DATE: February 13, 2016

TO: James Holloway, Vice Provost for Global and Engaged Education

SUBJECT: Interim Report of the Course Evaluation Instrument Review Committee

The Committee's Charge

The Course Evaluation Instrument Review Committee (hereafter "Committee") was established in December 2015 with the charge of recommending a small set of closed-ended survey questions to be used on course evaluation instruments across the university for the purpose of collecting data on student ratings of teaching (SRT). The Committee was asked to explore the existing literature on best practices in SRT measurement, to consult with campus stakeholders regarding the purposes for which SRT data are used at the University of Michigan (hereafter "University"), and to examine the SRT questions that are used most frequently across the University.

The Committee has been asked to provide an interim report by mid-February 2016, focusing specifically on recommending a small set of SRT questions to collect information intended to be useful to students, in the event that the University elects to release SRT results to students. The Committee has been asked to provide a final report in April 2016, recommending a larger set of SRT questions for the broader purposes for which such data are collected at the University.

This document is the interim report of the Committee, presenting the Committee's initial findings on the question of SRT data that would be informative and helpful to students. Given the interconnectedness of the objectives of the interim and final reports, the Committee anticipates that the initial findings discussed here may evolve and, consequently, be revised to some degree in the final report.

Overview of the Committee's Process

The members of the Committee were convened by Dr. James Holloway, Vice Provost for Global and Engaged Education, on December 10, 2015, to receive their charge. The chair and several members of the Committee then met with the Vice Provosts and Associate Deans Group (VPADG) on December 11, 2015, to discuss how SRT data are used in the units across campus. Subsequent meetings of the Committee were held on January 21, January 28, February 4, and February 11, 2016.

At the initial meeting on December 10, 2015, the Committee agreed that, for the purposes of informing student decision-making about courses, it would be sensible to focus on SRT questions that address *dimensions* of teaching effectiveness, rather than *global* assessments of teaching effectiveness. Furthermore, given the short timeline for the work, the Committee agreed that it would be sensible to focus primarily on existing SRT systems that are well tested or used widely, such as the IDEA System, the Student Evaluation of Educational Quality (SEEQ) system, and the Student Instructional Report (SIR II) system. The Committee also considered the University's existing list of SRT questions, a similar system used by the University of Illinois, and selected literature on dimensions of teaching effectiveness (e.g., Abrami, d'Apollonia & Rosenfield, 1996; Murray, 2007).

To organize the Committee's initial work, the various SRT systems were reviewed in concert with selected relevant literature to distill a reasonably common set of dimensions of teaching effectiveness. That is, across SRT systems and the pertinent empirical literature, there is variability in the number of identified dimensions of teaching effectiveness, though there is considerable overlap in dimensions. Viewing these sources holistically, an initial set of seven dimensions of teaching effectiveness that captured commonalities across the sources was proposed for consideration by the Committee. These seven dimensions are listed below.

1. self-assessed learning
2. course organization
3. instructor approachability/rapport
4. instructor dynamism
5. instructor clarity
6. assessment methods
7. encouraging student interaction with instructor and/or fellow students

Committee members were asked to add to the list any additional dimensions of teaching effectiveness that they perceived to be inadequately addressed in these initial seven, and then to rank order all dimensions in terms of perceived importance and relevance for collecting data that would be informative and useful to students. In addition, members were asked to identify and document two SRT questions --- a preferred first choice and a second choice --- that they believed best addressed each dimension of teaching effectiveness, drawing on the various sources (e.g., SIR II, SEEQ, IDEA, University of Michigan, University of Illinois) and any other sources that they deemed appropriate and relevant.

Finally, the Committee noted the empirical evidence concerning differences between courses that may influence SRT results but that are not under the instructor's control, such as student motivation, perceived workload, and course size (e.g., Benton & Cashin, n.d.). The Committee recognized that it would be important to control or adjust for such differences when reporting SRT results. Information on some of these differences, such as course size, presumably can be obtained from the University's records. Information on other differences, however, would need to be collected in an SRT survey format, and two of these are listed below. Committee members were asked to select and document first- and second-choice survey questions for each of these two factors, as they did for the dimensions of teaching effectiveness.

