

IN THE MATTER OF AN INTEREST ARBITRATION

BETWEEN:

Ryerson University

and

The Ryerson Faculty Association

(FCS & Related Issues)

Before: William Kaplan
Sole Arbitrator

Appearances

For the University: Simon Mortimer
Hicks Morley
Barristers & Solicitors

For the Association: Cathy Lace
Emma Phillips
Goldblatt Partners
Barristers & Solicitors

The matters in dispute proceeded to a hearing in Toronto on February 12 and April 3 & 10, 2018.

Background

It is fair to say that the content and use of Faculty Course Surveys (hereafter “FCS”) for employment related decisions such as promotion and tenure has been a live issue between the parties since at least 2003. There have been good faith collaborative discussions, including a joint committee, not to mention an ongoing on-line pilot project, as the parties grapple with matters of mutual concern. However, notwithstanding the allocation of time and resources, agreement could not be reached. In 2009, a grievance was filed, and hearing days were scheduled, and adjourned. There was mediation and the parties continued to attempt to work out their differences – differences that centre mainly on the use of FCS data and personal, Departmental, Faculty and University averages to assess faculty teaching effectiveness. A second grievance was filed in 2015. However, fundamental matters remained in dispute and that dispute made its way to the 2015-2016 collective bargaining round.

At that time, the parties were able to resolve everything – everything, that is, except for this issue. The Memorandum of Settlement provided: “...the parties agree that the FCS grievance and the FCS issues discussed in this round of negotiations including the Association’s Article 5 teaching assessment proposals will be placed before Arbitrator Kaplan for an interest mediation-arbitration.” The anticipated mediation was unsuccessful and the matters, therefore, proceeded to a hearing in Toronto on February 12 and April 3 & 10, 2018.

Association Submissions

For its part, and in brief, the Association takes the position that FCS averages should not be used in assessing teaching effectiveness and that this practice – the subject of

longstanding disagreement – must immediately stop. Averages provided no useful information. Making matters worse, the FCS itself was failing to achieve its stated objective of evaluating teaching effectiveness as the answers to the questions revealed nothing relevant about that. Indeed, the results were actually skewed by bias and their use quite possibly contravened the *Human Rights Code*. The Association submitted detailed interest arbitration proposals about this and ancillary issues.

University Submissions

For its part, the University submits that notwithstanding any of the identified problems with the FCS tool, it provided relevant information – information that was germane to faculty assessment. Likewise, averages were collective agreement compliant and provided a useful data point. They could not be determinative of a tenure and promotion decision, but they identified trends and concerns – they raised flags – requiring further investigation. Accordingly, they should be considered along with other information for purposes of assessing teaching effectiveness in tenure and promotion. This was, after all, a process undertaken by experienced colleagues who were well versed and skilled in reviewing and appropriately assessing information – and doing so in context – using a variety of sources as part of their evaluation of a professor’s teaching effectiveness.

In the University’s submission, interest arbitration principles were also important in the adjudication of this dispute – and the most important one to be considered here was gradualism. Any change to a long-standing universally accepted evaluation tool must be careful, deliberate and subject to extensive study and review. This was not an appropriate case for the award of a breakthrough provision introducing enormous

change. Change can occur, the University argued, but it was best achieved meaningfully and thoughtfully through institutions of collegial governance.

Discussion

It is axiomatic that demonstrated evidence of high quality in teaching is an essential requirement for tenure and promotion. Traditionally, substantial reliance has been placed on student evaluations of teaching (hereafter “SETs”) to assess teaching effectiveness. Indeed, SET results are ubiquitous and heavily relied upon in academic personnel decisions. Given their use and impact, on faculty and their careers, a high standard of justice, fairness and due process is self-evidently required. The evaluation of teaching effectiveness for purposes of tenure and promotion is so important – to both the faculty member and the University – that it has to be done right. Tenure and promotion decisions need to be made on the best possible evidence.

SETS Have Value

Significantly, SETs are the main source of information that comes directly from students about their educational experience. Student satisfaction is important to the mission of the University. Learning about it is appropriately considered in making personnel decisions. This information should be considered – it is relevant. The Association takes no issue with this. It agrees that the FCS can continue to be used. And, quite clearly, SET results have a role to play in providing data about many things such as the instructor’s ability to clearly communicate, missed classes made up, assignments promptly returned, the student’s enjoyment and experience of the class, and its difficulty or ease,

not to mention overall engagement. SETS capture student experience. Both professors and the University need to know about that.

