loadings between 0.451 for knowledge in the major and 0.666 for communication skills. A second factor loaded onto items concerned with social, cultural, and health knowledge with loadings between 0.490

for health and wellness knowledge and 0.669 for understanding diversity. Math skills loaded weakly onto both factors (0.247 and 0.323, respectively). The sum of squared loadings was strong as before (43.0%). While there was no overall assessment item, correlation with a question about the overall evaluation of the respondents' university experiences was moderate (r = 0.418, $p \leq 0.001$), suggesting that overall evaluation bias may play a role in these items.

It is possible that the overall evaluation is caused by rather than the cause of the connection across variables. If each item was understood as a consideration, the overall evaluation would be expected to be the average evaluation of each item. This is not likely to be the case, however. Analytically, one would expect this to be more likely if the overall evaluations followed the specific ones. The overall university rating question was the first question in the survey, and the overall program satisfaction item was the first of the items. Additionally, one would expect very strong correlation between the overall satisfaction rating and the mean of the specific satisfaction scores. The correlation is robust (r = 0.627, $p \leq 0.001$) but not overwhelming, suggesting at the least that some of these items are not considerations in overall satisfaction and that other considerations are present. The overall satisfaction bias in the acceptance stage of the RAS model is thus the most likely explanation of the factor analysis results.

3.3 Implication

The implication inference can be difficult to evaluate when surveys are developed through a process built on institutional validity because institutionally valid items are likely to have ill-defined target scores. A construct tested directly (e.g., measuring satisfaction by asking, “Are you satisfied?”) will unquestionably show logical consistency between the target score (as always, assuming that the observed score is a sound estimate of the target score because the scoring, generalization, and extrapolation inferences have been satisfied) and the construct itself, but only in the sense that the equivalence of satisfaction with the quality of instruction as a construct and satisfaction with the quality of instruction as the entirety of the target score is a truism. Hence, insofar as each item is taken as a measure of a distinct construct (which is the intent of the survey design) as opposed to an element of the target domain of all items collectively, assumption 4.1 is satisfied by both questions, but only trivially so. This does little to support the implication inference. A more robust argument could be made if satisfaction had been operationalized into more specific claims, such as whether respondents would change how their programs address a particular issue or whether the quality of the practice met their expectations.

Assessing assumptions 4.2 and 4.3 are more straightforward. Students are obviously well placed to express an evaluative claim about their satisfaction with the program, and perhaps the only ones who can do so. But a student opinion survey may not be the ideal means to assess personal and intellectual growth. On the one hand, this construct is less of an evaluative claim than satisfaction is; certainly end users want to say that there is a real, if not necessarily directly or easily observable, state of the world about students’ growth. On the other hand, students may not be the most competent evaluators of that growth. Self-reporting bias may lead students to overestimate their growth generally, divergent salience of conditions across students will lead some to think more seriously about some areas than others, and even graduates may still lack the expertise needed to make the kinds of claims that the question anticipates. Indeed, more than one professor has been satisfied with graduates who know how little they know about their field.

The properties of the items unequivocally support the characteristics of the construct. This is an area, in fact, where the construct validity may actually exceed the needs of the research program. The opinions are expressed on a scale of (generally) five points. But in analysis, responses are often collapsed into categories reflecting favorable and unfavorable responses except where this would eliminate variation in the responses. The items could be reduced to a simple scale identifying favorable, neutral, and unfavorable attitudes without compromising the ability of the observed score to support the characterization of opinions. Previous studies did report mean values for each item, which is not statistically meaningful for the ordinal measurements used in these items, but this practice has been discontinued. Descriptive properties such as value labels are also a consideration here. Given the use of data as described above, however, the verbal descriptions of each point on the response scale orient the respondent to the scale rather than describe attitudes substantively.

4 UNITING CONSTRUCT AND INSTITUTIONAL VALIDITY

The institutional validation process has clearly limited the construct validity of UVU's Graduating Student and Graduated Alumni Surveys. The problems that emerge in evaluating the interpretive arguments for the program satisfaction and personal and intellectual growth questions are precisely those that one would expect to see consequent to an institutional validation of measures. While the scoring inference is as sound as it would be for any closed-ended survey question, none of the remaining three inferences can be accepted without substantial reservations across all of the items. The ability to generalize the observations is challenged by response instability. To be sure, this is a problem that plagues most surveys; it is simply impossible to identify an average response across a representative sample of considerations in a single response. While treating this as a necessary simplification of reality is a tolerable *modus vivendi* for researchers the limitations it places on survey research must be acknowledged far more often than they are. But the UVU surveys show that this problem can be exacerbated by poorly operationalized constructs that do not cue considerations reliably.

