Protocol Paper: The Impact of Open-Ended Student Evaluations of Teaching on University Professors' Psychological Health and Well-Being: A Systematic Review Protocol

Watson, Ibrahim B.

Abstract

Background: Open-ended student evaluations of teaching (SETs) have become ubiquitous in higher education quality assurance systems worldwide. While these evaluations serve institutional accountability purposes, emerging evidence suggests that negative, abusive, or non-constructive student comments may significantly impact faculty psychological well-being, professional confidence, and career trajectories. The psychological toll of hostile feedback, particularly in digital contexts where anonymity can amplify aggressive language, represents an understudied occupational health issue in academia.

Objectives: This systematic review aims to synthesize evidence on how open-ended student evaluation comments affect university professors' mental health, psychological well-being, and professional confidence, while identifying evidence-based interventions and automated detection systems for managing abusive feedback.

Methods: We will systematically search MEDLINE, PsycINFO, ERIC, Web of Science, and Scopus databases from inception to [date] using comprehensive search strategies combining terms related to student evaluations, faculty well-being, and psychological impact. We will include quantitative, qualitative, and mixed-methods studies examining the relationship between student feedback and faculty psychological outcomes. Two reviewers will independently screen studies, extract data, and assess methodological quality using appropriate tools. Meta-analysis will be conducted where appropriate, with narrative synthesis for heterogeneous data.

Discussion: This review will provide the first comprehensive synthesis of evidence on an increasingly recognized occupational health issue in higher education, informing evidence-based policies for ethical feedback systems and faculty support mechanisms.

PROSPERO Registration: [To be completed]

Keywords: student evaluations; faculty well-being; academic mental health; occupational stress; higher education; systematic review

1. Introduction

Background

Student evaluations of teaching (SETs) have evolved from informal feedback mechanisms into standardized institutional requirements across global higher education systems (Uttl et al., 2017; Boring et al., 2016). While traditional Likert-scale ratings dominate quantitative assessment frameworks, open-ended comment sections have become increasingly prevalent, ostensibly providing nuanced qualitative insights into teaching effectiveness (Clayson, 2009; Spooren et al., 2013). However, the anonymity inherent in these systems, combined with digital delivery platforms, has created conditions where students may express feedback in ways that extend beyond constructive criticism into personally attacking, discriminatory, or psychologically harmful territory (Miller & Seldin, 2014; Peterson et al., 2019).

Recent qualitative studies have documented faculty experiences of receiving student comments containing personal attacks, inappropriate sexual commentary, discriminatory language targeting protected characteristics, and threats to professional reputation (Boring et al., 2016; MacNell et al., 2015). These hostile communications represent a form of workplace harassment that occurs within officially sanctioned institutional processes, creating complex ethical and legal implications for university administrations (Heffernan, 2022; Schmidt, 2019). The psychological impact of such experiences parallels documented effects of cyberbullying and workplace mobbing, including increased anxiety, depression, reduced self-efficacy, and career-related stress responses (Kokkinos et al., 2015; Nielsen et al., 2012).

The gendered and racialized nature of abusive student feedback has been extensively documented, with women faculty and faculty of color disproportionately receiving comments focused on personal appearance, authority questioning, and competence challenges unrelated to pedagogical performance (Boring et al., 2016; Reid, 2010; Anderson & Smith, 2005). This pattern suggests that student evaluation systems may inadvertently perpetuate workplace discrimination and contribute to documented disparities in faculty career advancement and retention (Mengel et al., 2019; Peterson et al., 2019).

From an occupational health perspective, the chronic stress associated with anticipating, receiving, and processing negative student feedback represents a previously unrecognized workplace hazard in academic environments (Kinman & Jones, 2008; Winefield et al., 2003). The cyclical nature of evaluation periods creates predictable stress peaks, while the permanence of written comments in personnel files amplifies long-term psychological impact (Uttl et al., 2017). Unlike other forms of workplace feedback, student evaluations often lack reciprocal dialogue opportunities, leaving faculty unable to clarify misunderstandings or address concerns constructively (Spooren et al., 2013).

