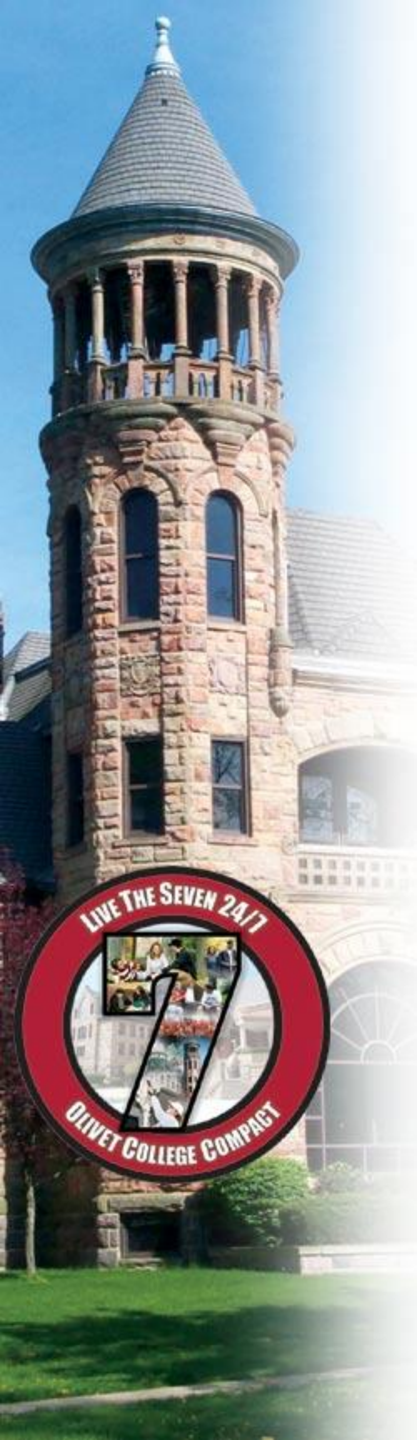


Invalidities in Causal Assessment and Questionnaire Analysis

Charles A. Graessle, Ph.D.
Director of Institutional Research

Olivet College, Olivet, Michigan

2015 MI/AIR Annual Conference, Traverse City, MI Nov 4-6

Validity in the modern age

Multiple definitions (*e.g.*, Haladyna & Downing, 2004)

1. Stability of *concept*

   over time, items, & raters, over *Ss*, internal…

2. Extensibility

   construct, criterion, predictive… , functional

Unanticipated importance & rise of

1. Respondent methods

   surveys, case studies, interviews, qualitative

2. Organizational assessment (*e.g.*, Juran, 1998)

   IR: knowledge-activity, users, producers, society

Problems with questionnaires

Cause missed if:

1. Believe question captures precise "truth"

   truth latent, broader;

   "fuzzy" thinking (*e.g.,* Ziegler, *et. al.*, 2015)

2. Focus on finding "positive" resultsx

   confirmation bias (See Nickerson, 1998)

3. Unknown validity / theory

   predictors, "heuristics" (*e.g.,* Chickering,1987)

   ≠ theories (*e.g*., learning, Bjork, 2011 and stereotype vulnerability, Ihme & Moller, 2015)

Pressure to compare institutions

Validity Issues

1. Objectivity vs “We must do what they did”

   scientific correlation needs counter-examples:

	Result was	
	Failure	Success
Them		X
Us	?	X

2. Missing data

   fair sample & their failure data
   inter-institutional confidence intervals
   individual IDs to relate to outcomes

3. Often assumes that item meaning unchanged

Pressure for high stakes testing

Definitions and examples

1. Brief observation that harms, denying graduation, job entry, etc.

   e.g., min. score to advance to next class

   e.g., test score allows one to enter a career

2. Not all requirements are high-stakes

   e.g., credit requirement for B.A. degree

   120 s.h. is over years - no penalty

   30 – 40 different assessments – no penalty

High stakes respondent methods

Validity issues

1. Negative consequences (*e.g.*, attention)
2. Individual prediction = extraordinary claim

Implications: Requires

1. More types, qualitative & quantitative

   e.g., Colorado test & teaching (Taylor, 2003)

2. Higher minimum values

   (Jonsson & Svingby, 2007)

Seeing these in an example

Student teacher evaluations

1. End-of-education internship (years ago)
2. A 35-question survey completed

   Cooperating Teachers – at midterm & finals
3. Uses (* denotes high-stakes)

   giving feedback to students

   grading/passing students on teaching *

   improving teacher prep program

   "final means" (typical accrediting requirement)

Analyzing the evaluations

Factor Analysis

Finds item groups that "vary together"

+ items correlated with a factor

- can not correct item/sample-selection biases

Assumptions

1. "Truth" is *behind* the survey

   + "factors" can be "latent" or hidden

   - naming factor is the subjective moment

2. Supports qualitative & quantitative validity

   + reduces number of items to most essential

A brief intro to factoring…

Our results

Final means were >=85%

1. Positive, “final means”-focused conclusion
2. Analyses to help dept/college:

   which parts of survey are best? trusted?

