# Brief on Student Evaluations of Teaching

**Prepared for the TLARC Evaluation Working Group** **December 14, 2016**
**Carolyn Hoessler and Nancy Turner**

A key finding of Phase I of the Teaching Quality Framework (TQF) project was for reviewing the teaching evaluation methods currently employed at the institution. Progressing toward selecting a new student evaluation of teaching (SET) instrument was an identified priority as our current tool (SEEQ) was noted to not align with some of our institutional values and priorities. This brief provides a summary of relevant literature as a first step towards selecting a new SET tool.

To provide a basis for discussing potential tools, this brief summarizes literature on SETs in relation to 3 elements that, together, determine what SETs measure, namely: (1) items, (2) scales, and (3) processes; and includes recommendations arising from the review. Further details about SETs can be found in comprehensive published summaries (for example, see Benton & Cashin, 2012, and Gravestock & Gregor-Greenleaf, 2008).

Before proceeding to consider what student ratings measure, it is important to note that SETs can provide both **developmental feedback** for improving teaching and **measurement** of the quality of teaching (Marzano, 2012). The emerging TQF has been positioned as foregrounding development of teaching. This counters typical use of SETs that often privilege evaluative[1] uses. As a result, some have gone so far as to suggest separate developmental[2] and evaluative instruments (Gravestock & Gregor-Greenleaf, 2008). The TQF principle of foregrounding development, while leaving open their use for evaluative purposes, will be important throughout the SET selection process and leveraged to inform decisions on items, scale and process.

## What Do Student Ratings Measure?

SETs are measurement tools that seek to gauge the quality of teaching for student learning by having students rate specific qualities of their instructors, courses and experiences. To do so, SETs are designed as "a substitute or proxy for direct measurement of student learning… by assessing teacher or course characteristics that are:

- ✓ believed to contribute to student learning, based on evidence or logical argument;
- ✓ observable by students;
- ✓ widely applicable, and thus can be used in many different courses; and,
- ✓ under the control of the instructor, and thus are justifiable for use in faculty personnel decisions on salary, promotion and tenure." (Murray, 2005, p. 2)

The items chosen, the rating scale used, and the process employed in distributing and gathering student responses all impact the effectiveness and focus of SETs. By making informed choices about each of these elements in line with institutional considerations of

[1] Tools used primarily for evaluative purposes are called **summative**
[2] Tools used primarily for developmental purposes are called **formative**

teaching quality, we can leverage SETs to best effect in supporting quality teaching at the institution. Literature on each of these elements is summarized in turn below.

## What SETs measure = Items + Scale + Process

## Items: Asking about specific elements of teaching & courses

*"Validity and utility depends strongly on the ability for institutions to identify questions that reflect the goals and practice of teaching in their institution"* (p.22, Gravestock & Gregor-Greenleaf, 2008)

### What can students meaningfully tell us?

Marsh (1987) and more recent researchers argue for measuring multiple aspects of teaching quality (for example, see Gravestock & Gregor-Greenleaf, 2008, section 4: Reliability, Validity and Interpretation of Course Evaluation Data). Of the potential criteria, the literature recommends students rate only some of these aspects (Benton & Cashin, 2012; Felder, 2004; Richmond et al., 2014); suggesting two key questions:

1) **Which aspects of teaching quality are meaningful to assess?** and;

2) **Of these, which aspects should students rate?**

Building on Richmond et al.'s (2014) facets of teaching quality, the table below summarizes the literature regarding which indicators of teaching quality SETs and colleagues should evaluate. Criteria are from Richmond et al. unless otherwise footnoted.

Students can only adequately rate what is observable or experienced by them.