The extrapolation inference must be considered unsound with respect to both survey items. Too much interpretation of some items by respondents is needed before they can provide answers, introducing both construct underrepresentation and construct-irrelevant variation and clouding the relationship between the measurement and the target domain. There is also quite substantial dependence among items, suggesting that items do not effectively discriminate the construct in question from other constructs. It thus be said neither that a given condition of the construct will result in a specific observation nor that the observed score is representative of the target domain. Moreover, though this is not a problem directly associated with institutional validity, the extrapolation inference is undermined by the fact that respondents are biased in their acceptance of considerations, both toward those considerations that reflect poorly on themselves and toward those considerations that are not consistent with their overall views of the institution.

The interpolation inference is, for many of these items, quite weak. Student opinion surveys are probably not the best tool to evaluate educational outcomes such as personal and intellectual growth as such (though it might be reasonable to seek information about students' perceptions of their growth as a distinct construct), and the connections between constructs and their target domains offer no analytical utility because the two are undifferentiated. But regardless of its strength in isolation, the implication inference cannot be sound in the absence of sound generalization and extrapolation inferences, as these inferences are necessary to link the observed score to the target domain. In the absence of such inferences one might link construct to target domain but have no idea whether the target domain corresponds to the universe of generalization or the observed score. The clear lack of a sound extrapolation inference (as well as a questionable generalization inference) this undermines the implication inference in these surveys.

The result is a clear lack of demonstrable construct validity in the UVU student opinion surveys. The interpretive argument for opinion surveys can offer little support for the claim that the survey items measure the associated constructs with any reasonable accuracy, in large part because the quite serious problems to which survey items developed with a view toward institutional validity are prone run contrary to the assumptions that support the key inferences establishing the equivalence of measurement and construct. To the extent that the process used to develop these surveys at UVU is representative of that used to develop institutionally-based surveys generally—and experience suggests that this is the case—the construct validity of student opinion surveys must be seriously questioned.

This does not mean, however, that institutional validity is unimportant. The stakeholders and policy statements that drive institutional validity represent the key objectives of colleges and universities. Kane's warning about the lack of face validity applies especially strongly to institutional validity as well: "The appearance of relevance does not go far in supporting the appropriateness of a trait interpretation, but a serious lack of such relevance can lend credibility to certain challenges to the extrapolation inference." (2006, p. 36) Surveys that cannot "be viewed by those implementing it as applicable, feasible, and useful" (Mertens, 2010, p. 212) will prove themselves as practically irrelevant as they are methodologically sound. The ideal student opinion survey will manifest both construct and institutional validity in survey items that measure, rather than simply invoke, institutional priorities.

The foremost step in preventing institutional validity from undermining construct validity is operationalization of constructs. Institutional priorities are usually, for good reason, expressed as general principles that await translation into practices appropriate to a specific context. Institutional researchers seek responses

> not as a simple predictor of behavior but as a basis for inferring the degree to which the individual possesses some characteristic presumed to be reflected in the test performance. The presumed characteristic being reflected is not something

> which can be pointed to or identified with some specific kind of behavior; rather, it is an abstraction, a construct. (Selltiz, Jahoda, Deutsch, & Cook, 1959, p. 159)

As professionals with often more expertise in survey design than they typical university administrator, institutional researchers are in positions to find concrete measures that represent the institutional priorities without requiring interpretation or background information from respondents. Where institutional researchers effectively operationalize the constructs with which stakeholders are concerned, they can produce data that is both meaningful institutionally and robust methodologically.

5 WORKS CITED

Jacobellis v. Ohio, 378 U.S. 184 (United States Supreme Court 1964).

Donaldson, S. I., & Grant-Vallone, E. J. (2002, December). Understanding Self-Report Bias in Organizational Behavior Research. *Journal of Business and Psychology, 17*(2), 245-260.

Donovan, N. J., Kendall, D. L., Young, M. E., & Rosenbak, J. C. (2008). The Communicative Effectiveness Survey: Preliminary Evidence of Construct Validity. *American Journal of Speech-Language Pathology, 17*, 335-347.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth ed., pp. 65-110). Westport, Connecticut, United States: American Council on Education/Praeger.

Johnson, J. A. (2000). Abductive Inference and the Problem of Explanation in Social Science. *Midwestern Political Science Association Annual Meeting.* Chicago.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth ed., pp. 17-64). Westport, Connecticut, United States: American Council on Education/Praeger.

