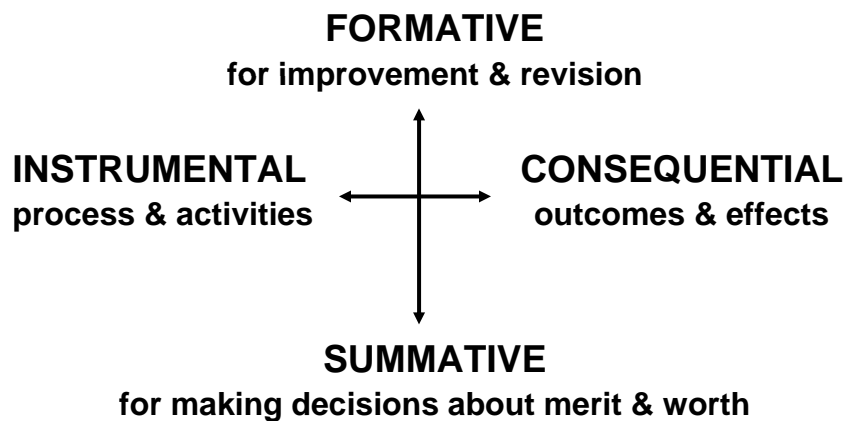


Choosing and Developing Reliable Instruments

Evaluating Faculty Performance
in the Classroom
June 24, 2008
Michael Theall, Ph.D.

1

Evaluation Purposes and Data



2

ACADEMIC - IMPRESSIONS

Uses of Data

Personnel Decisions

Overall performance

Quantitative outcome oriented

Comparative database

Empirical direct unambiguous

Global items

Public

Teaching Improvement

Assessable modifiable behaviors

Qualitative process oriented

Informative database

Comprehensive/detailed/suggestive

Specific "low inference" items

Confidential

Supporting data

3

ACADEMIC - IMPRESSIONS

COMPARISON OF DATA TYPES FOR EVALUATING CLASSROOM PERFORMANCE

TYPE OF DATA	RELIABLE	VALID	GENERALIZABLE	MANIPULABLE	SECURE	USEFUL (F)	USEFUL (S)
STUDENT RATINGS	mostly	mostly	maybe	possible	mostly	yes	yes
SELF RATINGS	maybe	yes	n/a	possible	mostly	yes	maybe
PORTFOLIO DATA	maybe	mostly	n/a	possible	mostly	yes	maybe
PEER RATINGS	maybe	maybe	n/a	possible	mostly	yes	maybe
ADMINISTRATOR RATINGS	maybe	maybe	n/a	possible	mostly	yes	maybe
EXTERNAL EXPERT RATINGS	maybe	maybe	n/a	possible	mostly	yes	maybe
CLASSROOM TESTS	maybe	mostly	maybe	possible	mostly	yes	maybe
VALIDATED EXTERNAL TESTS	maybe	maybe	yes	possible	mostly	yes	maybe
MEDIA DOCUMENTS	yes	maybe	n/a	possible	mostly	yes	maybe
OTHER ASSESSMENT DATA	maybe	maybe	maybe	possible	mostly	yes	maybe
SCHOLARLY PRODUCTIVITY	yes	no	n/a	possible	yes	no	no
ALUMNI OR EMPLOYER DATA	maybe	maybe	maybe	possible	maybe	maybe	maybe
HONORS & AWARDS	maybe	maybe	n/a	possible	yes	yes	yes

4

ACADEMIC - IMPRESSIONS

Common instrumentation needs

1. Student ratings
2. Peer ratings
3. Administrator ratings
4. Self ratings
5. External observer ratings

May use identical or similar instruments with special, additional sets of questions