**Table: Teaching Quality Criteria and who can/should evaluate them**

| Aspects | SETs recommended to rate: | Recommend colleagues evaluate: |
|---|---|---|
| Instructor Training | | ✓ subject knowledge<br>✓ pedagogical knowledge<br>✓ continuing education in pedagogical knowledge |
| Instructional methods and related interactions | ✓ Teaching Skills *(i.e., effective communication, preparation, listening, respectfulness, technology competent)*<br>✓ Instructor's teaching behaviors & actions[2]<br>✓ Classroom instruction[1]<br>✓ Out of class interactions *(availability and helpfulness)*[1]<br>✓ Advising and mentoring[1] | ✓ Teaching skills – via observation<br>✓ Pedagogy *(i.e., effectively employs instructional methods)*<br>✓ Classroom instruction[1]<br>✓ |

| Aspects | SETs recommended to rate: | Recommend colleagues evaluate: |
|---|---|---|
| Assessment process | ✓ Evaluation directness *(aligns assessment of student learning with learning objectives)*<br>✓ Evaluation utility *(provides constructive feedback).*<br>✓ Assessment tools and methods[1]<br>✓ Value seen in assessment[3]<br>✓ *Clarity of instructions*[3] | ✓ Student learning goals and objectives (*outcomes*[2])<br>✓ Assessment of student learning outcomes; assessment materials including assignments, tests and grades[2]<br>✓ Reflection on assessment (changes)<br>✓ Scholarship of teaching & learning<br>✓ Evaluation directness *(aligns assessment of student learning with learning objectives)*<br>✓ Evaluation utility *(provides constructive feedback).* |
| Syllabi | | ✓ Course transparency *(i.e., provides clear and complete information about the course in the syllabus)*<br>✓ Course planning *(i.e., demonstrates intentional selection of activities, evaluations, and assignments to achieve course goals)* |
| Content | | Content contains sufficient and relevant:<br>✓ Disciplinary knowledge base and application<br>✓ Development of broader skills including critical thinking, information literacy, collaboration and speaking<br>✓ Values in discipline |
| Instructor reflection and continuous improvement | Note: SETs are the basis of instructor reflection in this facet. | ✓ Student feedback *(i.e., solicits formative and summative feedback from students on teaching effectiveness) – includes having summaries of student evaluations*<br>✓ Reflection on student feedback *(i.e., utilizes formative and summative student evaluations of teaching to improve teaching and learning) – changes in teaching materials and methods* |
| Student learning/ experience | ✓ Amount they have learned[2]<br>✓ Difficulty of learning experience[2]<br>✓ Workload[2]<br>✓ Changed motivation toward the subject matter[2] | |

[1]Felder, 2004 [2] Gravestock & Gregor-Greenleaf, 2008 [3]Theall & Franklin, 2001

Richmond et al. (2014) identified specific sources of evidence to be considered (e.g., records of continuing education for subject knowledge, sample assessment methods and results for assessment processes) along with references to relevant research for each aspect.

## Assessing Active learning

SETs are often critiqued as poor measures of active learning. This disconnect can result, in part, from students' interpretations of teaching and learning. Most students interpret good teaching as what the instructor does rather than "their own active role in learning" (Parpala, Lindblom- Ylänne and Rytkönen (2011, p. 559). To improve interpretation, validity and applicability of SETs for active learning, item wording can explicitly direct students to consider their own active involvement in learning.

### Measuring the course or the instructor?

It is important to distinguish between items that are most related to a course, and thus vary with the course or course type (introductory courses, mandatory courses) and those that assess the behavior of the instructor across all types of courses. When collecting and reviewing course evaluations, looking for patterns across courses and the contextual factors of the course can be valuable in providing clarity on the impact of the nature of the course from the actions of the instructor. There are initiatives, including at Simon Fraser University, to provide aggregate SET scores from comparable courses to provide context for interpreting an instructor's ratings. This would be useful for contextualizing SETs scores, as recommended by Gravestock and Gregor-Greenleaf, 2008, particularly when used for formative purposes to support the development of teaching practice.

### Mix and Match Option

Rather than a single form, SETs can include a short set of standard questions and modules that departments or faculty select based on the nature of the course or focus of the evaluations (SFU, 2013). This may allow a tailoring of evaluations that addresses some of the noted issues of fit.

