Rate my attitude: research agendas and RateMyProfessor scores

Michael Carlozzi

To cite this article: Michael Carlozzi (2017): Rate my attitude: research agendas and RateMyProfessor scores, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2017.1348465

To link to this article: http://dx.doi.org/10.1080/02602938.2017.1348465

Published online: 02 Jul 2017.
Rate my attitude: research agendas and RateMyProfessor scores

Michael Carlozzi
Independent Researcher, USA

ABSTRACT
The literature on student evaluations of teaching (SETs) generally presents two opposing camps: those who believe in the validity and usefulness of SETs, and those who do not. Some researchers have suggested that ‘SET deniers’ resist SETs because of their own poor SET results. To test this hypothesis, I analysed essays by 230 SET researchers (170 lead authors) and classified the researchers as having negative, neutral or positive attitudes towards SETs. I retrieved their RateMyProfessors.com (RMP) scores and, using logistic regression, found that lead authors with negative attitudes towards SETs were 14 times more likely to score below an estimated RMP average than lead authors with positive attitudes towards SETs. Co-authors and researchers with neutral attitudes, on the other hand, did not significantly differ from the RMP average. These results suggest that personal attitudes towards SETs may drive research findings.

Introduction
Despite a near century of study, student evaluations of teaching (SETs) remain contentious. The seemingly endless debate relates to SET's role in high-stakes employment decisions: tenure allotments, merit-pay raises and, especially in the case of contingent labour, contract renewal (Williams and Ceci 1997; Olivares 2004; Crumbley and Reichelt 2009; Heller 2012). Researchers connect SETs to broader educational issues such as grade inflation (Goldman 1985; Eiszler 2002), institutional racism (Nast 1999) and increasing levels of consumerism and entitlement (Titus 2008). While these features give reasonable rise to policy-level debate, they fail to explain how researchers lack agreement on essentially everything about ‘the most thoroughly studied of all forms of personnel evaluations’ (Marsh 1987, 369).

One reason for the ongoing controversy might be rooted in research agendas. Some authors present research studies to advance explicit positions; thus, Nargundkar and Shrikhande (2012) found that ‘the (SET) literature tends to either extol the virtues of (SETs) or denigrate them as useless’ (69). In other words, the field of SET research can be understood as roughly divided between two opposing camps: (1) those who defend SETs, so-called ‘SET apologists’ (Stark 2015), and (2) those who attack SETs, so-called ‘(SET) deniers’ (Benton and Ryalls 2016). The latter typically reject the validity of SETs, often vehemently, and may even do so in public fora. In general, ‘deniers’ centre their rejections around SET bias, noting how research finds significant differences among certain groups (e.g. female and male instructors). Some go so far as to argue that students should not rate instructors (Browne et al. 1997; Vohra 2016) and that SETs carry problematic, even ‘dangerous’ assumptions (Grayson, 2015). They worry that students with ‘an axe to grind’ will punish faculty. Apologists, meanwhile, defend SETs against ‘wild statements of personal opinions’ (Theall 2010, 45), describing the experience of debating SETs as frustrating and
disappointing (Benton and Ryalls 2016, 1). They promote certain SET instruments as valid formative and summative tools and argue that their shortcomings are overblown (Feldman 1993; Marsh and Roche 1997; Centra and Gaubatz 2000).

The scholarly stalemate prompted my research question – how could these two positions endure despite decades of research, public attention and high-stakes outcomes? Some persistent disagreement arises from differences in interpretation. For example, virtually all researchers acknowledge that SETs correlate with students’ expected grades (Aleamoni 1987). SET deniers interpret this finding as evidence of grade inflation, i.e. the ‘leniency’ hypothesis. SET apologists, on the other hand, propose that this correlation might reflect meaningful student learning or a student’s prior interest in the subject matter (Marsh 1987; Feldman 2007). Some deniers propose a reciprocity effect whereby students quid pro quo offer high evaluations for high grades (Clayson, Frost, and Sheffet 2006), a conclusion rejected by other researchers (Norvilitis and Zhang 2009). Since so few studies triangulate student learning with teaching effectiveness and student ratings – and definitions of student learning remain contested – neither perspective has conclusive evidence.