5

ACADEMIC - IMPRESSIONS

General Characteristics of Good Systems/Instruments

- **COMPREHENSIVE**
 - **FAIR**
 - **FLEXIBLE**
 - **ADAPTABLE**
 - **EVALUATED**

6

Technical Characteristics of Good Systems/Instruments

- RELIABILITY
- VALIDITY
- GENERALIZABILITY
- FEASIBILITY
- SKULLDUGGERY

7

Choosing Instruments

FACTORS TO CONSIDER

- SCOPE
- FUNCTIONALITY
- COST
- FEASIBILITY
- APPLICABILITY
- ACCEPTABILITY

8

ACADEMIC - IMPRESSIONS

Instrument types compared

<u>LOCAL</u>	<u>NATIONAL</u>
Specific	General
Content control	Few options
Must be validated	Validated
Analysis needed	Analysis provided
No norms	Established norms
Reports created	Reports standard
Interpretation?	Guidelines provided

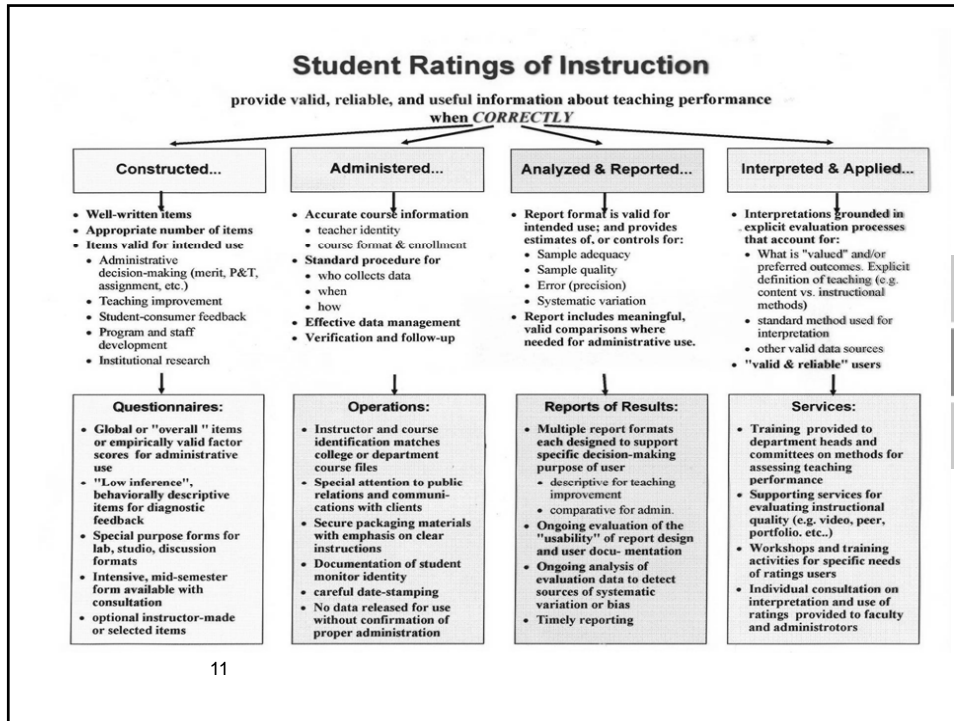
9

ACADEMIC - IMPRESSIONS

National/Commercial Evaluation Instruments

- IDEA (IDEA Center, Manhattan KS)
- SIR II (ETS, Princeton, NJ)
- CIEQ (Aleamoni, U AZ)
- SEEQ (Marsh, U Oxford)
- Institutional instruments: Purdue, Illinois, Washington, etc.

10



ACADEMIC IMPRESSIONS

BREAK

- Questions?
- Comments?
- Issues?

- An activity

12

ACADEMIC - IMPRESSIONS

On-campus ratings

VARIOUS TYPES OF RATINGS							
LOCAL POLICY AND PRACTICE							
TYPE OF DATA	CRITICAL?	USED?	INSTRUMENT?	SECURE?	PUBLIC?	USEFUL (F)?	USEFUL (S)?
STUDENT RATINGS FROM A LOCAL SYSTEM							
STUDENT RATINGS FROM AN EXTERNAL VENDOR							
STUDENT RATINGS FROM A SURVEY TOOL							
PEER RATINGS							
ADMINISTRATOR RATINGS							
SELF RATINGS							
EXTERNAL OBSERVER RATINGS							

13

ACADEMIC - IMPRESSIONS

On my campus...

