Extracted from:

Graniero, P.A., Hamilton, B. and Cramer, K. (2014). SRI data aggregation and visualization: An evaluation of potential uses. In A. Wright and B. Hamilton (Eds.), The Ontario Universities' Teaching Evaluation Toolkit: Feasibility Study (227-283). Ontario Ministry of Training, Colleges and Universities.

The full report is available at: http://ctl.uwindsor.ca/provincial-innovation-fund-reports/

ETHICAL PRINCIPLES IN DATA AGGREGATION

There are some obvious potential risks involved in the wholesale adoption of aggregate SRI data analysis, risks that are in essence a magnified version of the many challenges and tensions involved in all SRI activities. Firstly, the data involved are highly sensitive, and impact people's professional lives in significant ways. Just because it is possible to calculate, for example "the ten worst SRI scores on campus" does not make it constructive, statistically valid, or ethical. Further, the creation of tools to facilitate comparisons must be accompanied by mechanisms that guide people towards effective and appropriate data use, and limit the capacity for misinterpretation, bias, and the drawing of inappropriate conclusions. Finally, different institutions have significantly different understandings of how SRI data can and should be used, as reflected by the variability in collective agreements and policy guidelines across the province: these must be taken into account in considering how to approach data aggregation and the uses to which it can, and cannot, be put. These variations mean that in practical terms, tools built for actual institutional use must allow for customization at the institutional level.

It is clear that there are ethical issues to be addressed in the development of aggregate data tools, but a fully articulated set of guidelines is well beyond the scope of this study. We must find ways to establish practices that are collegial, appropriate, respectful, and beneficial. Some preliminary fundamentals are:

- practices must be in keeping with the ethical principles of the University as well as all policies and collective agreements;
- practices should be consistent with the stated purposes for which data has been gathered;
- practices must describe, clarify, and emphasize the limitations of data and tools, and limit user capacity to draw invalid conclusions where possible;
- practices must be respectful of instructors as central agents in teaching, and in many cases, as the owners of the data;
- drawing comparisons among and ranking individuals should be discouraged without extremely good reason;
- practices must be based on classification of data in terms of access rights and ability to drill down, and must also respect the need for confidentiality;
- data used in the aggregate should be anonymized, and under no circumstances should it be possible to disaggregate data in ways that make the identification of individuals possible; and
- practices must be in accordance with the Freedom of Information and Protection of Privacy Act.

There is a substantial and evolving body of literature in cognate fields such as business intelligence, health analytics, and learning analytics that could be drawn upon for the further development of appropriate guidelines. Slade and Pinsloo (2013), for example, identify the following in a discussion of the ethics of learning analytics (the large-scale use of student data for predictive purposes):

- learning analytics is a moral practice which should focus not only on what is effective, but on what is morally necessary;
- learning analytics should engage students as collaborators, co-interpreters and agents, rather than as mere recipients of interventions;
- data should be understood as a snapshot view at a particular time and place, and identity and performance should be understood as dynamic and changing;
- student success is complex and multi-dimensional. Data are incomplete and analyses vulnerable to misinterpretation and bias;
- there should be transparency regarding the purposes for which data will be used, who will have access to data, the conditions under which data will be used, and how and under what conditions privacy will be protected; and
- higher education cannot afford not to use these data (p. 12-13).

These kinds of principles appear to resonate well with the possible concerns that might arise in pursuing aggregate SRI data analysis.

It is one thing to build tools founded on and intended to promote ethical, methodologically sound data use: it is another for stakeholders to put faith in them. In general, aggregate data analysis is most likely to be effectively integrated into institutional practice if its use is of value to faculty members in pursuing their own goals and needs, and if their rights are protected through careful, consultative and incremental development of approaches to data use (Alderman & Melanie, 2012; Joughin & Winer, 2014). A process where tools are designed with faculty and administrators, and with sustained and proactive consultative processes with faculty associations (Alderman, in conversation, June 18, 2014) is more likely to produce a system that is sustainable, uncontroversial, and effective. Instructors, who are described by the data and also often own them, should also have opportunities to annotate the data so that contextual factors – first courses, introduction of innovative practices, team teaching, illnesses – can be introduced to support accurate interpretation of the narrative. It is impossible to predict all of the possible ways that tools like these need to be framed and delimited in advance of development and testing with real populations, so a thoughtful and responsive approach is necessary.