## Bias in Human Judgment

The notion that humans are not machines is hardly novel, yet concern about objectivity arises each time we look at measures of perceived quality. When a human perceives their world, they are shaped by their prior experiences, expectations, and inherent shortcuts. Monotone voices are perceived as slower, less interesting and less informational, in the same way that we are socialized to identify objects with specific characteristics as chairs. The concern about the potential for bias and thus validity of student ratings increases when student ratings are used for summative measurement purposes. The validity and reliability of ratings can be determined through several approaches included in Benson's and Cashin's (2012) detailed summary of recommended approaches to ensuing measurement quality in student ratings including:

- ✓ assessing correlations of student ratings with alumni ratings with moderate or higher correlations being an indicator of validity,
- ✓ determining construct validity by, for example, gathering trained observers and students' written comments to assess in relation to student ratings, and
- ✓ analyzing the interrelationships between a SET's responses to confirm that SET items represent distinct aspects of teaching and course quality to access reliability and validity.

In addition to these approaches, SETs can benefit from strategies for measurement development such as confirming interpretation using focus groups or think aloud approaches.

The discussion about bias in SETs has also been framed by concerns that particular students may approach evaluations differently or that irrelevant instructor actions might unduly influence scores. Benton and Cashin's (2012) IDEAS paper provides a nuanced summary of the instructor, student and course characteristics that have been found to more likely influence or not influence student ratings. Several characteristics are briefly noted here. See their paper for further descriptions of each variable and the relevant cited research.

**Characteristics found to be unrelated, and thus unlikely to bias SETs, included**:

- Course time of day
- Student GPA,
- Instructor age, and research productivity.

**Characteristics found to be at least weakly related, thus more likely to bias SETs, included**:

| Related to increased SETs | Related to decreased SETs | Mixed results |
|---|---|---|
| • Higher instructor rank/position<br>• More instructor expressiveness<br>• Greater prior student interest in the subject matter<br>• Higher course level, especially graduate courses | • Larger class size | • Instructor's gender (interacting with student gender and other factors)<br>• Students' anticipated grade |

*See Appendix 1 for Simon Fraser University' one-page summary of factors.*

General trends indicate some student, instructor and course characteristics are statistically related (and predictive) of SET scores, however these variables can interact to create a more complicated influence on SET's response patterns, even for the assumed to be clear link between SETs and grade expectations (See "Bias, It's complicated" box).

**Bias, It's complicated** One of the challenges with pinpointing biases is that their effects are contextually dependent. Take for example, the effects of anticipated grades on SETs. While anticipated grades have been found to predict ratings, **students' expectations** of difficulty were more predictive. Students who saw the class as more difficult than expected provided lower SET ratings (Addison, Best & Warrington, 2006). An easy high grade, in comparison, made no difference in course evaluation scores, other variables considered (Addison et al., 2006).

Variations also occur **across disciplines** (e.g., political studies showed no effect of grade expectation on student ratings in Boring, Ottoboni & Stark, 2016 though the effect did exist in other disciplines).

Effects of potential bias depend on interaction effects like those in the box above, and on the SET item wording. Each SET item can be affected differently by these characteristics, with some items being less susceptible to bias. Focusing on items that are most appropriate for students to rate and avoiding items that are overly related to irrelevant course, student, and instructor characteristics could reduce the effect of bias.

## Scales: Defining "Good"

*The choice of scale dictates both*
*the focus of the evaluation, and the items that can be rated.*

The SEEQ uses a *Strongly Disagree, Disagree, Neutral, Agree,* and *Strongly Agree* scale. The items thus must be statements that students agree or disagree with (e.g., "I have learned something which I consider valuable." or "Instructor was dynamic and energetic in conducting the course.").