Technological advances in natural language processing and sentiment analysis have created new possibilities for automated detection and filtering of abusive content in student evaluations (Chen et al., 2020; Zhang et al., 2018). Machine learning approaches to identifying toxic language, harassment, and discriminatory content show promise for implementation in educational feedback systems (Founta et al., 2018; Davidson et al., 2017). However, the integration of such systems into institutional evaluation processes raises questions about academic freedom, feedback authenticity, and the potential for algorithmic bias in content moderation (Gillespie, 2020; Roberts, 2019).

Rationale and Contribution

Despite growing recognition of student evaluation systems' potential for psychological harm, no systematic review has comprehensively synthesized evidence on the relationship between open-ended student feedback and faculty mental health outcomes. Existing reviews have focused primarily on evaluation validity and measurement properties rather than psychological impact (Spooren et al., 2013; Uttl et al., 2017). This knowledge gap represents a critical limitation in understanding and addressing occupational health risks in higher education.

Furthermore, while individual studies have documented instances of abusive student feedback and proposed technological solutions, no systematic synthesis has evaluated the effectiveness of intervention strategies or automated detection systems. This evidence gap impedes the development of evidence-based policies for ethical evaluation systems and faculty support mechanisms.

The current review addresses these limitations by providing the first comprehensive synthesis of evidence on student evaluation impact on faculty psychological well-being, while systematically evaluating intervention approaches and technological solutions. This synthesis will inform institutional policy development, support system design, and future research priorities in an area of growing concern for academic occupational health.

2. Objectives

Primary Objectives

1. **Psychological Impact Assessment:** To systematically synthesize evidence on the relationship between open-ended student evaluation comments and university faculty psychological health outcomes, including mental health symptoms, well-being indicators, and professional confidence measures.
2. **Harm Characterization:** To identify and categorize types of abusive, discriminatory, or non-constructive student feedback and their differential impacts on faculty psychological outcomes across demographic groups and career stages.

Secondary Objectives

3. **Intervention Evaluation:** To assess the effectiveness of institutional interventions, support mechanisms, and policy modifications designed to mitigate negative psychological impacts of student feedback on faculty well-being.
4. **Technology Assessment:** To evaluate the accuracy, feasibility, and ethical implications of automated detection systems for identifying abusive or inappropriate content in student evaluation comments.
5. **Moderator Analysis:** To examine factors that moderate the relationship between negative student feedback and faculty psychological outcomes, including institutional context, support availability, career stage, and demographic characteristics.

Research Questions

RQ1: What is the magnitude and nature of the relationship between exposure to negative/abusive student evaluation comments and faculty psychological health outcomes?

RQ2: Which characteristics of student feedback content predict greater psychological harm to faculty recipients?

RQ3: What interventions have been implemented to reduce psychological harm from student evaluations, and what is their effectiveness?

RQ4: How accurate and feasible are automated systems for detecting abusive content in student evaluation comments?

RQ5: What factors moderate the impact of negative student feedback on faculty well-being?

3. Methods

Study Registration and Reporting

This protocol follows the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) guidelines (Moher et al., 2015). The review will be registered with PROSPERO prior to study commencement and reported according to PRISMA 2020 guidelines (Page et al., 2021).

Eligibility Criteria

Inclusion Criteria

Population: University faculty, academic staff, teaching personnel, graduate teaching assistants, or adjunct instructors in higher education institutions globally.

Intervention/Exposure: Exposure to open-ended student evaluation comments, particularly those characterized as negative, abusive, discriminatory, or non-constructive. Studies examining automated detection or processing systems for such comments.

Comparison: Studies with comparison groups (faculty not exposed to negative feedback, pre/post intervention comparisons) will be prioritized, but single-group studies will be included if they provide relevant outcome data.

