Reliability and Validity of Instructor Course Evaluations: Exploring the Myths

VANGUARD UNIVERSITY

* your story matters:

John Kim, PhD, Associate Director of Institutional Research

Philip Newlin, Institutional Research Analyst

Ludmila Praslova, PhD, Professor & Director of Organizational Psychology & ALO

NOVEMBER 15, 2018 2018 CALIFORNIA ASSOCIATION FOR INSTITUTIONAL RESEARCH

VANGUARD UNIVERSITY

- A private, Christian university of liberal arts and professional studies in Costa Mesa
- Founded in 1920 by the Assemblies of God to train military chaplains
- 4-year Bachelor programs in 30 majors and graduate and professional programs in 12 majors, enrolling about 2,200 students
- University Motto = TRUTH, VIRTUE, SERVICE

SURVEY TIME!

tinyurl.com/CAIRVU2018

INTRODUCTION

- Most higher education institutions use instructor course evaluations to evaluate and improve the teaching effectiveness of their faculty
- However, its **usefulness and validity have been frequently challenged**
- Many people have claimed that evaluations are affected by several factors, such as gender, physical attractiveness, race/ethnicity, and academic discipline (course difficulty), etc.

COURSE EVAL CONCERNS

- “Student ratings are unreliable and invalid”
- “Student ratings are just popularity contests”
- “Students will not appreciate good teaching but just want easy courses” (Benton & Cashin, 2012, p.2)
- “Language students use in evaluations regarding male instructors is different than language used in evaluating female instructors” (Falkoff, 2018)
- “We must stop relying on student ratings of teaching” (Benton & Cashin, 2012, p.2)

PURPOSE

- **To examine the reliability and validity of instructor course evaluations**
  - Using the results of traditional UG courses administered in EvaluationKIT, separately for **4 consecutive semesters** (2016FA – 2018SP) to check for **cross-validation**
  - A total of 1,364 courses with 28,181 student responses
- **To explore some of the myths surrounding student ratings of teaching effectiveness**
  - In a private, non-profit religious affiliated college

PART 1.

Reliability and Validity of Course Evaluation

RELIABILITY

- Reliability: accuracy, consistency, and prerequisite to validity (Mitchell & Jolley, 2010)
- Used **10 opinion questions** asking instructor's teaching effectiveness
- Showed very high reliability continuously across the 4 semesters

: **Cronbach's alpha >.90**

Reliability Statistics

Term	Cronbach's Alpha	N of Items
2016 Fall	.941	10
2017 Fall	.941	10
2017 Spring	.926	10
2018 Spring	.939	10