LaNasa, S. M., Cabrera, A. F., & Trangsrud, H. (2009). The Construct Validity of Student Engagement: A Confirmatory Factor Analysis Approach. *Research in Higher Education, 50*, 315-332.

Mertens, D. M. (2010). *Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods* (3rd ed.). Thousand Oaks, California: SAGE Publications.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Meausrement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

National Survey of Student Engagement. (2011). *Construction of the NSSE Benchmarks.* Retrieved May 16, 2011, from NSSE: National Survey of Student Engagement: http://nsse.iub.edu/_/?cid=403

National Survey of Student Engagement. (2011). *NSSE Survey Instrument: 2011 US English Paper Version.* Retrieved May 17, 2011, from National Survey of Student Engagement: http://nsse.iub.edu/pdf/survey_instruments/2011/NSSE2011_US_English_Paper.pdf

Pollack, P. H. (2008). *The Essentials of Political Analysis.* Washington, D.C., United States: CQ Press.

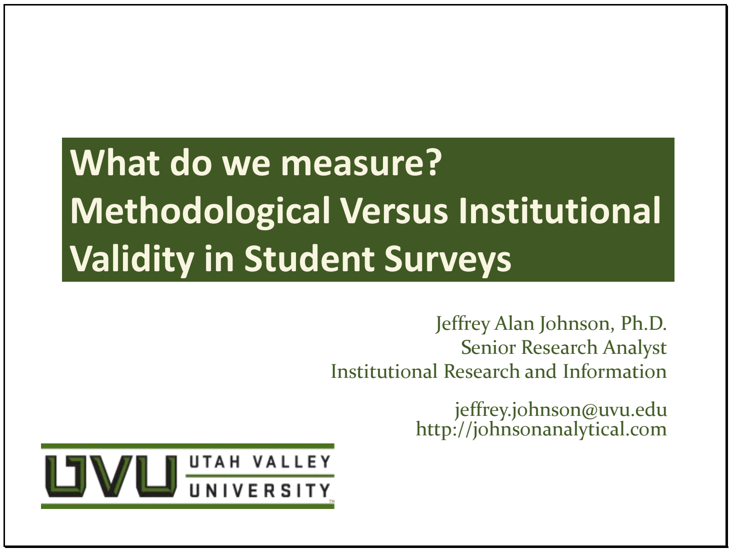
Scriven, M. (1991). *Evaluation Thesaurus* (Fourth ed.). Newbury Park, California, United States: SAGE.

Selltiz, C., Jahoda, M., Deutsch, M., & Cook, S. W. (1959). *Research Methods in Social Relations* (Revised On-volume ed.). New York: Holt, Rinehart, and Winston.

Utah Valley University. (2010, June). *Core Themes and Objectives--Endorsed.* Retrieved May 17, 2011, from Utah Valley University Strategic Planning: http://www.uvu.edu/planning/documents/Core%20themes%20and%20objectives%20-%20endorsed.pdf

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion.* New York: Cambridge University Press.

6 APPENDIX: AIR FORUM POWERPOINT PRESENTATION

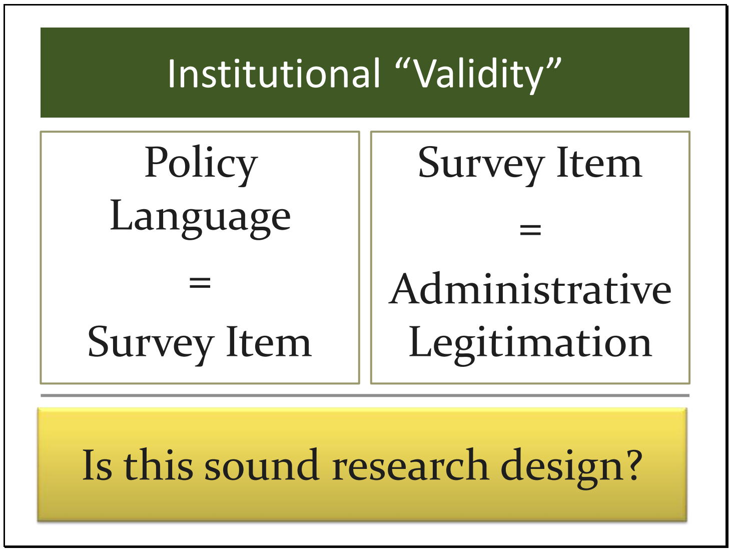


What do we measure? Methodological Versus Institutional Validity in Student Surveys