Indeed, even subjective student responses can be insightful and useful to the faculty member and the University in charting individual progress: Would they recommend the professor? How much work was involved? Was the course difficult or easy? Was the professor accessible? But interpreting these responses is challenging – they tell a story and add depth but the information needs to be carefully contextualized. The strengths and weaknesses of the SET need to be fully understood.

Strengths and Weaknesses

While SETs are easy to administer and have an air of objectivity, appearances are somewhat deceiving. SET results make it easy to compare Professor A with Professor B, and to others in the Department, Faculty and University. SETs seem scientifically sound: the objective correlation of numerical data. But upon careful examination, serious and inherent limitations in SETs become apparent (discussed further below). The expert evidence led at the hearing persuasively demonstrates that the most meaningful aspects of teaching performance and effectiveness cannot be assessed by SETs. Insofar as assessing teaching effectiveness is concerned – especially in the context of tenure and promotion – SETs are imperfect at best and downright biased and unreliable at worst.

According to the evidence, which was largely uncontested, and which came in the form of expert testimony and peer reviewed publications, numerous factors, especially personal characteristics – and this is just a partial list – such as race, gender, accent, age and “attractiveness” skew SET results. It is almost impossible to adjust for bias and stereotypes. Student and faculty gender affects outcomes, as does grade expectation. Other systemic problems were identified by the experts, and in the literature. One example, and there were many, is the reliability of SETs completed on line versus those completed in class. There are differences between the results of absent students who complete SETs online and those who complete the forms in class. These differences need to be understood. Overall response rates also need to be considered: the lower the response rate the less reliable the results. There is certainly no reason to believe that the views of responders can be extrapolated and applied to non-responders.

Course characteristics – elective versus required – class size – small versus large – hard versus soft – quantitative versus humanities – traditional teaching versus innovative – these are just a few of the other factors that can impact results. Some traditional SET questions about teaching effectiveness are highly problematic, as are questions that seek to measure the breadth of the instructor’s knowledge and scholarship, i.e. whether the curriculum was up to date. It is far from clear whether students have the expertise to comment on course content and teaching methods and assignments. Moreover, if faculty believe SETs will be used to assess performance, they may modify behavior – teaching to SETs – to achieve desired outcomes, ones that have little to do with what is supposedly being assessed. SET results also have a demonstrated correlation with student grade expectations, and the timing of the administration of the SET may

influence its reliability. The list goes on. And this is just the briefest of summaries of a small part of the voluminous evidence adduced and filed at the hearing about the myriad problems in relying on SETs to evaluate teaching effectiveness.

The fact of the matter is that SET results have demonstrable limitations that raise real issues about their use as a measure of teaching effectiveness in tenure and promotion decisions. A further complication is the practice of reducing the FCS results to averages and then comparing individuals with other individuals, the Department, Faculty and University. The evidence is clear, cogent and compelling that averages establish nothing relevant or useful about teaching effectiveness. Averages are blunt, easily distorted (by bias) and inordinately affected by outlier / extreme responses. Quite possibly their very presence results in inappropriate anchoring.

There is no demonstrated value in comparing average results across course formats, levels, topics and disciplines. This is statistically unhelpful and only exacerbates the confounding biases earlier referred to. The FCS can continue as it provides some important information (albeit not about teaching effectiveness), but the only relevant metric is frequency distribution as was demonstrated at length during the proceedings, and convincingly illustrated in evidence – the exhibits that were filed – and in argument.

Stated somewhat differently, SETs score generally low on any reliability marker, and while they can continue to be used, caution is in order. The data must be presented in a meaningful way: that means not as an average compared to other averages, which is close to meaningless for a long list of reasons set out in the expert evidence and exhibits, but as a frequency distribution together with response rates (to assess the usefulness/reliability of the data). Frequency distributions allow for a more accurate analysis and nuanced assessment.