Committees, administrators, and faculty members are already making decisions based on data. Whether they are doing this well, with a strong understanding of what the data mean and do not mean, is a completely different question. There are two approaches to addressing this challenge: the first is education, and the second is the simplification and improvement of data and data reporting. While there is some evidence (Villascusa, Franklin, and Aleamoni, 1997 cited Hativa 2013b; Ludlow, 2007) that training can significantly improve facility with the use of statistical information for decision-making in teaching evaluation, there is also evidence that faculty members are not pre-disposed to engage with this kind of professional development (Ryan, 1997, cited Hativa, 2013). Visual representations, with clear markers of significance and limits, and tools which disallow inappropriate or insignificant disaggregations or comparisons, can provide decision makers with clearer and more compelling data to work from. The goal of better analytical and visualization tools should also be to enhance equity, accuracy, and fairness. It is possible that access to data tools such as the ones described in this report, would impact the data culture and pre-disposition towards data use at Ontario universities, a much-desired outcome.

Extracted from:

Graniero, P.A., Hamilton, B. and Cramer, K. (2014). SRI data aggregation and visualization: An evaluation of potential uses. In A. Wright and B. Hamilton (Eds.), The Ontario Universities' Teaching Evaluation Toolkit: Feasibility Study (227-283). Ontario Ministry of Training, Colleges and Universities.

The full report is available at: http://ctl.uwindsor.ca/provincial-innovation-fund-reports/

USING NUMERICAL DATA IN CONTEXTS OF UNCERTAINTY

Statistical tools for making comparisons and other analyses can be incredibly powerful for making decisions in a complex, heterogeneous world, but only if the measurements are made in ways that are accurate, sufficiently representative of the population being examined, and both mathematically and statistically compliant with the tools' calculation methods and assumptions about the world that the measurements describe. There are many inherent obstacles to designing and administering SRI instruments that meet these requirements, as has been discussed above. In summary:

- Students interpret the subjective questions and define or understand the available qualitative response categories in different ways, which introduces **measurement error**.
- Only a portion - and often too small a portion - of the class responds to the survey, so there are questions of how well the sampled group's responses represent the class as a whole, which introduces **sampling error**.
- Many statistical tests assume a random sample, but typically classroom response to an SRI constitutes a **convenience sample**. This violates assumptions about the sample used in statistical analysis.
- **Scales may involve too few items** to allow for statistical calculation with certainty.
- A variety of factors appear to impact the validity of student ratings of instruction. A number of differences between respondents and non-respondents have in fact been noted (Goyder 1987; Richardson 2005), in particular for students, in their attitudes and behaviour (Goyder 1987) and in their study behaviour and academic attainment (Astin 1970; Neilsen et al. 1978; Watkins and Hattie 1985). **Non-random samples may be biased by differences in the individual characteristics of students**, such as disciplinary differences in response patterns, gender differences in response patterns, and student year might affect the representativeness of a given sample (Hativa, 2013b). **Comparison among different courses may be affected by differences among the courses that impact ratings, such as level, class-size, or delivery mode**. Issues of bias in student ratings of instruction, however, are hotly debated, with considerable evidence on each side of the debate (Hativa, 2013b). One critical element of establishing the validity and reliability of SRI is establishing regular analysis at each institution of context-specific data to identify or disconfirm theories about bias within student responses (Winer et al., 2012; Joughin & Winer, 2014), a practice that is not widely employed.
- SRI instruments rely heavily on the use of **Likert and Likert-type items which produce ordinal-level data, but the responses are used as interval-level data even though the degree of difference between response choices is not uniform**, and is very difficult to characterize. Most of

the calculation and statistical methods we use to describe and make comparisons between distributions, rely entirely on having uniform differences between values.

- **Statistical tools must be used in ways consistent with their particular functions and requirements, a practice which is not always ensured**. Some tests, for example, assume a normal distribution of scores, which is certainly not the case for many SRI distributions. Other tests should not be applied to small populations.
- **Statistical measures are frequently provided without an indication of whether they are significant**, or whether the population was sufficient to justify the calculation involved.
- The use of **statistical measures of central tendency** (means, medians, and modes), even when valid for the level of measurement involved, **can mask important information about score distributions**: a bimodal distribution with many students at each extreme can produce the same mean score as a tight clump of scores around the middle, but these distributions have significantly different implications, both for decision making and for instructional improvement.