Jeffrey Alan Johnson, Ph.D.
Senior Research Analyst
Institutional Research and Information

jeffrey.johnson@uvu.edu
http://johnsonanalytical.com

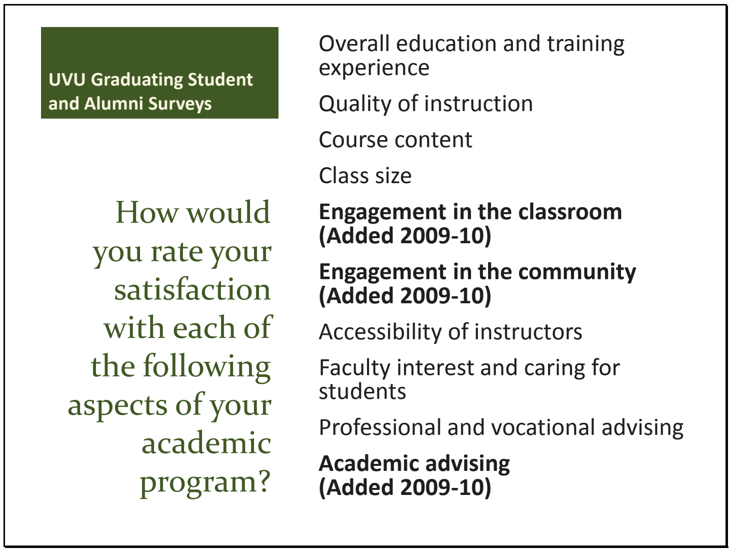
UVU UTAH VALLEY UNIVERSITY



Institutional "Validity"

Policy Language = Survey Item	Survey Item = Administrative Legitimation

Is this sound research design?



UVU Graduating Student and Alumni Surveys

How would you rate your satisfaction with each of the following aspects of your academic program?

Overall education and training experience

Quality of instruction

Course content

Class size

Engagement in the classroom (Added 2009-10)

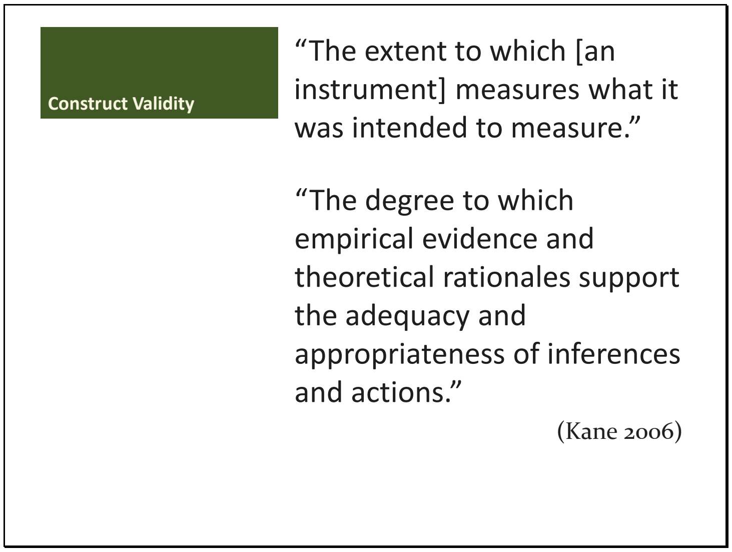
Engagement in the community (Added 2009-10)

Accessibility of instructors

Faculty interest and caring for students

Professional and vocational advising

Academic advising (Added 2009-10)

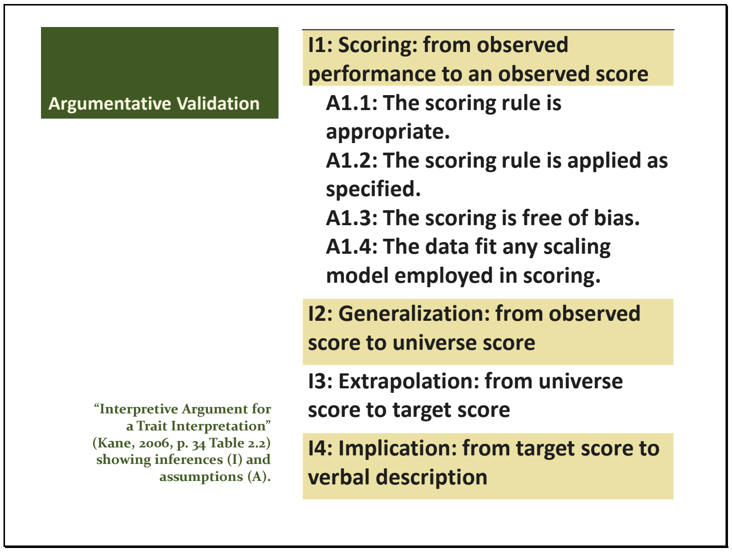


Construct Validity

"The extent to which [an instrument] measures what it was intended to measure."