In contrast, Marzano, Frontier and Livingston (2011) argue that evaluations focused on instructor's development should use a scale reflecting levels of development, including *not using*, *beginning* (uses strategy with some errors), *developing* ("without significant error and with relative fluency"; Marzano, 2012), *applying*, and *innovating*. The items are specific strategies or tasks within the classroom (e.g., "Providing clear learning goals and scales to measure these goals.").

A third approach is a likert-scale with frequency of behavior (e.g., how frequently are students invited or encouraged to speak in class: not at all, rarely (1-2 times an hour), occasionally (2-3), often (3-5) or frequently (>5)) or teaching checklists of instructor behaviours or syllabi completeness.

## Process: SET collection, communication & response rate

The data collection process shapes SET response rates and quality of responses by providing clear instructions and communicating to students the meaning and importance of the evaluations.

### SRI data collection

- **Instructions** to students are clear, and items pretested to ensure reliably accurate interpretation.
- **Dedicating class time** increases participation when class time is provided to complete surveys (Nevo, McClean & Nevo, 2010), including by conveying the importance and value of these evaluations.

### Institutional Endorsement

- Participation is increased when university leaders and instructors indicate the value of ratings and students see SETs as important, including when academic leaders promote SETs (SFU, 2013).

### Instructor invitation and earlier use of student feedback

- Participation increases when instructors invite and encourage completion (see Crews & Curtis, 2011, list of strategies).
- Seeking student feedback earlier in the course through midterm evaluations and making changes, increases participation rates in final course evaluations (Davis, 2009).

## Response rates & Online Evaluations

There are two defining features of a good response rate statistically:

1. **Representative population** (no response bias) where those who complete SETs represent the distribution of the class.

2. **Sufficient proportion** of the population to try to reduce error in interpreting from a sample to the population, or in the case of SETs, interpreting what the full class would rate based on the SETs completed.

Consider, for example, online course evaluations, which have lower participation rates than paper-based evaluations (Gravestock & Gregor-Greenleaf, 2008). Online response rates range from 30% to 53% (Nowell, Lewis & Handley, 2010), which is 20-30% lower than paper-based. However, instructor scores are equivalent (according to peer-review research summarized by Dawn, 2008 in UBC 2010 report), and students are also more descriptive in online surveys (Kelly, 2012).

Looking specifically at one institution, a very recent study at the University of Ottawa (Groen & Herry, manuscript in preparation) where the online surveys are just being introduced did find significantly lower response rate with online delivery (51% instead of 63%). In the same study there was no significant difference in evaluation scores for 10,417 students in 318 courses when comparing online (score of 4.0) and paper-based (scores of 3.9). Similar findings have been found in other studies (e.g. Burton, Civatono & Steiner-Grossman, 2012).

## Recommended response rates

Recommended response rates vary based on class size and desired accuracy (confidence level and probability of consistent rating). For example, a response rate of 15% to 25% is recommended for a course with 200 students, while 40% to 53% is recommended for a course with 30 students. A summary of recommended response rates based on an 80% confidence interval are included in Appendix 2.

# Supplementing SETs with additional data collection

Wright (2006) recommended evaluations of teaching include additional data collection to:

- **Allow for follow-up** for very high or low evaluations and completion of correlations with grades to investigate concerns of faculty that student evaluations are related to students' grades. This would require student evaluations to be classed as confidential but not anonymous (as is currently the case with the SEEQ tool at the U of S) and for instructors to be involved and supported (to ensure confidentiality) in use of results in this way.

- **Complement student evaluations** for untenured faculty (or others by request) with in-depth interviews of low-, medium- and high- raters for all students for small classes or a sample of students in large classes. This supplementary data collection could address concerns and "protect faculty members who are very demanding in the classroom, but skilled teachers" (p. 421). Implementation could, according to Wright, vary depending on the number of years of good teaching already established.

# Recommendations

To create meaning and accurate SETs as part of course evaluations and teaching quality at the U of S, the following recommendations, based on the reported literature, are proposed for consideration by the Vice-Provost, Teaching and Learning.