Essentially, in typical SRI implementation and analysis, necessary mathematical and statistical requirements of the tools are not met, which means that results are simply not as accurate as our faith in statistics tends to lead us to believe. Nor are they as accurate as the statistical measurements of accuracy tell us they are. Given these challenges, the results of statistical analyses of SRI data must be used with informed caution, and with an understanding that their results are not as precise, accurate, or certain as they are in other fields or applications where the requirements of statistical calculation are more easily met. It doesn't mean that the results are unusable and any analysis is pointless. Rather, the results should be used as sign posts to broader patterns, trends, or potential differences: persuasive, not conclusive, evidence. Unfortunately, their appearance of numerical precision can be beguiling.

While the guidance offered by statistical information can be helpful if used appropriately, other visual tools can significantly enhance our understanding of what the data and accompanying analyses tell us. For these reasons, we have adopted an approach to tool design that uses visualization methods to document the nature of the patterns within the data and to display statistical measures in a more accessible fashion. In many respects, the visualizations present a complementary view of what the statistics articulate with numbers (which in many cases are less easily comprehended). While, for example, SRI data reporting may provide the reader with a mean and standard deviation, actual study of the scatterplot those numbers represent can first of all concretize the information, and secondly offer a more nuanced representation than the standard deviation. Further, the use of visualization acts as a check for the appropriateness of the statistical shorthand. Finally, visualization is more democratic: it allows all users to reflect more effectively on their data, regardless of their level of familiarity with statistics. The visualizations afford an opportunity to provide other contextual information that simply cannot be captured in a numerical fashion. These contextual factors play an important role in uncovering and telling the teaching narrative the numbers summarize, a summary that often has limited effectiveness. The visualization tools employ fundamental statistical concepts to ensure that the visual story is compatible with the statistical story, to aid the users' understanding of statistical results and their accuracy, and to help users avoid drawing inappropriate conclusions. Future work may further explore effective ways to integrate more advanced statistical practices with visual tools for exploring aggregate data.

For a more detailed exploration of the fundamental terms, methods and requirements of statistical calculation, please see Appendix A in the original report.

REFERENCES

Alderman, L., & Melanie, L. (2012). REFRAME: A new approach to evaluation in higher education. *Studies in Learning, Evaluation, Innovation & Development, 9*(1), 33-41. from http://sleid.cqu.edu.au

Astin,A. (1970). The methodology of research on college impact (Part 1). *Sociology of Education, 43*(3), 223-254.

Goyder, J. (1987). *The silent minority: Nonrespondents on sample surveys*. Camridge: Polity Press.

Hativa, N. (2013b). *Student Ratings of Instruction: A practical approach to designing, operating, and reporting.* Charleston, SC: CreateSpace Independent Publishing Platform.

Joughin, G., & Winer, L. (2014). *Response to Teaching Evaluation Toolkit preliminary reporting*. Panel discussion presented March 5, 2014 in Toronto, Ontario.

Ludlow, L. H., (2005). A longitudinal approach to understanding course evaluations. *Practical Assessment, Research & Evaluation, 10*(1). Retrieved from http://pareonline.net/pdf/v10n1.pdf

Neilsen, H., Moos, R., & Lee, E. (1978). Response bias in follow-up studies of college students. *Research in Higher Education, 9*, 97-113.

Richardson, J.T.E. 2005. Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, 30(4), 387-415.

Ryan, K. (1997). *Making student ratings comprehensible to faculty: A review of alternative reporting approaches*. Paper presented at the 78th American Educational Research Association. Chicago.

Slade, S., & Prinsloo, P. (2013). Learning analytics: ethical issues and dilemmas. *American Behavioral Scientist, 57*(10), 1509–1528. Retrieved from http://oro.open.ac.uk/36594/2/ECE12B6B.pdf

Villascusa, Franklin & Aleamoni (1997) p20 Don't need it, cited from Hativa.

Watkins, D.A., & Hattie, J. (1985). A longitudinal study of the approach to learning of Australian tertiary students. *Human Learning, 4*, *127-142*.

Winer, L., di Genova, L, Vungoc, P., & Talsma, S. *Interpreting end-of-course evaluation results.* Montreal: Teaching and Learning Services, McGill University