Factor Analyses performed on both sets of data

1. Example does not label items or factors
2. Interpretation based on

   number of factors found

   items which compose each factor

Midterm Evaluation

Factor matrix of cooperating teachers evaluations

(part of a rotated matrix shown –data no longer used)

	Factors At Midterm		
	1	2	3
Item 1	.281	.470	.543
Item 2	.216	.831	.154
Item 3	.298	.554	.469
Item 4	.547	.205	.552
Item 5	.328	.261	.746
Item 6	.525	.341	.410
Item 7	.545	.085	.505
Item 8	.601	.228	.160
Item 9	.764	.145	.335
Item 10	.396	.783	.271
Item 11	.431	.756	.151

Three factors identified (the overall score on this survey has 3 components)

Partial correlations (item is heavily linked to a factor if value >=.6 and low values on other factors)

Valid midterm survey needs only circled items (10-15 needed)

Final evaluation

How does this compare to “Final scores”?

(same students, class, instrument, and cooperating teachers)

	Factors at Midterm			At Finals		
	1	2	3	1	2	3
Item 1	.281	.470	.543	.340	**.635**	.144
Item 2	.216	**.831**	.154	**.625**	.284	.354
Item 3	.298	.554	.469	**.620**	.401	.307
Item 4	.547	.205	.552	**.709**	.362	.208
Item 5	.328	.261	**.746**	.208	**.843**	.174
Item 6	.525	.341	.410	.320	.231	**.818**
Item 7	.545	.085	.505	.167	.120	**.885**
Item 8	**.601**	.228	.160	.383	.344	.440
Item 9	**.764**	.145	.335	.537	.373	.409
Item 10	.396	**.783**	.271	.571	**.608**	.266
Item 11	.431	**.756**	.151	.562	**.623**	.266

In this example only 2 of 11 items remain associated. All other item-loadings changed

Factor means can not be compared. Instead, we must explain why the *factors* differ.

We must describe a qualitative change

Notice items that are necessary/ which are not

(same students, class, instrument, and cooperating teachers)

	Factors at Midterm			At Finals		
	1	2	3	1	2	3
Item 1	.281	.470	.543	.340	**.635**	.144
Item 2	.216	**.831**	.154	**.625**	.284	.354
Item 3	.298	.554	.469	**.620**	.401	.307
Item 4	.547	.205	.552	**.709**	.362	.208
Item 5	.328	.261	**.746**	.208	**.843**	.174
Item 6	.525	.341	.410	.320	.231	**.818**
Item 7	.545	.085	.505	.167	.120	**.885**
Item 8	**.601**	.228	.160	.383	.344	.440
Item 9	**.764**	.145	.335	.537	.373	.409
Item 10	.396	**.783**	.271	.571	**.608**	.266
Item 11	.431	**.756**	.151	.562	**.623**	.266

Science tells us that items 1, 3, 4, 6, and 7 are newly-emphasized at finals

items 8 & 9 are now less important

Is factoring related to score increase?

More items became important than less important

(same students, class, instrument, and cooperating teachers)