## Formative and Flexible

1. Design/select a SET tool that:
    a) includes a short set of standard questions as well as modules that departments or instructors can select based on the nature of the course or focus of the evaluations
    b) has an instructor-friendly interface
    c) allows for addition of questions
    d) allows for creation and production of easily understood reports, and
    e) allows for multiple displays of results (tables or graphs; standard deviations) (see Theall & Franklin, 2001).
2. Clearly delineate developmental and evaluative (end of course) evaluations, and keep formative/developmental evaluations confidential and distinct (Theall & Franklin, 2001).
3. Provide instructors with a simple system to deliver developmental SETs to students during the course with questions that align with later evaluative (end of course) evaluations.

## One Piece of Richer Picture

4. Develop guidelines/support for U of S review committees on interpreting teaching evaluations as part of a holistic assessment of quality teaching.
5. Supplement SETs with other forms of student evaluations (possibly including focus groups and assessment of students' work) particularly for those where evaluations have the greatest impact (e.g. pre-tenure faculty, teaching award applicants; Wright, 2006).
6. Complement SETs with self and peer-ratings and peer assessment of artifacts (see Richmond et al.'s, 2014, for example sources of evidence).

## Grounded in Shared Understanding of Teaching Quality

7. Establish systems that allow individual SET results to be easily contextualized with aggregate results from comparable courses and faculty (e.g.; other required or large courses; Benton & Ryallis, 2016; Simon Fraser University has an example system). This would need to be implemented in conjunction with guidelines/support (#4).
8. As part of SET reports, include descriptive information about the course (e.g., is it elective or required) to allow for interpretation to be done in consideration of context (Theall & Franklin, 2001). This also should be implemented with guidelines/support (#4).
9. Consider the unique contribution SETs make to the institutional approach to evidencing/evaluating aspects of teaching quality. SETs should be included explicitly within the teaching quality framework (TQF) ensuring appropriate positioning as one of several indicators of teaching quality.
10. Communicate the importance, purpose and uses of SETs in the context of the TQF to stakeholders (students, instructors, academic leaders; Theall & Franklin, 2001).

## Appropriate Items, Scales and Process

11. Select items that are appropriate and feasible for students to provide input on with attention to the items that improve SET validity and are least likely to be biased and/or

least likely to use language that increases the likelihood of biased response (see Benton & Cashin, 2012; Benton & Ryallis, 2016; SFU summary in Appendix 1).

12. Use scales that gather the information appropriate for developmental and evaluative (end of course) versions of SETs (as recommended by Marzano, Frontier, & Livingston, 2011)
13. Encourage good response rates and show value placed on student feedback by:
    a) Encouraging a consistent practice of providing class time for completion of SETs and ensuring that students have access to an electronic platform that can be readily accessed in class to complete the evaluations.
    b) Encouraging instructors to communicate about changes they have made in the past based on feedback (Benton & Ryallis, 2016).
14. Consider item wording to allow students to rate their own active involvement in learning (particularly relevant for courses that employ active learning strategies).

## References for Cited Literature


Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal*, *40*(2), 409 - 416.

Benton, S. L. & Cashin, W. E. (2012). *Student Ratings of Teaching: A Summary of Research and Literature*. IDEA Paper #50. Manhattan: KA. http://ideaedu.org/wp-content/uploads/2014/11/idea-paper_50.pdf

Benton, S. L. & Ryalls, K. R. (2016). *Challenging Misconceptions About Student Ratings of Instruction*. IDEA Paper #58. Manhattan: KA. www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_58.pdf

Boring, A., Ottoboni, K., & Stark, P.B. (2016) Student evaluations of teaching (mostly do not measure teaching effectiveness

Burton, W., Civitano, A., & Steiner-Grossman, P. (2012). Online versus paper evaluations: differences in both quantitative and qualitative data. *Journal of Computing in Higher Education*, *24*(1), 58-69.