What kinds of ratings are collected?
 Are the data equally weighted?
 How are they used?

Please respond and compare your response with those of 2 other people

14

Constructing Instruments

BASIC PROCESS

- Establish the purpose
- Identify the domain/focus
- Identify the components
- Consider the overall organization
- Weigh/prioritize elements
- Consider analysis & reporting
- Consider feasibility

15

Constructing Instruments

ASSEMBLING ITEMS

- Factor structures
- Global items
- Teacher items
- Course items
- Student demographics / information
- Open-ended items
- Additional (free) items
- Assessment / other items

16

Constructing Instruments

IMPLEMENTATION ISSUES

- Initial review (faculty, students, administrators)
- Pilot test
- Revision
- Field test
- Validation studies
- Data banking
- Norm development

17

Constructing Instruments

IMPLEMENTATION ISSUES

- Analysis / reporting formats
- Interpretive guides
- User training
- Norm vs. criterion referencing
- Coordination with development
- 2-3 year review
- Item banking

18

Constructing Items

ITEM STEM

- Clear / concise
- Focused / direct
- Single concept
- Individual response
- Available or created?
- Global or low-inference?
- Duplicate other items

19

Constructing Items

ITEM STEM

- Simple sentence
- Understandable (reading level)
- Correct syntax
- Avoids strong/loaded words/terms
- Match the intended behavior
- Applicable to all students
- States a fact

20

Constructing Items

ITEM STEM

- Answerable by all
- Unambiguous
- Identify positive/negative behavior
- Use absolute terms
- Use vague terms
- Use unfamiliar terms
- Elicit variant responses

21

BREAK

- Questions?
- Comments?
- Issues?

- An activity

22

ACADEMIC - IMPRESSIONS

Writing item stems

- With a team member or person(s) from a similar institution/situation, write 2 item stems that address a similar issue (e.g., clarity; organization; rapport; etc.)

23

ACADEMIC - IMPRESSIONS

Constructing Items

RESPONSE SCALE

- Odd (balanced) or even (forced choice)
- Number of items (4-7)
- Numeric or verbally anchored or both
- Degree of inference required
 1. Frequency
 2. Likert SA-SD
 3. Verbal labels
 4. Semantic differential
- Standard or reverse scoring?

24

Constructing Items

RESPONSE SCALE

- Consistent with stem
- Logical / appropriate
- Consistent in wording/syntax
- Concrete
- Elicit variant responses
- Balanced (bipolar)
- Understandably graduated (polar)

25

BREAK

- Questions?
- Comments?
- Issues?

- An activity

26

ACADEMIC - IMPRESSIONS

Creating response scales

- With a team member or person(s) from a similar institution/situation, create a response scale for the two item stems you just developed. If you need to create two scales, why so? Would one scale suffice for both items?

27

ACADEMIC - IMPRESSIONS

Operations

- Data collection / sampling
- Analysis issues
- Report formats
- Interpretation & use
- Additional investigation / analysis
- Decision-making

28

Sample size recommendations

<u>Class size</u>	<u>Minimum safe sample**</u>
5 – 10	at least 90%
10 - 20	at least 80%
20 – 30	at least 75%
30 – 50	at least 66%
50 – 100	at least 60%
100 +	more than 50%

** = assuming no systematic bias (e.g., non-response)

29

Analysis possibilities for reports of results

- **Descriptive statistics**
(item distributions in # and %)
- **Central tendency (mean, mode, median)**
 - 1 2 3 3 4 5 (3, 3, 3)
 - 1 1 2 3 4 5 5 (3, 1 & 5, 3)
 - 1 2 3 4 5 5 (3.33, 5, 3.5)
 - 1 2 3 4 5 5 5 (3.57, 5, 4)
- **Standard deviations & sampling error**

30

Analysis possibilities for reports of results

- **Standard scores**
- **Comparative data (norms, criterion references, self-ratings)**
- **Ranges (%ile rank, %ile group, confidence intervals for self and comparison groups)**
- **Enrolled / response #s and ratio**

31

Analysis possibilities for validation

- **Item analysis**
- **Reliability coefficients**
- **Correlational analysis**
- **Factor analysis**
- **Regression analysis**

on entire database and on subsets as soon as enough data is available

32

ACADEMIC - IMPRESSIONS
INSTRUCTIONAL REPORT of EDUCATIONAL
SATISFACTION: (I.R.E.S.)