"The degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions."

(Kane 2006)



Argumentative Validation

I1: Scoring: from observed performance to an observed score

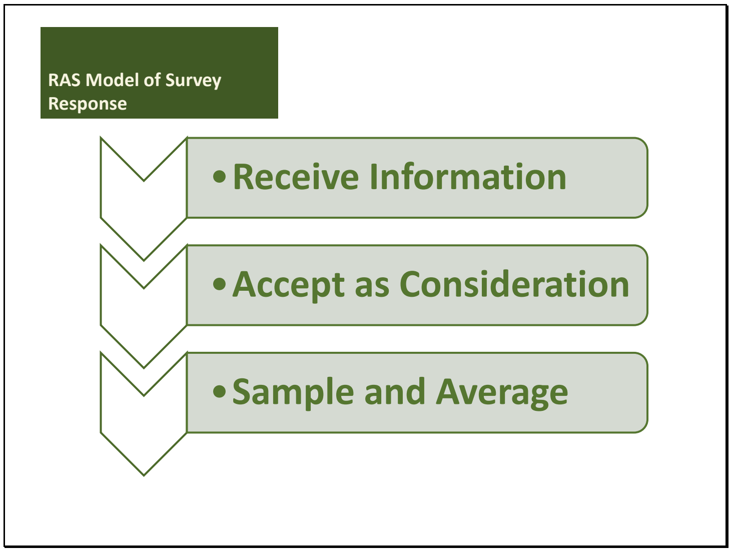
A1.1: The scoring rule is appropriate.
A1.2: The scoring rule is applied as specified.
A1.3: The scoring is free of bias.
A1.4: The data fit any scaling model employed in scoring.

I2: Generalization: from observed score to universe score

I3: Extrapolation: from universe score to target score

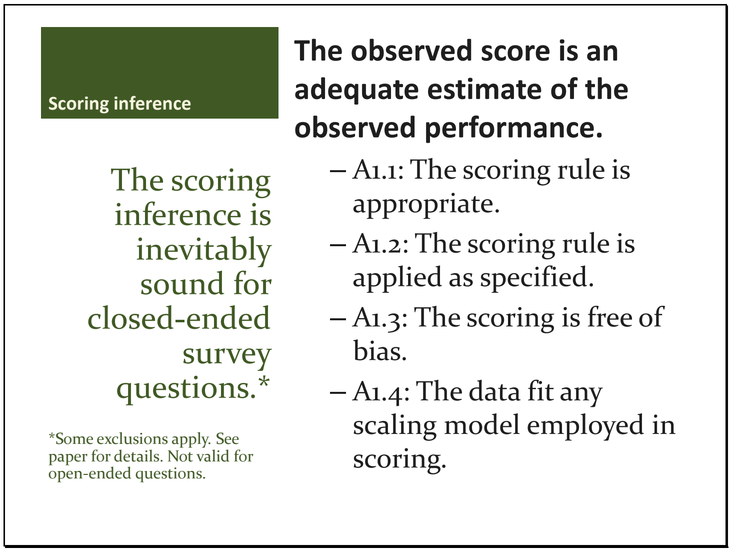
I4: Implication: from target score to verbal description

"Interpretive Argument for a Trait Interpretation" (Kane, 2006, p. 34 Table 2.2) showing inferences (I) and assumptions (A).



RAS Model of Survey Response

- Receive Information
- Accept as Consideration
- Sample and Average



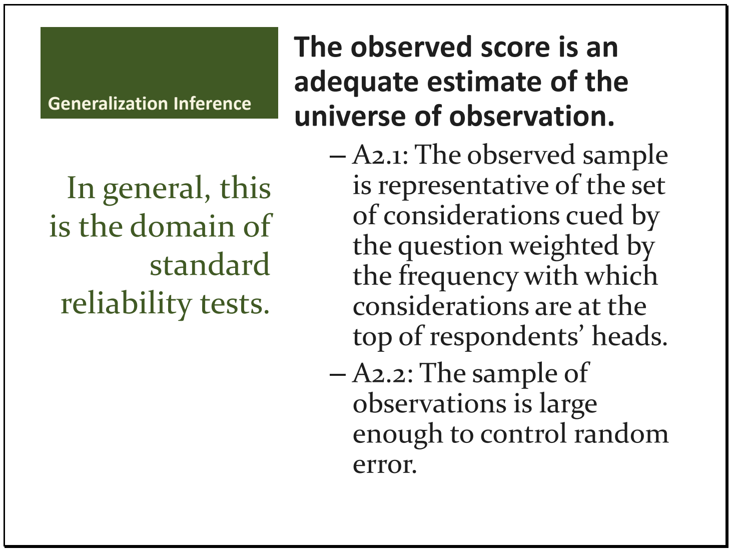
Scoring inference

The observed score is an adequate estimate of the observed performance.