1. student motivation
2. perceived workload/difficulty

The results of the initial ranking revealed a strong consensus on the three most important dimensions of teaching effectiveness for informing student decision-making: *self-assessed learning*, *course organization*, and *instructor clarity*. Further discussions led to two substantive changes to the list of focal dimensions. First, the dimension of *course organization* was split into a course-focused component and an instructor-focused component.[1] Second, two additional dimensions of teaching effectiveness were added: *impact on students* and *classroom climate*. The final list of six dimensions of teaching effectiveness that the Committee deemed most relevant to informing student decision-making is provided below. Added to this list are the two adjustment factors (marked with an asterisk) that the Committee considered especially important for making sense of SRT results and for which information would need to be collected in an SRT survey format.

1. self-assessed learning
2. impact on students
3. course organization (course-focused)
4. course organization (instructor-focused)
5. instructor clarity
6. classroom climate
7. student motivation*
8. perceived workload/difficulty*

Focusing on these six dimensions of teaching effectiveness and the two adjustment factors, the Committee reviewed the various sources of existing SRT survey questions and deliberated about the best question to measure each dimension. As noted earlier, the governing objective was a short list of questions that would elicit information that would be useful to students if the University elects to release SRT data to students. In this context, "short list" was defined by the Committee, with input from Vice Provost Holloway, as eight or fewer questions.

Committee members compared survey questions with respect to their perceived centrality to the particular dimension, applicability across a range of course types (e.g., undergraduate and

[1] The Committee recognized that a course may be well organized, yet may have an instructor who is perceived by students to not use class time especially well or prepare adequately for class meetings. Likewise, an instructor may have carefully organized materials for each class session, but the sequencing of topics, assignments, etc., may lead students to experience a sense of disorganization.

graduate, lecture and studio, academic and professional), transparency in wording, level of abstraction (with minimal abstraction preferred), and attention to topics that students would be able to assess knowledgeably. Possible alternative wording for preferred survey questions was discussed by the Committee, and, in some cases, a modified version of the original question was selected by the Committee.

In developing its recommendations regarding SRT survey questions, the Committee assumed that the information reported to students would continue the unit of reporting that currently is used by the University, namely the intersection of course, instructor, and term, except in cases of team-taught courses.[2] Additionally, the Committee assumed that the SRT survey questions must be closed-ended and employ a five-point response scale, although it need *not* be confined to a response scale of *strongly agree* to *strongly disagree.* The Committee assumed that SRT information would be collected at the end of a course, as is currently practiced. Finally, the Committee assumed that its recommendations would apply to instructional faculty, not to graduate student instructors.

With regard to response scales, the Committee was apprised of the fact that the University's current SRT data collection system is limited to a scale of *strongly agree* to *strongly disagree.* However, or a number of the proposed SRT questions, the Committee perceived meaningful advantages to responses scales other than *strongly agree* to *strongly disagree*. Consequently, as the University explores alternative SRT data collection systems, the Committee recommends that the University specify a system that supports variation in the response scales of SRT questions. In the interim, the Committee notes that it may be reasonable in some cases to offer a response scale in the SRT question itself, as is done in University question 891, "The workload for this course was (SA=LIGHT...SD=HEAVY)." When such a modification is not reasonable, however, the Committee's assessment is that the *strongly agree* to *strongly disagree* response scale will be adequate, though still less than ideal, for most of the SRT questions recommended in this interim report.