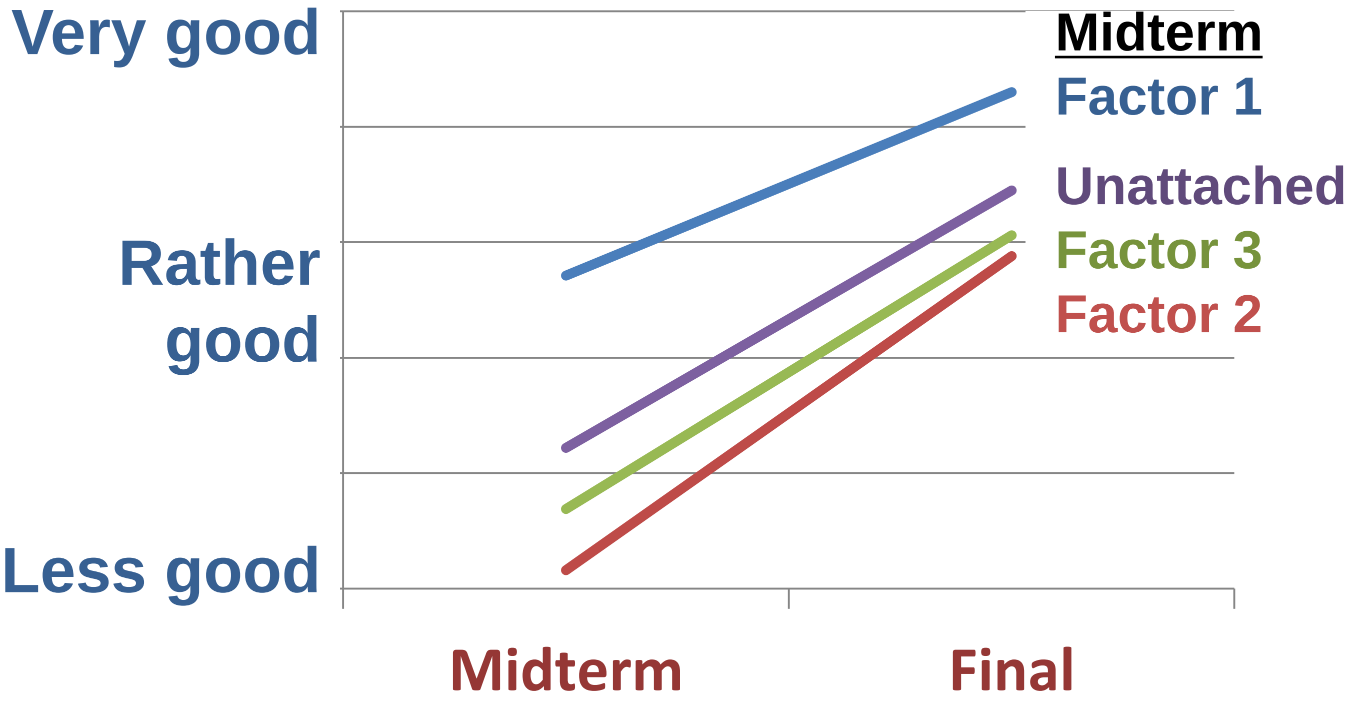
	Factors at Midterm			At Finals		
	1	2	3	1	2	3
Item 1	.281	.470	.543	.340	**.635**	.144
Item 2	.216	**.831**	.154	**.625**	.284	.354
Item 3	.298	.554	.469	**.620**	.401	.307
Item 4	.547	.205	.552	**.709**	.362	.208
Item 5	.328	.261	**.746**	.208	**.843**	.174
Item 6	.525	.341	.410	.320	.231	**.818**
Item 7	.545	.085	.505	.167	.120	**.885**
Item 8	**.601**	.228	.160	.383	.344	.440
Item 9	**.764**	.145	.335	.537	.373	.409
Item 10	.396	**.783**	.271	.571	**.608**	.266
Item 11	.431	**.756**	.151	.562	**.623**	.266

If so, then new item scores were lower on midterm

And item 8 & 9 scores were higher at midt.

An objective test of that prediction

Relation of qualitative change to evaluations



No. Unattached improved at about same rate. See ANOVAs
All items, factors improved & were not different at Final.

Conclusions from this data

Student teacher success based on means

1. All improved, but all same by final

   performed a lot in last half ?

   work remembered better by final?

Possible confounds/ validity concerns

1. Untheorized factor structure effects
2. Less discriminating at final

   less time => less serious

3. More likely to hurt student (high stakes eval)
4. Evaluator may be hurt

Putting IR on the screen

Advising about knowledge-activity

1. Be theoretical about respondent methods

   imprecision of soc science knowledge

   qualitative & modern analyses

2. Remove invalidity pressures

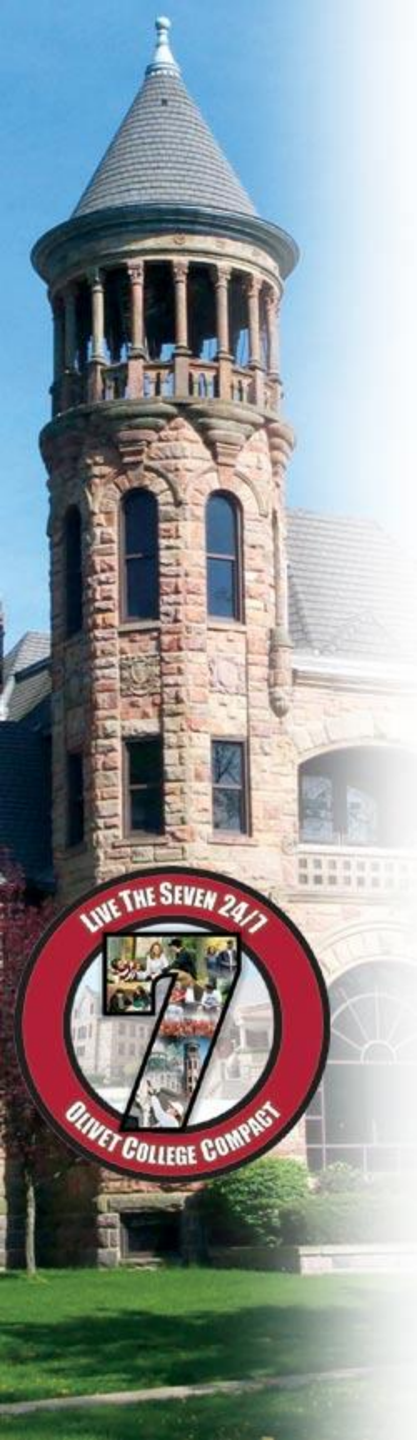
   design equivalency (*e.g.*, factor structures)