– A1.1: The scoring rule is appropriate.
– A1.2: The scoring rule is applied as specified.
– A1.3: The scoring is free of bias.
– A1.4: The data fit any scaling model employed in scoring.

The scoring inference is inevitably sound for closed-ended survey questions.*

*Some exclusions apply. See paper for details. Not valid for open-ended questions.

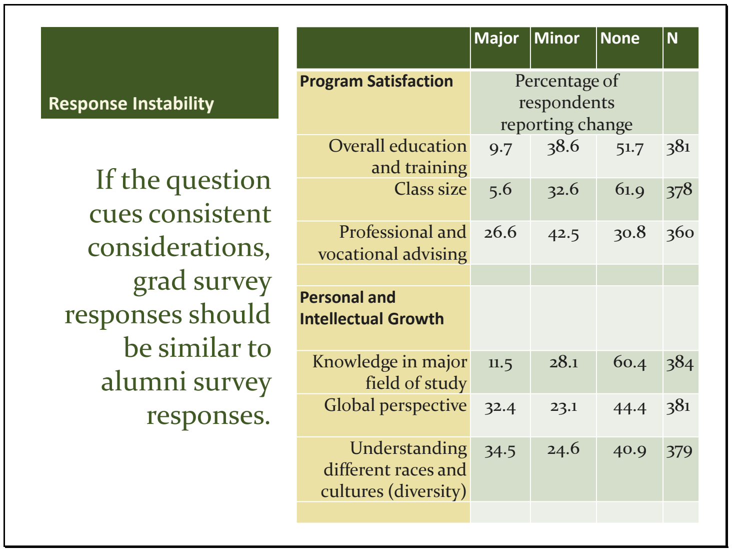


Generalization Inference

The observed score is an adequate estimate of the universe of observation.

– A2.1: The observed sample is representative of the set of considerations cued by the question weighted by the frequency with which considerations are at the top of respondents' heads.
– A2.2: The sample of observations is large enough to control random error.

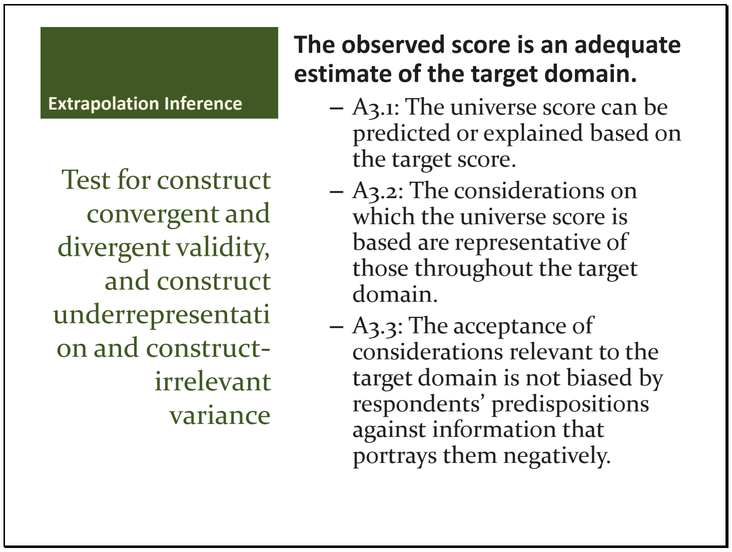
In general, this is the domain of standard reliability tests.



Response Instability

If the question cues consistent considerations, grad survey responses should be similar to alumni survey responses.

	Major	Minor	None	N
Program Satisfaction	Percentage of respondents reporting change			
Overall education and training	9.7	38.6	51.7	381
Class size	5.6	32.6	61.9	378
Professional and vocational advising	26.6	42.5	30.8	360
Personal and Intellectual Growth				
Knowledge in major field of study	11.5	28.1	60.4	384
Global perspective	32.4	23.1	44.4	381
Understanding different races and cultures (diversity)	34.5	24.6	40.9	379