But not all disagreement is interpretive. The well-acknowledged contradictory results of SET research stem from a ‘great variety of methods, measures, controlling variables, SET instruments, and populations’ (Spooren, Brockx, and Mortelmans 2013, 617). In this regard, SET research has been remarkably constant; earlier research identified the literature to be ‘very extensive … contradictory, and … of highly variable quality’ (Dowell and Neal 1982, 51). The end result is an environment whereby ‘almost any conclusion’ can be reached by marshalling individual studies (Benton and Cashin 2012, 12).

I wondered how certain researchers kept reaching the same conclusions, radically different from researchers on the ‘other side’. Theall and Franklin (2001) suggested that low SETs may lead certain instructors to ‘develop negative attitudes toward students and student ratings’ (47). Could a researcher’s beliefs, and the arguments and research informing those beliefs, stem from his or her own experiences with SETs? Could it be that SET deniers are ‘professors who receive low evaluations’ (Moore 1990, 260)?

If Moore (1990) and others were right, I hypothesised that this speculation could be observed in SET data. While classroom-based SET scores can be difficult to obtain, the online repository RMP provides a publically available data-set that has been found to approximate reliably actual SET scores (Coladarci and Kornfield 2007; Timmerman 2008). RMP allows users to rate instructors in both perceived quality (‘overall quality’) and perceived difficulty (‘level of difficulty’); an instructor’s difficulty score does not factor into his/her quality score. Their quality score is intended to measure an instructor’s overall effectiveness.

I made the following hypotheses:

1. In general, SET deniers and SET apologists would have different RMP scores.
2. Co-authors of SET articles would have RMP scores that resembled the RMP average. I assumed that lead authors would endorse the views of their research and I hypothesised that co-authors might not share this enthusiasm. Given the dynamics of academic research, it is impossible to judge a priori whether co-authors believe in their work’s ideology. Co-authors may be graduate students facilitating data collection, statisticians providing expert analysis or simply prolific colleagues. SET research especially invites multidisciplinary collaboration, as some academic researchers may lack adequate sufficient expertise to analyse SET data-sets.
3. Conclusions reached in ‘thought pieces’ would be more influenced by a researcher’s attitude towards SETs than would research since research should be more ‘objective’ than personal argument.
4. Corroborating prior research (Constand and Pace 2014), lower SET scores would be associated with higher RMP difficulty scores.

I analysed 230 total SET researchers (lead authors = 170) and classified them as having negative, neutral or positive attitudes towards SETs. After placing researchers into one of these three groups, I collected their RMP ratings and determined, through logistic regression, the extent to which SET attitudes predicted RMP scores. This study’s primary finding is that lead authors with ‘negative’ attitudes towards
SETs were 14 times more likely to receive below average RMP scores than ‘positive’ researchers. No such differences were found for authors without apparent strong interests in the subject matter, i.e. co-authors and ‘neutral’ researchers. Animus and amity, it seems, correlate with research findings.

**Method**

**Data collection and coding**

I built a data-set of eligible SET researchers through two methods. The first involved extensive citation chaining through the SET literature. This allowed me to access popular and seminal essays so that I would not omit major, or even modest, figures in the field of SET research.

The second method involved searching library discovery portals (PRIMO and OneSearch), Google, Google Scholar and academic databases (ERIC, ProQuest and JSTOR). Keywords used were *SET, student ratings of instruction, student teaching evaluations* and similar combinations (e.g. ‘evaluating instructors’).