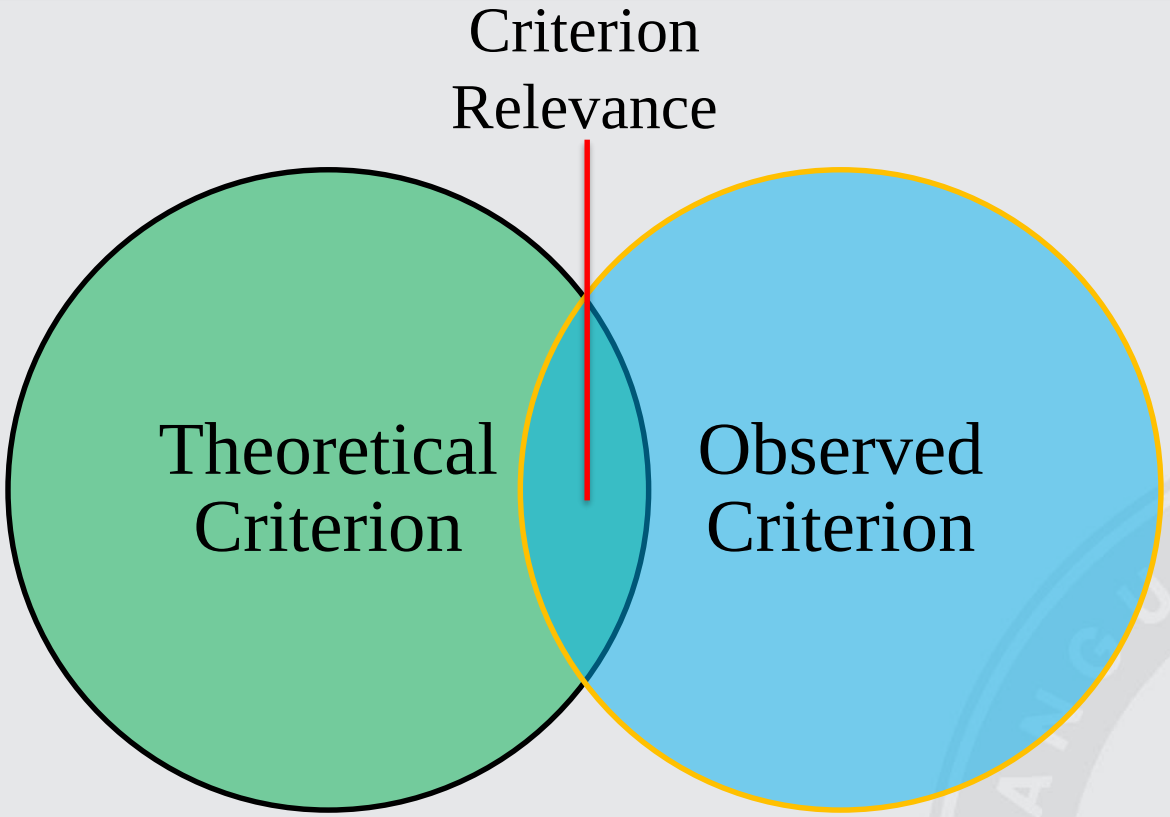
CONSTRUCT VALIDITY

- The degree to which the scale measures what it is supposed to measure
- Agreement between a test score or measure and the quality it is believed to measure (Kaplan & Saccuzo, 2001)
- Soundness and relevance of a proposed interpretation (measurement) (Cronbach, 1990)
- What the course evaluation is supposed to measure is **TRUE TEACHING EFFECTIVENESS (θ)**

Teaching Effectiveness

Teacher's Popularity

CRITERION PROBLEM



: A large part of the observed criterion that is being used to represent the theoretical criterion may actually be measuring other things

CRITERION PROBLEM

- Students cannot always effectively assess their own learning, and grade point is not generalizable or standardized
- This study serves as an *initial exploration of our course evaluation system* using available data
- **Additional objective criteria** are required to effectively assess true teaching effectiveness; both course grade and course evaluations are subject to substantial criterion contamination

CONSTRUCT VALIDITY

Estimator of true teaching effectiveness ($\hat{\theta}$)

- Average course evaluation score of all teaching courses over 4 semesters for each instructor
- Used courses with n ≥10 only in order to reduce SE
- Used data of **42 faculty** for a total of **440 courses**
- Aggregate data does not control for **systematic factors,** however:
- A large enough data set could control for unwanted variance

CONSTRUCT VALIDITY(CONT)

- Individual course evaluation scores showed **high correlations (> .60) with $\hat{\theta}$**
- Results indicate **acceptable evidence** of construct validity

		$\hat{\theta}$
2016FA	Ind. Course Eval	**.619****
2017SP	Ind. Course Eval	**.752****
2017FA	Ind. Course Eval	**.701****
2018SP	Ind. Course Eval	**.768****

- Based on Cohen's guideline of the effect size of correlation coefficient as follows: small=0.10, medium=0.30, large=0.50.

CONSTRUCT VALIDITY(CONT)

- **"Theoretically, the best indicant of effective teaching could be student learning outcome.** Other things being equal, the students of more effective teachers should learn more" (Benton & Cashin, 2012, p.3)
- **$\widehat{\theta}$ showed significant correlations (> .30) with Average Course Final Grade** for each instructor over 4 semesters, which supports the claim that "students of more effective teachers should learn more"
- **Individual course evaluations, however, had no or very low correlations with course grades.** Individual course evaluation can be affected by various factors (gender, class size, division, discipline, faculty type, etc.)

CONSTRUCT VALIDITY(CONT)

		Avg Course Grade	Ind. Course Grade
2016FA	Ind. Course Eval	.128	.112*
	$\hat{\theta}$	**.308****	.116
2017SP	Ind. Course Eval	.159*	.094
	$\hat{\theta}$	**.301****	.130
2017FA	Ind. Course Eval	.051	.071
	$\hat{\theta}$	**.316****	.168*
2018SP	Ind. Course Eval	**.329****	.115*
	$\hat{\theta}$	**.333****	.038

* p <.05, ** p <. 01

PART 2.