Crews, T. B., & Curtis, D. F. (2011). Online course evaluations: faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education, 36*(7), 865–878.

Davis, B. G. (2009). *Tools for teaching*. 2nd ed. San Francisco, CA: Jossey-Bass.

Felder, R. M. (2004). How to evaluate teaching. *Chemical Engineering Education, 38*(3), 200-202.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models and Trends*. Higher Education Quality Council of Ontario: Toronto, ON. www.heqco.ca/SiteCollectionDocuments/Student Course Evaluations.pdf

Groen, J., & Herry, Y. (manuscript in preparation). *The online evaluation of courses: Impact on participation rates and evaluation scores*. University of Ottawa.

Kelly, M (2012). *Student evaluations of teaching effectiveness: Considerations for Ontario universities.* Council of Ontario Universities report No. 866. http://cou.on.ca/papers/student-evaluations-of-teaching-effectiveness/

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research , 11*(3), 253-388.

Marzano, R. J. (2012). The Two Purposes of Teacher Evaluation. *Educational Leadership,* 70(3), 14-19. www.ascd.org/publications/educational-leadership/nov12/vol70/num03/The-Two-Purposes-of-Teacher-Evaluation.aspx

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: ASCD. www.ascd.org/publications/books/110019.aspx

Murray, H. G. (2005). Student evaluation of teaching: Has it made a difference? Paper presented at the Annual Meeting of the Society for Teaching and Learning in Higher Education. www.stlhe.ca/wp-content/uploads/2011/07/Student-Evaluation-of-Teaching1.pdf

Nevo, D., McClean, R., & Nevo, S. (2010). Harnessing Information Technology to Improve the Process of Students' Evaluations of Teaching: An Exploration of Students' Critical Success Factors of Online Evaluations. *Journal of Information Systems Education, 21*(1), 99-109.

Nowell, C., Lewis R. G., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education*, *35*(4), 463–475.

Parpala, A., S. Lindblom- Yläänne and H. Rytköönen (2011): Students conceptions of good teaching in three different disciplines. *Assessment & Evaluation in Higher Education 36*(5), 549-563.

Richmond, A. S., Boysen, G. A., Gurung, R. A. R., Tazeau, Y. N., Meyers, S. A., & Sciutton, M. J. (2014). Aspirational model of teaching criteria for psychology. *Teaching of Psychology, 41*(4), 281-295.

Simon Fraser University. (2013). Student evaluation of teaching and courses: The teaching and courses evaluation project final report. Retrieved from www.sfu.ca/content/dam/sfu/teachingandcourseeval/documents/TCEP Final Report 1.7.pdf

Smith, M. K., Jones, F. H. M., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a New Instrument to Characterize University STEM Classroom Practices. *CBE-Life Sciences Education*, *12*(4), 618 - 627. Available at: http://www.cwsei.ubc.ca/resources/COPUS.htm

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research, 109* , 45-56.

University of British Columbia (April 15, 2010). *Student evaluation of teaching: Response rate*s. Report prepared for SEoT Committee.

Wright, R.E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal*, *40*(2), 417-422.

Zumrawi, A. A., Bates, S. P., & Schroeder, M. (2014). What response rates are needed to make reliable inferences from student evaluations of teaching?, *Educational Research and Evaluation, 20*(7-8), 557-563.

## Appendix 1 – SFU Quick reference guide[3]

# Student Evaluation of Teaching and Courses (SETC) – Summer 2015
# Quick reference guide: Factors that may influence student ratings