Identified essays were first checked for relevance. A relevant essay discussed SETs as the primary object of interest. For example, I would exclude essays that discussed SETs as the vehicle of a university’s assessment outcomes. Similarly excluded were essays about developing robust teaching evaluations. An essay, however, that discussed teaching evaluations *with an argument about the adequacy or inadequacy of SETs* would be included. Relevant essays also needed to be published in peer-reviewed journals.

Essays were then categorised as either ‘research’ or ‘thought’. ‘Thought’ referred to an essay that was principally argumentative where the author(s) did not statistically analyse SETs, though he/she may have relied on the statistical analysis of others. ‘Research’ referred to an essay that performed statistical analysis on SETs, conducted an experiment related to SETs or tested the validity of a particular SET instrument.

The essay’s first author was classified as its ‘lead author’ and other authors were classified as ‘co-author’. I then read the essay (or essays, in the case of multiple results for one author) to classify the researcher’s attitudes towards SETs as negative, neutral or positive. Essays classified as ‘thought’ frequently made bold claims; for example, one essay labelled SETs as an ‘inaccurate, misleading, and shamming procedure’ (Gray and Bergmann 2003, 46). ‘Research’ essays more often demanded interpretative judgement because academic writing avoids confrontation and promotes hedging and indeterminate claims (Myers 1989; Hyland 1994; Salager-Meyer 1994; Martin-Martín 2008; Lancaster 2016).

Furthermore, both sides make various concessions, which could not themselves be interpreted as positive or negative claims. SET apologists concede the following points:

1. SETs should be part, not the whole, of an instructor’s summative evaluation (Benton and Ryalls 2016).
2. Certain non-teaching factors correlate with SETs, i.e. expected grades, student motivation, subject discipline and class size (Centra and Creech 1976; Feldman 1978; Cashin 1990; Benton et al. 2015). Many researchers, SET apologists included, advocate adjusting SET scores for these non-teaching factors (Benton and Ryalls 2016). Some apologists concede that certain factors (e.g. gender) introduce bias but they maintain that these effects matter little to an overall analysis (Centra and Gaubatz 2000).
3. Validated instruments should be used to deliver SETs because idiosyncratic instruments may introduce unintended biases (Theall and Franklin 2001); SET instruments are often regarded as tools (Boysen 2016).
4. Sound administrative and statistical practice should guide SET application, such as analysing instructors within appropriate contexts (Theall 2010; Benton and Ryalls 2016). For example, the performance of physics instructors should be analysed within the context of physics or at least the physical sciences.

Similarly, many staunch SET deniers make these concessions:
Student input has value. Most SET deniers contend that administrations over-rely on simplified, quantitative input that fails to measure teaching effectiveness (Greenwald 1997; Stark and Freishtat 2014).

SETs are pervasive and carry significant weight with regard to high-stakes personnel decisions (Clayson 2009). A researcher who argues that teachers should take SETs seriously does not necessarily hold 'positive' feelings towards them; by itself, this is a neutral, pragmatic stance.

If a researcher argues, for example, that SETs should not solely determine faculty promotion, then this reasoning does not classify the researcher as having 'negative' feelings towards SETs. Such reasoning may be expressed by those who do have negative feelings towards SETs, but this particular statement is no more extreme than anything expressed by SET apologists. Similarly, a researcher who discovers a correlation between expected grades and SETs is not 'negative.' If this researcher presents the correlation as evidence of 'buying' grades, however, then he/she would be classified as 'negative'. On the other hand, authors who argue against this interpretation would be classified as 'positive', as in the case of Gerkin and Kierkus (2011): 'Our study also helps dispel myths like: teachers can “buy” high evaluation scores with high grades, or teachers who are assigned “difficult” classes like quantitative research methods are doomed to receive low evaluation scores' (35). In short, to be classified as positive or negative, researchers had to express sentiments at least as extreme as the archetypes on either 'side'.