Prediction Modeling for Course Evaluations

REGRESSION MODELING

- Regression modeling of Course Evaluation on the following various factors was conducted:

1) **Gender** (Male, Female)
2) **Faculty Type** (Adjunct, Term-contract, Tenure-track, Tenured)
3) **Ethnicity** (White, Non-White)
4) **Degree** (Bachelor, Master, Terminal, Doctoral)
5) **Course Division** (Lower, Upper)
6) **Class Size**
7) **Course Final Grade**
8) **Discipline** (BUSN, Humanities, Fine Arts, Social Science, Natural Science, & Religion)

*Data of 1364 courses with 28,181 student responses 2016FA -2018SP were used

SIG/NON-SIG PREDICTORS

	2016FA	2017SP	2017FA	2018SP
Discipline1 (BUSN < Relig)	***	***	*	*
Discipline5 (Natural Sci < Relig)	*	***	***	***
Discipline2 (Humanities < Relig)	*	***	NS	*
Discipline3 (Fine Arts < Relig)	NS	NS	*	NS
Ethnic1 (Asian < White)	***	*	NS	**
Ethnic2 (Afr Ame < White)	*	NS	***	***
Ethnic3 (Hispanic < White)	NS	***	NS	NS

* $p < .05$, ** $p < .01$, *** $p < .001$, NS=Not Significant

SIG/NON-SIG PREDICTORS

	2016FA	2017SP	2017FA	2018SP
Course Final Grade (positive corr)	**	NS	NS	**
Degree2 (Master < Doctoral)	NS	NS	*	**
Degree3 (Terminal < Doctoral)	*	NS	*	***
Class Size (negative corr)	NS	NS	*	NS
Gender (Female > Male)	NS	NS	NS	***
Division (Lower < Upper)	*	NS	*	NS

* p <.05, ** p <. 01, *** p <. 001, NS=Not Significant

RESULT SUMMARY

After controlling for all the factors listed above,

- **Discipline**, especially for **Business & Natural Science**, was a very significant factor for course evaluation scores. Business and Natural Sciences showed lower scores than Religion for all 4 semesters.
- **Ethnicity** (White vs others), **Faculty degree** (Doctoral vs others), **Course final grade** (positive corr.) were significant in 2 or 3 semesters
- However, **Gender, Class size, Division**, and **Faculty type** were NOT significant in 3 or all 4 semesters.

PART 3.