   trust & respect ≠ high stakes decision-making

3. Promote sophistication in interpretation

   assessor, Board, administrator skills

   learn to help each other



Invalidities in Causal Assessment and Questionnaire Analysis

Questions/Comments

cgraessle@olivetcollege.edu

For example…

Group by how respondents answer items

Students in Michigan indicate amount of agreement with items where 4=Strongly agree and 1=Strongly disagree

	Respondent			
Item	#1	#2	#3	#4
I originally lived near Michigan	4	4	1	1
I originally lived near Texas	1	1	4	4
The sky is blue here	4	4	3	2
I live with non-Earth beings	1	1.5	1	1.5

These are negatively related, but are still responded to in the "same" way

But this item is not related to any others

Original survey responses…

	Respondent			
Item	#1	#2	#3	#4
I originally lived near Michigan	4	4	1	1
I originally lived near Texas	1	1	4	4
The sky is blue here	4	4	3	2
I live with non-Earth beings	1	1.5	1	1.5

yields 2 different factors(bold-faced)…

	Factor	
Item	#1	#2
I originally lived near Michigan	**.995**	.044
I originally lived near Texas	**-.955**	-.044
The sky is blue here	**.940**	-.279
I live with non-Earth beings	-.038	**.997**

Return

Objective tests of factor improvement

Items unattached at midterm improved like others

1. 2 x 2 repeated measures ANOVA on means

   2 different evaluation times (Midterm vs Finals) and whether items were or were not part a factor at midterm

Effect (Source)	SS	df	MS	F	*p*
Time: Midterm or Finals	12.986	1	12.986	113.24	.000*
Belonged to a Factor	.044	1	.044	2.013	.157
Time X Belonging to Factor	.022	1	.022	.120	.729
Error	3.481	267			

Objective tests of factor improvement

Midterm factors all improved, but at different rates

2. 4 x 2 repeated measures ANOVA on means

2 different evaluation times (Midterm vs Finals) and
3 midterm factor item was attached (or was unattached)
Greenhouse-Geisser adjusted *df* used

Effect (Source)	SS	df	MS	F	*p*
Time: Midterm or Finals	26.629	1	26.629	116.34	.000*
Error (Time)	59.74	261	.229		
Factor at Midterm	11.394	2.637	4.262	68.66	.000*
Error (Factor)	43.314	697.72	.062		
Time X Factor	1.005	2.858	.352	14.64	.000*
Error	17.921	746.04	.024		

Return

Citations

Bjork, E.L., and Bjork, R. A. (2011). On the symbiosis of learning, remembering, and forgetting. In A. S. Benjamin (Ed.), ***Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork,*** (pp. 1-22). London, UK: Psychology Press.

Chickering, A. W. and Gamosn, Z. F. Seven principles for good practice in undergraduate education. ***AAHE Bulletin.*** 3-7.

Haladyna, T. M., and Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. ***Educational Measurement, Issues and Practice, 23***(1), 17-27.

Ihme, T. I., and Moller, J. (2015). "He who can, does; he who cannot, teaches?": Stereotype threat and preservice teachers. ***Journal of Educational Psychology, 107***(1), 300-308.

Citations (continued)

Jonsson, Anders, and Svingby, Gunilla. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. ***Educational Research Review, 2***(2), 130-144.

Juran, J. M. (1998). How to think about quality. In Juran, J.M. & Godfrey, A. B. (Eds), ***Juran's Quality Handbook, 5th Edition.*** New York: McGraw-Hill.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. ***Review of General Psychology, 2***(2), 175-220.

Citations (continued)

Taylor, G, Shepard, L, Kinner, F., and Rosenthal, J. (2003). A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost. Technical Report #588, Los Angeles, Center for the Study of Education. Downloaded form internet: http://www.cse.ucla.edu/products/Reports/TR588.pdf

Zielger, Matthias, Kemper, Christoph J., and Lenzner, Timo. (2015). The issue of fuzzy concepts in test construction and possible remedies. ***European Journal of Psychological Assessment, 31***(1), 1-4.