Universitas pro Omnibus Discipuli
 et Facultitas in Excelcis

Instructor: U.N. Fortunate
Course #: HIS123
Course name: History of Everything
Term/year: Spring, 1994

	A	B	C	D	E	F	O
amount learned	3	16	46	21	14	0	1
overall teacher	1	12	40	29	18	0	0
overall course	2	8	49	20	11	0	0

Note: (A) =5= Best; (E)=6=Worst ... Enrolled: 120; Responded: 53

ACADEMIC - IMPRESSIONS
INSTRUCTIONAL REPORT of EDUCATIONAL
SATISFACTION: (I.R.E.S.)

Universitas pro Omnibus Discipuli et Facultitas in Excelcis

Instructor: U.N. Fortunate
Course #: HIS123
Course name: History of Everything
Term/year: Spring, 1994

% / # responses >	A	B	C	D	E	F	O	mean	s d	T	grp
amount learned	3/2	16/10	46/29	21/13	14/9	0/0	1/1	2.64	0.88	27	low
overall teacher	1/1	12/8	40/25	29/18	18/11	0/0	0/0	2.43	0.96	24	low
overall course	2/1	18/11	49/31	20/13	11/7	0/0	0/0	2.81	0.93	33	low

Raw score: (A) =5= Best; (E) =1= Worst; F= Not applicable; O = Omitted;
 Enrolled = 120; Responded = 63: (sample adequate)
 T-score: Standardized score where 40 - 60 = mean, and each 10 points in each direction is one standard deviation
 Group score: 0-10% = low; 10-30% - low middle; 30-70% = middle; 70-90% = high middle; 90-100% = high

ACADEMIC - IMPRESSIONS

Two evaluations of HIS 345

	mean	s d	T	group	
amount learned	3.35	0.87	<u>45</u>	low-mid	term/yr = spring, 1995 instr = UNFortunate course = his 345 resp/enr = 29/61 % resp=48
overall teacher	2.76	0.76	35	low	
overall course	2.85	0.90	37	low	
<hr/>					
amount learned	3.97	1.40	<u>56</u>	hi-mid	term/yr= fall, 1995 Instr = UNFortunate course = his 345 resp/enr = 20/42 % resp=48
overall teacher	3.57	1.30	<u>47</u>	mid	
overall course	3.63	1.24	<u>50</u>	mid	

35

ACADEMIC - IMPRESSIONS

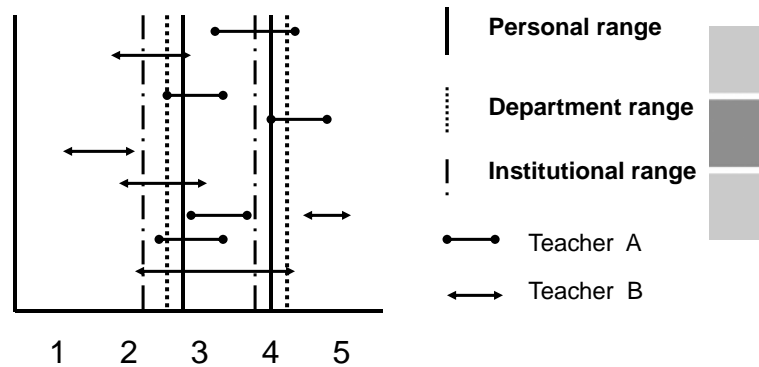
Enrollment profiles for HIS 345 in two semesters