A researcher who takes no discernible stance – essentially presents the data and acknowledges its ambiguity, perhaps explaining the primary camps of interpretation – becomes 'neutral'. If I could not confidently determine a researcher's attitudes because of conflicting statements, I erred on the side of caution and classified him/her as 'neutral'. This happened infrequently, as most researchers extended descriptive findings to policy-level arguments in their introductions, discussions or concluding sections. ‘Neutral’ ratings came mostly from descriptive research stances, such as researchers assessing the impact of online response rates on SET scores.

After classifying all eligible researchers, I stopped adding researchers to the data-set and started collecting SET ratings. Because no publicly available database of institutional SETs exists, I relied on RMP. To mitigate the influence of small samples, I only selected researchers with at least five ratings, in line with criteria from previous research (Carter 2016). I used the instructor’s overall quality as his/her SET score, and I combined ratings for instructors who taught at multiple institutions. Difficulty scores ('level of difficulty') were also recorded. I stopped adding RMP entries on 20 January 2017, so the data-set excludes any ratings entered after that date.

**Statistical analysis**

Data were analysed at the ‘instructor-level’, as instructors were the unit of interest. I performed logistic regression to answer the primary research question, ‘Does an instructor’s SET score – as approximated by RMP – relate to his/her attitude toward SETs?’ Such a question is well-suited for the methods of logistic regression. By calculating an odds ratio, logistic regression offers a real-world interpretation of the data. I sought to answer the following research questions:

1. How much more likely were researchers to score above or below the RMP average when holding negative or positive attitudes towards SETs?
2. Will RMP scores differ for authors working on essays classified as ‘research’ as opposed to ‘thought’?
3. Will co-authors have RMP scores similar to lead authors?

Based on prevalent research lines within SET scholarship, I also analysed factors often hypothesised to affect SET scores:

4. How do non-instructional factors (instructor gender and subject discipline) affect SET scores?
(5) How does an instructor’s perceived difficulty relate to his/her odds of outperforming the RMP average?

The RMP average chosen (3.70) was supported by recent research (Clayson 2014; Carter 2016). For the purposes of comparing scores, I selected a RMP score of 3.65 to accommodate some level of variability. Predictor and outcome variables are explained on a case-by-case basis in the results section.

**Results**

Table 1 presents descriptive statistics for all relevant groups. The total number of student responses was 6316 (5255 for lead authors); the mean number of responses per instructor was 31.4 (median = 22).

Given that neutral researchers did not differ significantly from the RMP average (3.73 to 3.70), I removed them from the following model as they might suppress differences between positive and negative researchers. I ran a logistic regression model on all lead authors with positive and negative attitudes towards SETs ($N = 126$). The predictor variable was the author’s attitude (positive = 0; negative = 1) and the outcome variable was the author’s relationship to the RMP average (SET score ≥ 3.65 = 0; SET score < 3.65 = 1). This model was significant: $\chi^2 (1) = 40.86, p < 0.001$ (Table 2). Negative researchers were 14 times more likely to score below the RMP average than were positive researchers.

I turned to the research question involving co-authors, adding neutral authors back into the data-set. With the same predictor variable, but only for co-authors, a logistic regression model was not significant: $\chi^2 (1) = 0.004, p = 0.95$. With neutral authors removed from the data-set ($N = 40$), the model also lacked significance: $\chi^2 (1) = 0.073, p = 0.79$. Attitude classification did not predict co-author scores; as mentioned above, however, it did predict lead authors’ scores.

I then performed logistic regression with attitude alongside variables determined relevant by the SET literature, i.e. gender and subject discipline. Neutral researchers were present in this analysis. Gender was coded as male = 0 and female = 1. Subject discipline was coded as business = 0, social sciences = 1, physical sciences = 2, humanities = 3 and professional = 4 (law and social work). The model was significant due to the attitude variable: $\chi^2 (3) = 36.65, p < 0.001$ (Table 3). Gender and subject discipline were not significant predictors. The lower odds ratio for the attitude variable (down from 14.11) comes from contamination by neutral researchers, not the influence of other predictors.