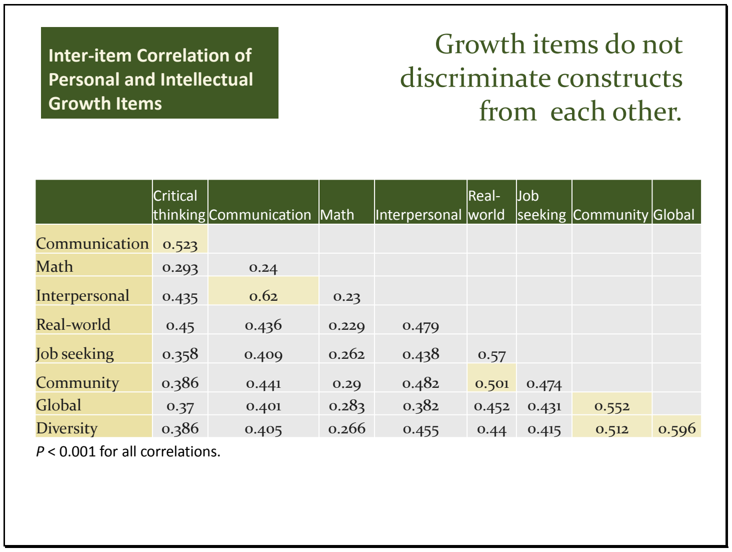


Extrapolation Inference

Test for construct convergent and divergent validity, and construct underrepresentation and construct-irrelevant variance

The observed score is an adequate estimate of the target domain.

- A3.1: The universe score can be predicted or explained based on the target score.
- A3.2: The considerations on which the universe score is based are representative of those throughout the target domain.
- A3.3: The acceptance of considerations relevant to the target domain is not biased by respondents' predispositions against information that portrays them negatively.

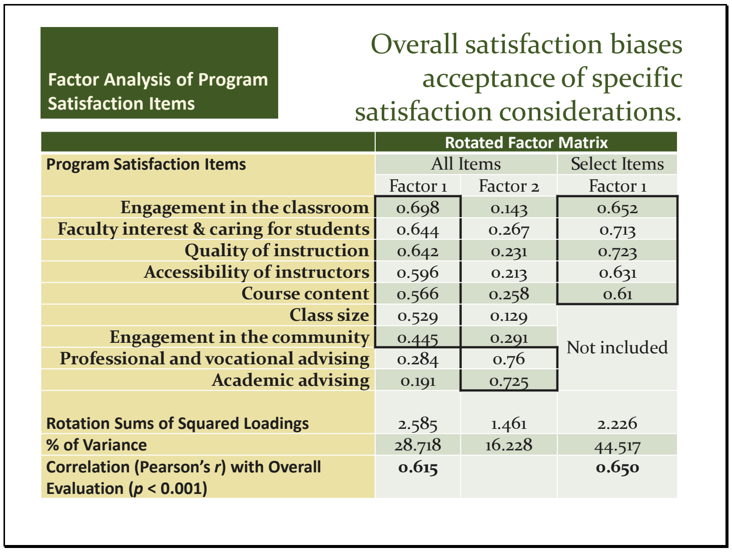


Inter-item Correlation of Personal and Intellectual Growth Items

Growth items do not discriminate constructs from each other.

	Critical thinking	Communication	Math	Interpersonal	Real-world	Job seeking	Community	Global
Communication	0.523							
Math	0.293	0.24						
Interpersonal	0.435	0.62	0.23					
Real-world	0.45	0.436	0.229	0.479				
Job seeking	0.358	0.409	0.262	0.438	0.57			
Community	0.386	0.441	0.29	0.482	0.501	0.474		
Global	0.37	0.401	0.283	0.382	0.452	0.431	0.552	
Diversity	0.386	0.405	0.266	0.455	0.44	0.415	0.512	0.596

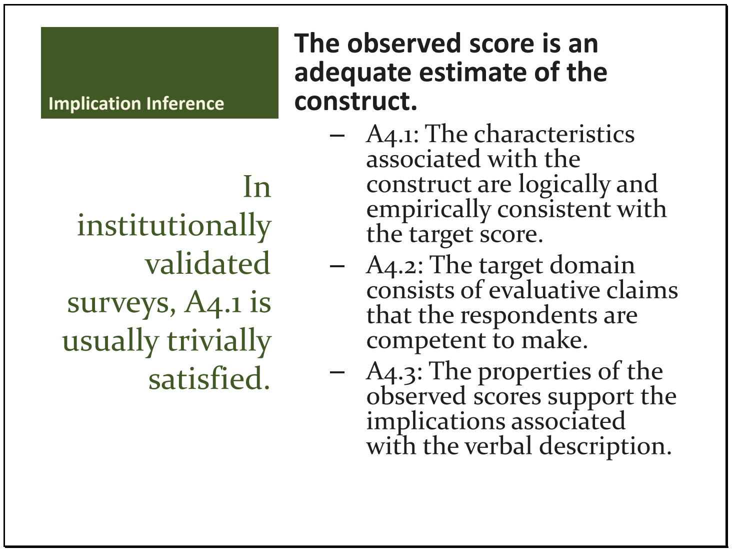
P < 0.001 for all correlations.