Item & Factor Analysis

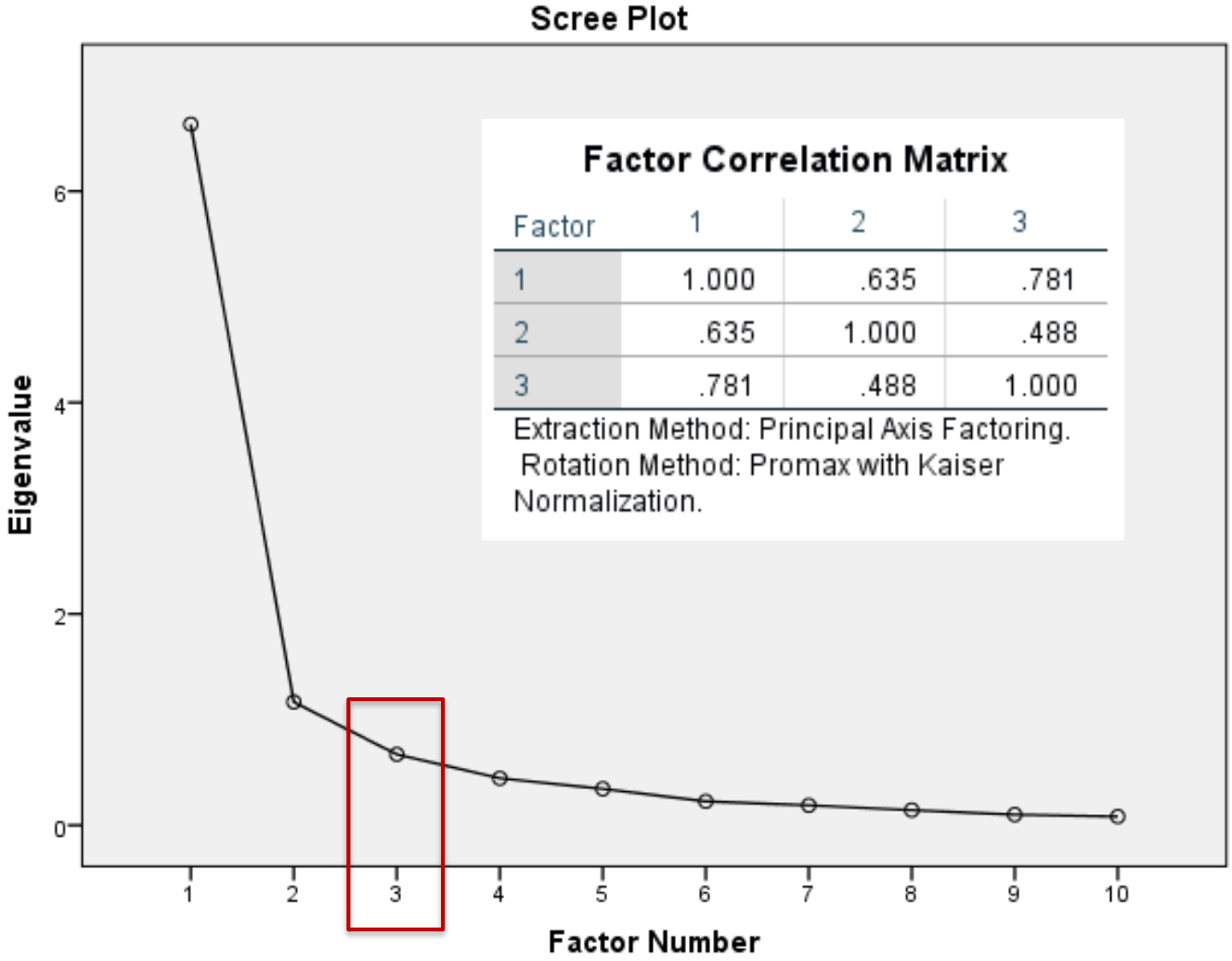
ITEMS AND SUB-SCALES

Sub-scales (3)	Items (10)
Instruction-related	#1 Explaining the course requirements. #2 Preparation for each class session. #3 Effective class time management. #6 Responsiveness to questions. #7 Availability to help outside of the classroom.
Assignment-related	#8 Grading & Returning assignments in a reasonable amount of time. #9 Following course syllabus for the course content and the pace. #10 Helpfulness of the assignment for learning.
Faith	#4 Exhibiting Christian worldview. #5 Integration of faith with course content.

ITEM & FACTOR ANALYSIS

- Used all data (1364 courses with 28,181 students responses for 4 semesters)
- **Factor Analysis**
  - Principal Axis Factoring,
  - Promax with Kaiser Normalization rotation method
  - 3 sub-scales ➔ 3 factor structure
- **Item Analysis**
  - Item-total correlations were computed for each subscale

3-FACTOR MODEL (OVERALL)



Factor Correlation Matrix

Factor	1	2	3
1	1.000	.635	.781
2	.635	1.000	.488
3	.781	.488	1.000

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

FACTOR PATTERN (OVERALL)

- Showed clear 3-factor structure
- Same pattern for all semesters

Pattern Matrix[a]

	Factor 1	2	3
Classtime_MEan Inst 3	1.000	-.084	-.028
Prepared_Mean Inst 2	.928	-.065	.033
Coursereq_Mean Inst 1	.773	.011	.160
Responsive_mean Inst 6	.728	.236	-.043
Available_mean Inst 7	.445	.331	.084
ChristWV_Mean Inst 4	-.025	.988	-.024
IntFaith_Mean Inst 5	-.013	.925	.039
Syllabus_mean SLO 2	.108	-.097	.925
AssignReturn_mean SLO1	-.074	.097	.728
Activities_mean SLO 3	.398	.066	.467

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

ITEM ANALYSIS (OVERALL)

Scales	Items	Item-Scale Corr.
Instruction-related (Alpha=.94)	#1 Course requirements.	.867
	#2 Class preparation.	.863
	#3 Time management.	.867
	#6 Responding questions.	.824
	#7 Availability to help	.714
Assignment-related (Alpha=.86)	#8 Grading & Returning assignments in time	.664
	#9 Observing Course syllabus	.826
	#10 Helpful assignment	.738
Faith (Alpha=.92)	#4 Christian worldview.	.852
	#5 Integration of faith	.852

CONCLUSION

- Course evaluation survey showed **high reliability** and **acceptable construct validity** in the preliminary study using a **multi-year average of evaluation scores as an estimator of true teaching effectiveness** (=theoretical criterion)
- The following factors appear to affect Individual course evaluation scores very or somewhat significantly when all other factors are controlled for:

  -Discipline (BUSN, Natural Science),

  -**Ethnicity** (White), **Faculty degree** (Doctoral)**,** and

  Course final grade (positive corr with course eval)

CONCLUSION (CONT.)

- However, **Gender, Class size, Division**, and **Faculty type** do not seem to affect course evaluation scores
- Student ratings of instruction can still provide insight that improves teaching ability
- We recommend using a multi-year average (2-3 years) of course evaluation scores with *at least 10 responses*
- Also, additional objective criteria (other than average course grade) will be necessary to test the true construct validity of the course evaluation in the future

REFERENCES

- Benton, S. L. & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50).
- Cronbach, L. (1990). *Essentials of Psychological Testing* (5th ed.). New York, NY: Harper Collins Publishers
- Falkoff, M. (2018). *Why we must stop relying on student ratings of teaching,* ChronicleVitae. Retrieved from http://chroniclevitae.com
- Kaplan, R. & Saccuzo, D. (2001). *Psychological Testing* (5th ed.). Belmont, CA: Wadsworth/Thompson Learning.