The Gold Standard for Measuring Teaching Effectiveness

The expert evidence convincingly establishes a few other things, the most important of which is that the best way to assess teaching effectiveness is through the careful assessment of the teaching dossier and in-class peer evaluations (as provided for in the collective agreement, although here too potential bias must be addressed and, for the reasons explained at the hearing, can be, at least to some extent, mitigated). Extremely comprehensive teaching dossiers – as is also already anticipated by the collective agreement – containing diverse pedagogical information drawn from the instructor and other sources should provide the necessary information to evaluate the actual teaching as an ongoing process of inquiry, experimentation and reflection. Together with peer evaluation, they help paint the most accurate picture of teaching effectiveness.

It is probably impossible to precisely measure teaching effectiveness. But the difficulties in doing so cannot serve as a justification for over-relying on a tool – the SET – that the evidence indicates generates ratings but has little usefulness in measuring teaching

effectiveness. At the same time, FCS results can continue to be used in tenure and promotion, when the results are presented as frequency distributions, and when the end users are appropriately educated and cautioned about the inherent limitations both about the tool and the information it generates. As noted at the outset, FCS results provide information about the student experience, and, contextualized, are appropriately considered for tenure and promotion although, to repeat, not for reaching conclusions about teaching effectiveness.

Award

While specific and detailed RFA proposals were advanced at the hearing, it is clear that the parties – the Association and the University – share the commitment of ensuring that faculty teaching effectiveness is fairly and accurately assessed. While gradualism in interest arbitration is important, this case is one of demonstrated need for change and that conclusion informs the directions that follow. Put another way, this issue has been contested for many years. Despite their best efforts, the parties have not, collegially, or with third party assistance, been able to resolve it. There has been no lack of trying. When those attempts failed the parties, as they are entitled, sought third party mediation and adjudication and this is the result – a result reached after a hearing that benefited from expert evidence.

That evidence, as earlier noted, was virtually uncontradicted. It establishes, with little ambiguity, that a key tool in assessing teaching effectiveness is flawed, while the use of averages is fundamentally and irreparably flawed. It bears repeating: the expert

evidence called by the Association was not challenged in any legally or factually significant way. As set out above, the assessment of teaching effectiveness is critical, for faculty and the University, and it has to be done right. The ubiquity of the SET tool is not a justification, in light of the evidence about its potential impact, for its continuation, or for mere tinkering. The evidence is dispositive that some of the questions do not elicit any useful information about teaching effectiveness and are subject to bias, while the use of averages – individual, Departmental, Faculty and University – provides no relevant information about teaching effectiveness.

Final Observation

A final observation is necessary: I acknowledge that the purpose of legal proceedings is to bring about finality, and I am very well aware that this issue has been ongoing for considerable time. The Association advanced detailed collective agreement provisions it asked be awarded. However, I am of the opinion that it makes the most sense to initially set out some general directions in order for the parties to at least try to reach agreement on how these directions should be best memorialized in their collective agreement.

In particular, I direct the following:

- The collective agreement is to be amended to ensure that FCS results are not used to measure teaching effectiveness for promotion or tenure. A necessary corollary of this direction is that the numerical weighting system in the FCS be replaced with an alphabetical one. Question 15 should be struck.
- The parties are directed to meet and agree upon an appropriate, user-friendly, intelligible and easily accessible mode of presentation of FCS data in the form of a frequency distribution together with response rates. The parties are also directed to ensure that Deans and DECs, FTCs, FPCs and any others charged with evaluating faculty are educated – a process already set out in the collective agreement – in inherent and systemic biases in SETs so that FCS results can be considered in light of their actual and inherent limitations and with necessary context. The parties are directed to meet and try to agree upon governing guidelines to be incorporated into the collective agreement. The parties are further directed to establish a properly resourced joint committee with hard time lines to consider the current FCS and the possible revision/addition to the questions that are asked, the provision and use of comments, the process and methodology for FCS administration including paper, online, in class, scheduled time, and data tracking of response rates.
- Until and unless the parties agree on a substitution for the online system for non-online courses, it is to be discontinued for probationary faculty.

Conclusion

This award will be effective on a date to be agreed upon by the parties. Failing agreement, I remain seized both with respect to determining the effective date and with any implementation issues that may arise.

DATED at Toronto this 28th day of June 2018.

“William Kaplan”

William Kaplan, Sole Arbitrator