*This document is a very brief summary of factors that may influence student evaluation of teaching and courses, as presented in Student Ratings of Teaching: A Summary of Research and Literature (Stephen L. Benton and William E. Cashin, The Idea Centre, 2012). These factors are sometimes popularly classified as "bias," but Benton and Cashin point out difficulties with definitions of bias. They suggest that a better way to approach variables that show a correlation with student ratings is to "distinguish between variables … that possibly require control versus those that do not require control, especially when making personnel decisions." The table below is intended to provide context for the interpretation of results in the summer and fall 2015 pilots of SFU's new online course evaluation system. It should be noted that once sufficient SFU data is available, the SETC project team plans to prepare a reference document based on student evaluation data from the SFU context.*

| Factor | Correlation | Possible impact |
|---|---|---|
| **Instructor-related variables** | | |
| Faculty rank | Stronger | Regular faculty members tend to receive higher ratings than graduate teaching assistants |
| Expressiveness (related to presentation style) | Stronger | Making the class interesting as well as informative can foster student attention and thereby enhance learning |
| Personal characteristics | Average | Research is mixed; positive correlations with self-esteem, energy and enthusiasm, neatness, organization and value placed on approval |
| Age and teaching experience | Weaker | Inconclusive; older faculty and first-year instructors tend to receive lower ratings, but reasons are unclear |
| Gender | Weaker | Research is mixed; female students may rate female instructors higher and male students may rate male instructors higher |
| Research productivity | Very weak | Little correlation found with student ratings |
| Race | None | Limited studies; no differences found |
| **Student-related variables** | | |
| Student motivation | Stronger | Instructors are more likely to receive higher ratings from students with a prior interest in the subject matter |
| Expected grades | Low | Research is mixed; low positive correlation found between student ratings and expected grades |
| Gender | Weaker | No consistent gender effect; however, some gender preferences found, particularly female students for female instructors |
| Level (year) | Weaker | Little practical effect found on ratings |
| GPA | Weaker | Little or no relationship found between student ratings and GPA |
| Age | None | Studies suggest age of student has little effect on student ratings |
| Personality | None | No meaningful relationships found |
| **Course and administrative variables** | | |
| Course level | Weaker | Higher-level courses tend to be rated higher than lower-level courses |
| Class size | Weak | Weak inverse relationship between ratings and class size |
| Academic discipline | Stronger | Humanities/arts courses rated higher than social science courses, which in turn are rated higher than math/science courses; not clear why |
| Workload/difficulty | Weaker | Students tend to give somewhat higher ratings to difficult courses that require hard work, but differences are not large |
| Time of day of course | None | No meaningful influence found |
| Timing of data collection | None | No impact from time during the term when ratings are collected |

## Appendix 2 – Recommended response rates

[3] Note that the associations noted in the column titled 'correlation' are actually descriptions of relationships found in the literature between the factor and SET outcomes, not statistical correlations.

Recommended response rates vary as class sizes, confidence intervals and the probability of a favourable instructor rating are taken into account. Zumrawi, Bates & Schroeder (2014)'s detailed analysis of UBC data indicates that for some items it reasonable to assume a probability of favourable ratings of 0.7 to 0.8. The desired response rates for 0.7 (70% favourable) are included in the table below.

| Class Size | McGill: Acceptable Response Rate %[1] | Nulty (2008): Recommended response rates with 80% Confidence Interval (10% Sampling Error)[1] | Zumrawi, Bates & Schroeder (2014): determined desirable response rates for 80% confidence interval (10% margin of error), 70% favourable rating[2] |
|---|---|---|---|
| 5-11 | minimum 5 responses | at least 75% | 78% |
| 12-30 | at least 40% | 74 – 48% | 53 – 78% |
| 31-100 | at least 35% | 47 – 21% | 26 – 53% |
| 101-200 | at least 30% | 20 – 12% | 15% – 26% |
| 201-1000 | at least 25% | 11 – 3% | 3%– 15% |

([1]McGill and Nulty columns are from UBC, 2010, adapted from Rawn, 2008)
([2] Zumrawi, Bates & Schroeder, 2014, recommendations are categorized to fit with class-sizes in column 1; this article compares desired response rates for 0.7, 0.8, and 0.9 probability of positive response – "0.8 would mean that, overall, 80% of the students in the institution rate their instructors favourably" p. 560)