Factor Analysis of Program Satisfaction Items

Overall satisfaction biases acceptance of specific satisfaction considerations.

	Rotated Factor Matrix		
Program Satisfaction Items	All Items		Select Items
	Factor 1	Factor 2	Factor 1
Engagement in the classroom	0.698	0.143	0.652
Faculty interest & caring for students	0.644	0.267	0.713
Quality of instruction	0.642	0.231	0.723
Accessibility of instructors	0.596	0.213	0.631
Course content	0.566	0.258	0.61
Class size	0.529	0.129	Not included
Engagement in the community	0.445	0.291	
Professional and vocational advising	0.284	0.76	
Academic advising	0.191	0.725	
Rotation Sums of Squared Loadings	2.585	1.461	2.226
% of Variance	28.718	16.228	44.517
Correlation (Pearson's *r*) with Overall Evaluation (*p* < 0.001)	**0.615**		**0.650**

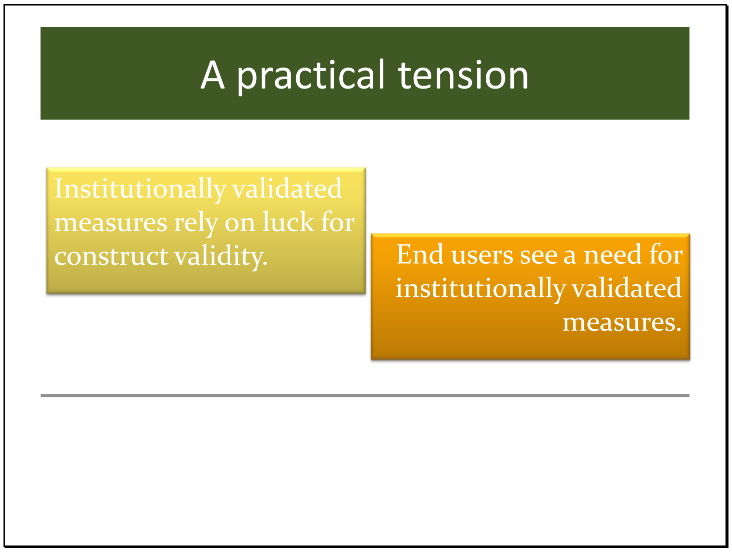


Implication Inference

The observed score is an adequate estimate of the construct.

In institutionally validated surveys, A4.1 is usually trivially satisfied.

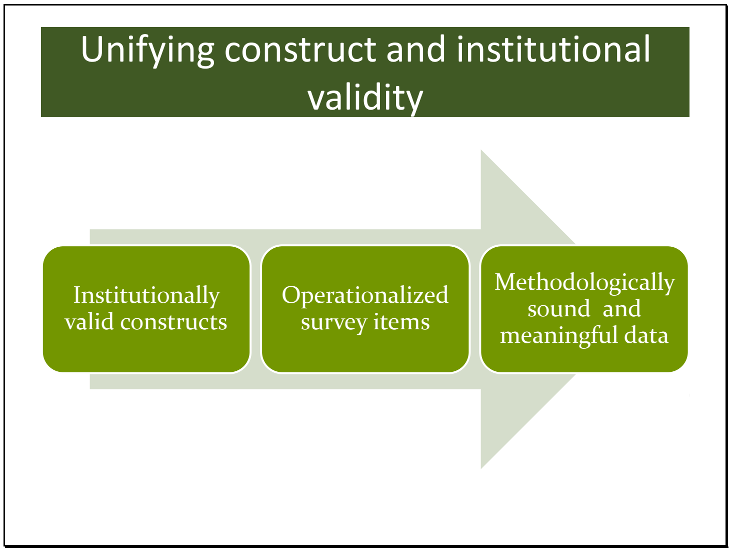
- A4.1: The characteristics associated with the construct are logically and empirically consistent with the target score.
- A4.2: The target domain consists of evaluative claims that the respondents are competent to make.
- A4.3: The properties of the observed scores support the implications associated with the verbal description.



A practical tension

Institutionally validated measures rely on luck for construct validity.

End users see a need for institutionally validated measures.



Unifying construct and institutional validity

Institutionally valid constructs

Operationalized survey items

Methodologically sound and meaningful data