Recommended SRT Questions

In the table below are the Committee's recommended SRT survey questions, organized by dimension. In most cases, the Committee offers a preferred question as well as a secondary alternative question (or alternative wording of the preferred question). In nearly all such cases,

[2] More specifically, the Committee assumed that reports to students of SRT results would *not* aggregate information for two or more instructors teaching the same course, would *not* aggregate information for different courses taught by a single instructor, and would *not* aggregate information for the same course taught in different terms by a single instructor. An important exception to this generalization is the need to aggregate *course-related* SRT information, but not *instructor-related* SRT information, in a team-taught course.

the Committee found the alternative question to be a strong option but less desirable than the preferred option.

Dimension	**Preferred Question/Prompt**	**Alternative Question/Prompt**
Self-Assessed Learning	This course advanced my understanding of the subject matter. [Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree]	This course enhanced my understanding of the subject matter. [Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree]
Impact on Students	My interest in the subject has increased because of this course. [Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree]	This course has increased my interest in the subject. [Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree]
Course Organization (course focused)	I knew what was expected of me in this course. [Almost Always, Frequently, Sometimes, Occasionally, Hardly Ever]	The expectations for this course were clear. [Almost Always, Frequently, Sometimes, Occasionally, Hardly Ever]
Course Organization (instructor focused)	The instructor seemed well prepared for class meetings. [Almost Always, Frequently, Sometimes, Occasionally, Hardly Ever]	The instructor was well prepared for class meetings. [Almost Always, Frequently, Sometimes, Occasionally, Hardly Ever]
Instructor Clarity	The instructor explained material clearly. [Almost Always, Frequently, Sometimes, Occasionally, Hardly Ever]	None.
Classroom Climate	The instructor treated students with respect. [Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree]	The instructor promoted an atmosphere conducive to learning. [Almost Always, Frequently, Sometimes, Occasionally, Hardly Ever]
Student Motivation	I had a strong desire to take this course. [Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree]	None.
Perceived Workload	As compared with other courses of equal credit, the workload for this course was… [Much Lighter, Lighter, Similar, Heavier, Much Heavier]	The workload for this course was… [Very Light, Light, Moderate, Heavy, Very Heavy]

The only dimension on which the Committee was meaningfully divided concerning survey questions was *classroom climate*. On this dimension, the majority of members favored the question (or prompt), “The instructor treated students with respect,” which is listed in the table below as the Committee’s preferred question. However, concerns were raised by some members about the high level of inference necessary for students to respond to this prompt. Additional concerns were raised about the effect of cultural differences in definitions of respect (in a classroom setting) on students' responses to this prompt, particularly in classes that address controversial or emotionally-laden subjects.

The Committee also recognized that there are a number of factors that influence classroom climate, and many of these are not under the control of the instructor. Thus, the preferred question and the alternative question offered here address only one aspect of the larger picture of classroom climate.

Additional Faculty Input

This interim report recommends a significant change in the current SRT instrument. Given the potential implications of this change, the Committee further recommends that the University seek feedback from key advisory bodies, such as the VPADG, and from the broader instructional faculty. Concerning the latter, one option that the University might consider is a survey of instructional faculty. If such a survey were to be conducted, it may be useful to ask faculty to rank proposed SRT questions within each dimension, but also to allow "none of these questions" as an option within each dimension.[3] It also may be useful to offer an open-response section of the survey in which faculty would have the opportunity to provide unconstrained feedback.

Interpretability of SRT Information

The Committee discussed a number of concerns about the interpretability of SRT results, particularly for audiences that do not read or use such information routinely. Among these concerns, the Committee expressed reservations about the University's current approach of summarizing SRT data for a particular instructor or course with a score determined through median interpolation. This implies a level of measurement precision that may be unjustified. In addition, the calculation of an interpolated median is sufficiently obscure that, given the original data, many would be unable to reproduce the corresponding SRT summary scores. Using the mean of responses as an alternative summary score would be easier to understand, but it exacerbates both problematic assumptions about measurement precision and, perhaps more importantly, the influence of extreme responses (either positive or negative) on the summary score.