Because I had hypothesised that research essays might involve less subjectivity than thought pieces, I performed logistic regression with the predictor variable of ‘research’ (0) and ‘thought’ (1). The model was not significant: $\chi^2 (1) = 3.77, p = 0.058$ (Table 4). Moreover, 79% (27) of thought pieces were written by authors with negative views towards SETs, introducing a strong confounding variable.

**Table 1.** Descriptive statistics for all groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>% Above average</th>
</tr>
</thead>
<tbody>
<tr>
<td>All instructors</td>
<td>230</td>
<td>3.62</td>
<td>0.84</td>
<td>51.3</td>
</tr>
<tr>
<td>Lead authors only (All)</td>
<td>170</td>
<td>3.55</td>
<td>0.87</td>
<td>47.1</td>
</tr>
<tr>
<td>Lead authors only (Research)</td>
<td>136</td>
<td>3.62</td>
<td>0.85</td>
<td>50.7</td>
</tr>
<tr>
<td>Lead authors only (Thought)</td>
<td>34</td>
<td>3.28</td>
<td>0.91</td>
<td>32.4</td>
</tr>
<tr>
<td>Co-authors only (All)</td>
<td>60</td>
<td>3.84</td>
<td>0.73</td>
<td>63.3</td>
</tr>
<tr>
<td>Negative authors only (All)</td>
<td>81</td>
<td>3.18</td>
<td>0.85</td>
<td>24.7</td>
</tr>
<tr>
<td>Neutral authors only (All)</td>
<td>44</td>
<td>3.73</td>
<td>0.79</td>
<td>52.3</td>
</tr>
<tr>
<td>Positive authors only (All)</td>
<td>45</td>
<td>4.05</td>
<td>0.65</td>
<td>82.2</td>
</tr>
</tbody>
</table>

**Table 2.** Negative versus positive attitude variable for positive and negative lead authors ($N = 126$).

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard error</th>
<th>$p$ Value</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−1.12</td>
<td>0.26</td>
<td>0.0000</td>
<td>–</td>
</tr>
<tr>
<td>Attitude</td>
<td>2.65</td>
<td>0.47</td>
<td>0.0000*</td>
<td>14.11 (5.64–35.26)*</td>
</tr>
</tbody>
</table>

*Significant at the $p < 0.01$ level.
Finally, I analysed the likelihood of scoring above the RMP average with a high difficulty rating. The model was significant: $\chi^2 (1) = 51.10, p < 0.001$ (Table 5). Of researchers with a difficulty score at least as high as the RMP average ($N = 52$), just 14% (7) scored above the RMP average; I chose the RMP average as a convenient benchmark, not because it related to the difficulty score in any way.

**Discussion**

These results present a relationship between a researcher’s RMP scores – which approximate actual classroom SETs – and his/her scholarship findings in SETs. Specifically, researchers with higher RMP ratings were more likely to approach SETs positively and reach positive conclusions about them, and researchers with lower RMP ratings were more likely to view them negatively and reach negative conclusions about them. Indeed, only 25% of SET deniers scored above the RMP average compared to 82% of SET apologists; their group means differed significantly from 3.18 to 4.05. Deniers were 14 times more likely than apologists to score below the RMP average. Neutral researchers and co-authors, meanwhile, scored around the RMP average. Such findings connect SET scores and SET attitudes, supporting speculation that researchers who deny the validity of SETs are also those with low SET scores.

RMP scores did not significantly differ ($p = 0.058$) based on broad classification (‘research’ versus ‘thought’). Slight mean differences likely stemmed from bias; 27 of the total 34 (79%) thought pieces came from SET deniers; the ‘attitude’ variable confounded this analysis. As hypothesised, co-authors did not differ significantly from the RMP average (3.73–3.70). This suggests that lead authors push research agendas with which co-authors may or may not agree; co-authors appear largely unattached to ideology.