Outcomes:

- **Primary:** Psychological health indicators (depression, anxiety, stress symptoms), well-being measures, professional confidence, self-efficacy, job satisfaction
- **Secondary:** Physiological stress markers, career-related outcomes, coping strategies, intervention effectiveness measures, automated detection system performance metrics

Study Design: Quantitative (experimental, quasi-experimental, observational), qualitative (phenomenological, grounded theory, case studies), and mixed-methods studies.

Publication Type: Peer-reviewed journal articles, conference proceedings, dissertations, institutional reports, preprints.

Language: English, Spanish, French, German, Portuguese.

Time Period: No date restrictions.

Exclusion Criteria

- Studies focusing exclusively on quantitative (Likert-scale) evaluations without open-ended components
- Studies examining student satisfaction or learning outcomes without faculty well-being measures
- Reviews, editorials, or opinion pieces without primary data
- Studies of K-12 teacher evaluations
- Conference abstracts without full-text availability

Search Strategy

Electronic Databases

- **MEDLINE (PubMed):** 1946 to present
- **PsycINFO (APA):** 1806 to present
- **ERIC (Education Resources Information Center):** 1966 to present
- **Web of Science Core Collection:** 1945 to present
- **Scopus:** 1970 to present
- **Academic Search Ultimate:** 1887 to present
- **Business Source Premier:** 1886 to present

Search Terms Strategy

((student* evaluation* OR student* feedback OR student* comment* OR teaching evaluation* OR course evaluation* OR instructor evaluation* OR faculty evaluation*)

AND

(mental health OR psychological OR well-being OR wellbeing OR stress OR anxiety OR depression OR burnout OR self-efficacy OR confidence OR job satisfaction OR occupational health)
AND
(faculty OR professor* OR instructor* OR academic* staff OR teaching* staff OR lecturer*))

COMBINED WITH

((abusive OR negative OR hostile OR inappropriate OR discriminatory OR toxic OR harassment OR cyberbull*)
AND
(student* comment* OR student* feedback OR evaluation* comment*))

COMBINED WITH

((automat* detection OR machine learning OR natural language processing OR sentiment analysis OR content moderation OR text classification)
AND
(student* evaluation* OR student* feedback OR comment* filter*))

Grey Literature and Additional Sources

- **Conference Proceedings:** Association for Institutional Research (AIR), European Association for Institutional Research (EAIR), Society for Teaching and Learning in Higher Education (STLHE)
- **Dissertation Databases:** ProQuest Dissertations & Theses Global
- **Institutional Reports:** National Center for Education Statistics, Higher Education Research Institute
- **Preprint Servers:** EdArXiv, PsyArXiv, bioRxiv
- **Reference Screening:** Forward and backward citation tracking of included studies
- **Expert Consultation:** Contact with subject matter experts for unpublished studies

Study Selection Process

Two reviewers (XX and XX) will independently screen titles and abstracts using predefined criteria in Covidence systematic review software. Full-text screening will be conducted independently by the same reviewers, with disagreements resolved through discussion or third reviewer consultation (XX). Inter-rater reliability will be calculated using Cohen's kappa, with κ ≥ 0.80 considered acceptable.

Data Extraction

Standardized Extraction Form

Study Characteristics:

- Author, year, country, study design, sample size
- Institutional context (type, size, setting)
- Participant demographics (age, gender, race/ethnicity, career stage, discipline)

Exposure/Intervention Details:

- Student evaluation system characteristics
- Types of negative feedback examined
- Intervention components and duration
- Technology specifications for automated systems

Outcome Measures:

- Psychological health instruments used
- Assessment timing and follow-up periods
- Primary and secondary outcome definitions

Results Data:

- Means, standard deviations, effect sizes
- Correlation coefficients, regression parameters
- Qualitative themes and supporting quotes
- Technology performance metrics (sensitivity, specificity, accuracy)

Quality Assessment

Quantitative Studies

- **Randomized Controlled Trials:** Cochrane Risk of Bias 2 (RoB 2)
- **Non-randomized Studies:** Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)
- **Cross-sectional Studies:** Joanna Briggs Institute Critical Appraisal Checklist

Qualitative Studies

- **Critical Appraisal Skills Programme (CASP) Qualitative Checklist**

Mixed-Methods Studies

- **Mixed Methods Appraisal Tool (MMAT)**

Quality assessment will be conducted independently by two reviewers, with disagreements resolved through consensus discussion.