Straightforward alternatives that the Committee discussed include [1] the percentage of responses to a particular question that report favorable perceptions of the course/instructor (e.g., the proportion of responses that were a score of 5 or 4 on a five-point response scale) or [2] a simple median without interpolation. However, the Committee recognized that, without frequent

[3] Given that this Committee is a response to a Faculty Senate resolution, this constrained set of choices on the survey arguably is one reasonable approach to soliciting faculty input. Other approaches, however, may be equally reasonable.

education to the contrary, the need for efficient decision-making will result in the tendency for oversimplified dichotomies between “acceptable” and “unacceptable” scores to be adopted in practice, even for the most straightforward of numerical summary scores. Thus, a third possibility that the Committee discussed that may reduce this tendency toward oversimplification is to report SRT information as a histogram of responses, in place of a single numerical summary measure.

Another aspect of interpretability that the Committee discussed was the need to ensure comparability in SRT results across courses. The literature on SRT data (e.g., Benton & Cashin, n.d.) suggests that student motivation and perceived workload influence SRT results.[4] Given the evidence in this regard, the Committee recommends adjustment for these factors and other important factors, such as course size, when SRT results are reported, whether to students, instructors, or administrators. Adjustment of this sort can be accomplished in several different ways that are not mutually exclusive:

- Offering guidelines for interpreting the data that draw attention to the need to condition results on important factors that are not under the control of the instructor.
- Providing a reporting interface that allows users to make comparisons of courses that share similar characteristics (e.g., comparing SRT results for courses of the same size, similar perceived workload, and similar student motivation).
- Adjusting summary scores directly through an algorithm that draws on key adjustment factors, such as multiple regression, and then reporting adjusted scores in place of the simple summary scores.[5]

Finally, the Committee discussed the importance of ensuring a minimum number of student responses and perhaps a minimum response rate to SRT survey items when reporting SRT results. That is, SRT results based on few responses may be less informative than are results based on many responses, and, in fact, may be misleading.

[4] An analysis of Michigan’s own SRT data supports the finding about student interest, but does not provide evidence of a significant relationship between perceived workload and overall course assessment. Please see Appendix A for more information.

[5] One advantage of this approach is that SRT results can be adjusted before they are viewed by users, reducing the prior knowledge necessary to make sense of the results. An important disadvantage, however, is that it depends on assumptions about the precision of the underlying data that the Committee finds problematic. As an example of this approach, the University of Washington adjusts the median of the overall course quality item with a motivation factor, the log of course size, and student expected grade. For more information, please see https://www.washington.edu/oea/services/course_eval/uw_seattle/adjusted_medians.html

In conclusion, the Committee looks forward to receiving your thoughts on the work done to date and continuing into the second stage of our work, leading to a recommendation of a larger set of SRT questions to address the broader needs of the University for such information.

Respectfully submitted,

Peter Bahr, Associate Professor and Committee Chair, Center for the Study of Higher and Postsecondary Education
Lisa Emery, Associate Registrar, Office of the Registrar
Jason Geary, Associate Dean, School of Music, Theatre & Dance
Matthew Kaplan, Executive Director, Center for Research on Learning and Teaching
Mika LaVaque-Manty, Associate Professor, Department of Political Science
Benjamin Paloff, Assistant Professor, Slavic Languages and Comparative Literature
Jamie Phillips, Professor, Electrical Engineering and Computer Science
Anushka Sarkar, Chief Programming Officer, Central Student Government
Gundy Sweet, Clinical Professor, College of Pharmacy
James Wagner, Research Associate Professor, Institute for Social Research

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. C. Smart (Eds.), *Higher education: Handbook of theory and research* (pp. 213-264). New York: Agathon.