Gender and course discipline were not found to be significant predictors, but this should not be interpreted to mean that these effects do not exist. Instead, this study – using a data-set much smaller than many other SET studies – likely lacked the power to detect these effects. Gender and course discipline have consistently been found to be small but noticeable effects (Centra and Gaubatz 2000; Beran and Violato 2005; Rosen 2017). After all, even staunch SET apologists concede that outside factors influence scores.

These findings may help address the SET research ‘stalemate’. Take, for example, the debate over SET forms. For some years now, a segment of SET scholars has argued for moving beyond questions

<table>
<thead>
<tr>
<th>Table 3. Attitude, gender and subject discipline on the outcome variable for all lead authors.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Attitude</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Subject discipline</td>
</tr>
</tbody>
</table>

$^*$ Significant at the $p < 0.01$ level.

<table>
<thead>
<tr>
<th>Table 4. Essay type on the outcome variable RMP average ($N = 170$).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Essay type</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5. Difficulty as a continuous predictor on the outcome variable ‘above’ or ‘below’ the RMP average.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Difficulty</td>
</tr>
</tbody>
</table>

$^*$ Significant at the $p < 0.01$ level.
of instrument validity. Other researchers, however, oppose any such development. My results suggest that this irreconcilable disparity might be deeply rooted in personal experiences and therefore will be difficult, if not impossible, to overcome. This is the current research climate, where little progress seems to be made despite insistence from both ‘sides’ to improve the use of SETs in high-stakes decisions.

These findings further challenge assumptions of scholarly open-mindedness as well as the quality control assumed in peer review. How could so many studies reflect predetermined attitudes rather than conscientious analysis? I would argue that not only is this possible but that the special position of SET research allows for these problems to persist and even to proliferate.

To start with, the abundance of SET research permits the existence of multiple competing agendas. No single literature review can represent the thousands of previous studies, increasing the likelihood of selective framing. As some authors note, virtually any conclusion can be reached by marshalling the ‘right’ constellation of studies (Benton and Cashin 2012). If a researcher wanted to frame SET scholarship in the light of a particular attitude, enough work exists to create a plausible narrative of such consensus.

Methodological complexities also confound SET analysis. SET scholarship is multidisciplinary, and standards of study design and analytical methods vary widely both in type and rigour. Statistical literacy cannot sufficiently deter suspect analysis. Even in fields that commonly publish statistical analysis, peer reviewers miss egregious statistical errors or provide spurious and irrelevant criticisms (Goodman, Altman, and George 1998; Bacchetti 2002; Strasak et al. 2007; Norman 2010). One might imagine that multiple perspectives on SETs would create nuanced, productive research trajectories that build towards a common set of principles. Yet, this has not been the case. Instead, we find a field of inquiry where researchers arrive at preferable conclusions by citing the right authors or running the right regressions.

Throughout this paper, I have assumed that high SET scores are desirable. Given the relationship of SETs to employment decisions, this seems undeniable. But the role of SETs in these decisions assumes that SET scores reflect teaching effectiveness. SET deniers frequently claim the opposite: low SETs reflect rigour, and, by extension, teaching effectiveness. Lower scores can thus be seen as a badge of honour because they evidence academic rigour.

RMP allows for an empirical test by allowing students to rate separately an instructor’s difficulty; the instructor’s difficulty score does not affect the quality score. My final hypothesis was that low SET scores would reflect high difficulty scores, and, indeed, the results show that RMP difficulty scores correlate with overall quality scores; each unit increase in difficulty made it almost seven times less likely to score above the RMP average. SET deniers would likely interpret this finding as further evidence that instructors will lower standards and rigour in the interest of increasing “the (SET) number” (Marks and O’Connell 2003, 259). But it is not so simple. Some SET research has suggested that difficulty does not necessarily entail grade leniency (Rizvi 2015). And after examining student rating patterns, I found that RMP raters discussed and interpreted course difficulty differently than SET deniers, consistent with work on how students perceive course difficulty and workload (Addison, Best, and Warrington 2006; Dee 2007). In other words, RMP difficulty scores, taken in the context of RMP’s explanatory comments, might relate to instructor attitudes and behaviours rather than course expectations and workloads.