	Fr	So	Jn	Sn	Tot	
original enr.	6	17	15	23	61	term/yr= spring, 1995 Instr = UN Fortunate course = his 345 resp/enr = 29/51 % resp=57
final enr.	5	14	12	20	51	
eval respondents	5	13	11	0	29	
<hr/>						
original enr.	3	11	12	16	42	term/yr= fall, 1995 Instr = UN Fortunate course = his 345 resp/enr = 20/29 % resp=69
final enr.	2	7	8	12	29	
eval respondents	2	4	5	9	20	

36

ACADEMIC - IMPRESSIONS

Graphic display of 95% confidence intervals or individuals vs. comparison groups



37

ACADEMIC - IMPRESSIONS

For access to more information, go to:

*Enhancing Your Teaching Through Use of the SIR II Report:
Suggestions for Improvement*

for a compendium of information on using evaluation for teaching improvement

(at the ETS website, go to the SIR II link & then to the "research" tab to find the PDF file)

<http://idea.ksu.edu>

for very useful discussions of teaching and evaluation, see the "POD-IDEA Notes" and "POD-IDEA Learning Notes"

<http://ntlf.com/pod/index.html>

for a review of the research and an extended/annotated bibliography

38

ACADEMIC - IMPRESSIONS

Basic References

- Arreola, R. A. (2007) *Developing a Comprehensive Faculty Evaluation System*. (3rd ed.) Bolton, MA: Anker Publishing Company.
- Berk, R. A. (2006) *Thirteen strategies to measure college teaching*. Sterling, VA: Stylus Publishing.
- Cashin, W. E. (1990) Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.) "Student ratings of instruction: issues for improving practice." *New Directions for Teaching and Learning # 43*. San Francisco: Jossey Bass.
- Centra, J. A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*. San Francisco: Jossey-Bass.

39

ACADEMIC - IMPRESSIONS

Basic References

- Franklin, J. & Theall, M. (2002) "Thinking about faculty thinking about teacher and course evaluation results." In N. Hativa & P. Goodyear. *Teacher thinking, beliefs, and knowledge in higher education*. Dordrecht, the Netherlands: Kluwer Academic Publishers
- Franklin, J. L., & Theall, M. (1990). Communicating ratings results to decision makers: Design for good practice. In M. Theall & J. Franklin (Eds.) Student ratings of instruction: Issues for improving practice. *New Directions for Teaching and Learning #43*. San Francisco: Jossey Bass.
- Franklin, J. & Theall, M. (1989) Who reads ratings: knowledge, attitudes, and practices of users of student ratings of instruction. Paper presented at the 70th annual meeting of the American Educational Research Association. San Francisco: March 31.
ERIC # ED 306 241

40

ACADEMIC - IMPRESSIONS

Basic References

- Marsh, H. W. (2007) Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and usefulness. In R. P. Perry & J. C. Smart (Eds.) *The scholarship of teaching and learning in higher education: an evidence-based perspective*. New York: Springer.
- Perry, R. P. & Smart, J. C. (Eds.) (2007) *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht, The Netherlands: Springer
- Scriven, M. (1967) "Methodology of evaluation." In R. Tyler, R. Gagne, & M. Scriven (Eds.) *Perspectives of curriculum evaluation*. Chicago: Rand McNally & Company.
- Scriven, M. (1994) Duties of the teacher. *Journal of Personnel Evaluation in Education* 8, 151-184.

41

ACADEMIC - IMPRESSIONS

Basic References

- Theall, M., Abrami, P. A. & Mets, L. (2001) (Eds.), "The student ratings debate. Are they valid? How can we best use them?" *New Directions for Institutional Research No. 109*. San Francisco: Jossey Bass.
- Theall, M. & Arreola, R. A. (2006) "The Meta- Profession of teaching." *NEA Higher Education Advocate*, 22 (5), 5-8, June.
- Theall, M., & Franklin, J. L. (Eds.) (1990). Student ratings of instruction: Issues for improving practice. *New Directions for Teaching and Learning #43*. San Francisco: Jossey Bass.
- Other New Directions volumes on evaluation/ratings issues:
N. D. Teaching & Learning #s: 48, 83, 87, 96;
N. D. Institutional Research #: 114

42