Benton, S. L., & Cashin, W. E. (n.d.) *Student ratings of teaching: A summary of research and literature, IDEA Paper, no. 50*. Manhattan, Kansas: IDEA. http://ideaedu.org/research-and-papers/idea-papers/idea-paper-no-50/

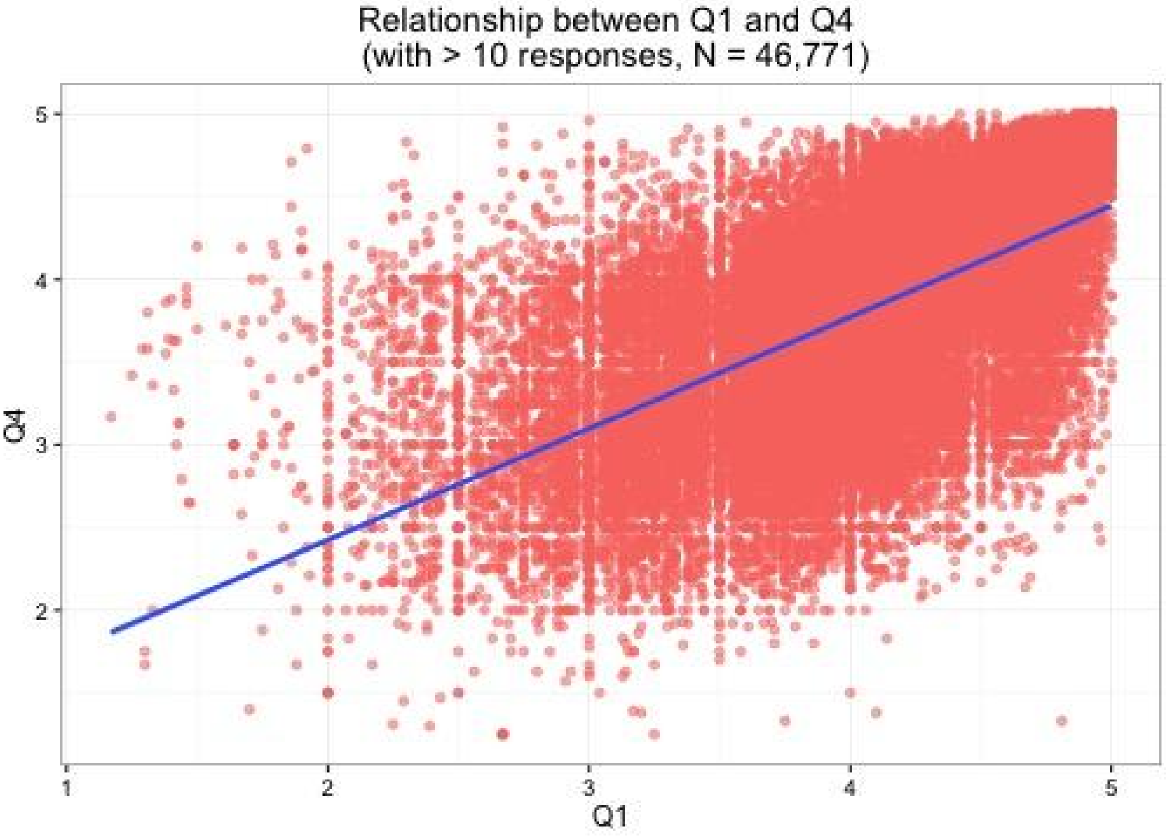
Murray, H. G. (2007). Low-inference teaching behaviors and college teaching effectiveness: Recent developments and controversies. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 145-200). New York: Springer.