When rating difficult instructors on RMP, students focused on several characteristics that reflect ineffectual – and potentially absent – teaching. A class was rated difficult because, for example, the instructor refused to answer emails and to meet with students, referring all questions to teaching assistants. I argue that this does not reflect academic rigour. It portrays a difficult class, but one which is logistically, even artificially, hard. Further analysis of RMP comments would be needed, but the odds ratios in this paper, alongside my admittedly cursory review of RMP comments, suggest that classes are rated ‘difficult’ based in large part on the instructors’ uninterested attitude towards students. Students find this attitude manifesting in several outcomes: an almost fanatical devotion to lecturing, an inability to return material and an overall lack of availability. These student comments provide some support for the view that SET deniers have a ‘general [insulting] attitude toward students’ (Benton and Ryalls 2016, 2). Of course, the subject requires further study; it was not this project’s primary object of interest.

Needless to say, any study of this design carries limitations. First, not all SET researchers are represented in this sample. Many SET researchers published long before the advent of RMP. Other researchers
did not have the requisite number of RMP ratings (5) to be included. Still other researchers were not based in United States’ institutions, and RMP, in general, lacks data on institutions outside of the United States. And some researchers do not work in academia at all.

Second, the study’s methodology, i.e. the manner by which researchers were classified, might have facilitated some errors. Classifications were subjective by their very nature. Many researchers voiced clear opinions on SETs, but for some researchers I had to infer feelings based on their discussion and conclusion sections. With close calls, I read essays multiple times, sometimes days apart to approach them with ‘fresh eyes’. When in doubt I classified researchers as ‘neutral’; to borrow the language of statistics, I felt it better to commit a Type II than a Type I error.

Third, the classification scheme of this paper simplifies SET discourse. Much SET analysis is nuanced and does not present straightforward ‘yea’ or ‘nay’ arguments. I am sure that many researchers would object to being labelled ‘apologists’ or ‘deniers’; it is, after all, the nature of academics to resist simplistic classification. As is to be expected with observational data, not all researchers fell in with their respective groups. Some SET deniers – and in argumentative essays they often assumed this very mantle – scored well on RMP. Some SET apologists, in contrast, performed poorly on RMP, such as one researcher who chastised faculty for refusing to accept their poor SETs; this researcher’s overall quality RMP score was 2.72 (N = 25). The trend, however, was that negative researchers scored below the RMP average, neutral researchers around the RMP average and positive researchers above the RMP average.

In conclusion, this study finds that researchers’ personal attitudes towards SETs might influence their research findings. Researchers with negative attitudes towards SETs were 14 times more likely to score below the RMP average than researchers with positive attitudes towards SETs. Groups without apparent vested interests in SETs – co-authors and ‘neutral’ researchers – did not differ significantly from the RMP average. The ‘great variety of methods’ of SET scholarship may result not so much from panoply of choice – methods ranging from structural equation modelling to Bayesian analysis – as from agendas to find the ‘right answer’. This conclusion echoes the sentiments of Gravestock and Gregor-Greenleaf (2008), who were ‘dismayed and concerned [by an] … apparent lack of objectivity’ among SET scholars (7). Perhaps it is not so much retaliatory students as faculty who have an ‘axe to grind’.

Acknowledgements

I would like to thank Alexis Teagarden and the paper’s two reviewers for their constructive feedback on all aspects of this project, as well as Joe Essey for his assistance in coding a program to parse RateMyProfessors.com.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Michael Carlozzi is an independent researcher and library director. He has research interests in assessment, library science, instructional design and the perfect game of cribbage.

ORCID

Michael Carlozzi http://orcid.org/0000-0003-0773-3867

References