Appendix A

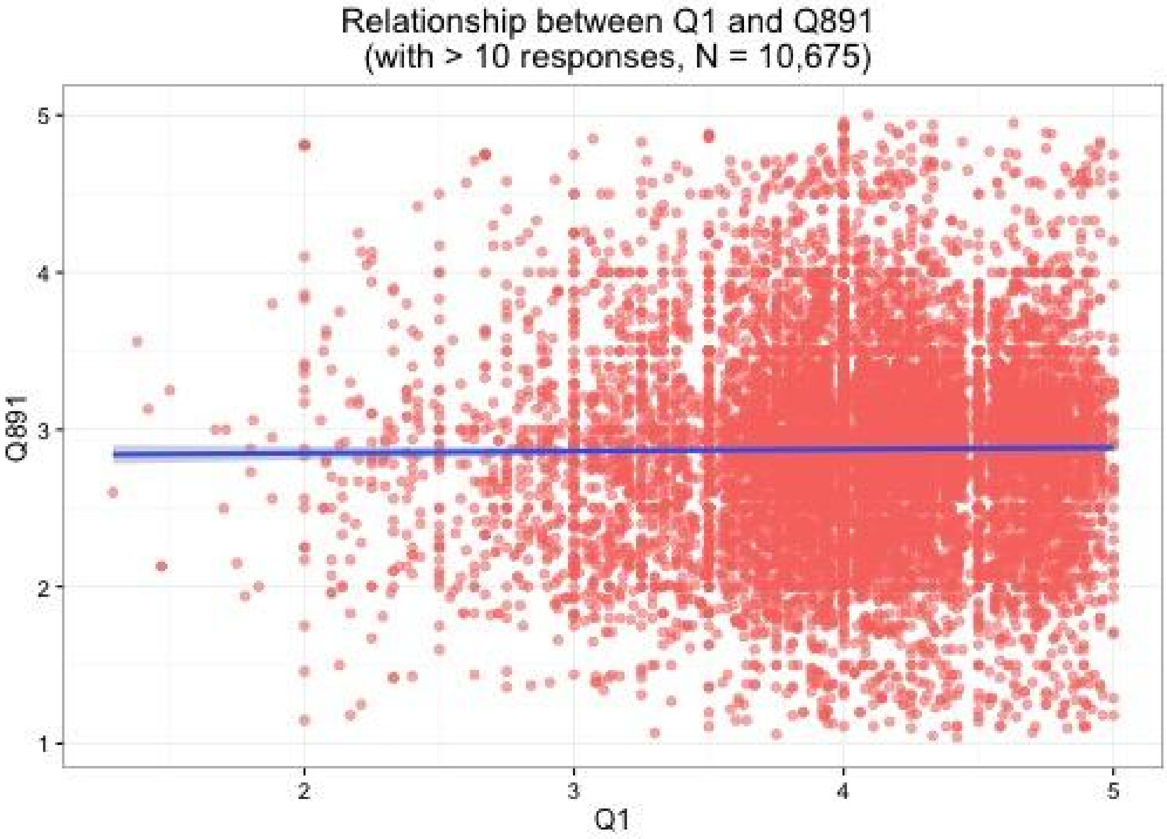
To illustrate why adjustment might be a concern, the Committee includes here two visualizations of the relationships between items on the Michigan SRT. Both of the relationships are frequently believed to be correlated; the illustrations below suggest that one of these beliefs is warranted, while the other is not.

Using data obtained from the Registrar's Office, Mika LaVaque-Manty has conducted exploratory analyses of Michigan SRT data. The data were limited to evaluations from the College of LSA and the College of Engineering for the period from Winter 2005 to Winter 2013. The entire dataset contains 108,000 course-terms (including labs and discussion sections). The results reported here are limited to data for courses for which more than 10 student responses were submitted.

There is a strong association between a student's interest (measured in Michigan's SRT by Q4, "I had a strong desire to take this course") and the student's overall rating of the course (Q1, "Overall, this was an excellent course"). The correlation coefficient between Q1 and Q4 is 0.60; the following plot demonstrates this graphically. Each dot is the average rating median in an iteration of a course, section, or lab.



However, at least in LSA and COE, there does not appear to be a relationship between students' overall rating of a course and their perception of the workload, measured by Q891, "The workload in this course was…," with a response scale of *Light* (1) to *Heavy* (5). The overall number of observations is lower because many courses do not collect the latter item, but the number of observations still is quite high, and the absence of association likely is meaningful.


Relationship between Q1 and Q891
(with > 10 responses, N = 10,675)
Q891
Q1