Data Synthesis and Analysis

Quantitative Analysis

Meta-Analysis: Where three or more studies report comparable outcome measures with sufficient statistical data, random-effects meta-analysis will be conducted using Review Manager 5.4. Effect sizes will be calculated as:

- Standardized mean differences (Cohen's d) for continuous outcomes
- Odds ratios for dichotomous outcomes
- Correlation coefficients (Fisher's z-transformed) for association studies

Heterogeneity Assessment: Statistical heterogeneity will be evaluated using:

- Cochran's Q test ($p < 0.10$ indicating significant heterogeneity)
- I^2 statistic (>50% indicating substantial heterogeneity)
- Tau^2 for between-study variance estimation

Subgroup Analysis: Planned subgroup analyses include:

- Faculty demographic characteristics (gender, race/ethnicity, career stage)
- Institutional context (public/private, research intensity, geographic region)
- Evaluation system characteristics (mandatory/voluntary, anonymous/identified)
- Type of negative feedback (personal attacks, competence questioning, discriminatory)

Sensitivity Analysis: Influence of study quality, outliers, and publication bias on pooled estimates.

Qualitative Analysis

Thematic Synthesis: Following Thomas and Harden's approach:

1. Line-by-line coding of study findings
2. Development of descriptive themes
3. Generation of analytical themes addressing review questions

Confidence Assessment: Using GRADE-CERQual approach for qualitative evidence synthesis.

Mixed-Methods Integration

Convergent Synthesis: Parallel synthesis of quantitative and qualitative findings with integration matrix comparing themes and statistical results.

Assessment of Publication Bias

- **Funnel Plot Analysis:** Visual inspection for meta-analyses with ≥10 studies
- **Egger's Test:** Statistical assessment of small-study effects
- **Trim-and-Fill Analysis:** Estimation of missing studies impact

Confidence in Evidence

GRADE Assessment: Evidence quality rating (high, moderate, low, very low) based on:

- Risk of bias
- Inconsistency
- Indirectness
- Imprecision
- Publication bias

4. Discussion

Expected Findings and Implications

This systematic review is anticipated to provide the first comprehensive evidence synthesis on a critical but understudied occupational health issue in higher education. Expected findings include documentation of significant psychological impacts from abusive student feedback, identification of vulnerable faculty populations, and evaluation of emerging technological solutions for content moderation.

The review will likely reveal substantial heterogeneity in how institutions handle problematic student feedback, highlighting the need for evidence-based policy development. Findings may inform institutional review board considerations for evaluation system ethics, faculty support service design, and professional development programming focused on resilience and coping strategies.

Potential Limitations

Anticipated limitations include publication bias toward studies reporting significant effects, heterogeneity in outcome measurement approaches, and potential confounding from broader academic workplace stressors. The relative novelty of automated detection research may limit available evidence for technology assessment objectives.

Dissemination and Impact

Results will be disseminated through peer-reviewed publication, conference presentations at relevant higher education and occupational health venues, and policy briefs for institutional administrators. Findings will inform ongoing debates about evaluation system reform and contribute to developing evidence-based approaches to faculty well-being support.

References

Anderson, K., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, 27(2), 184-201. https://doi.org/10.1177/0739986304273707

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Chen, L., Zhang, R., & Wilson, C. (2020). Detecting toxic language in student feedback: A machine learning approach. *Computers & Education*, 157, 103-116. https://doi.org/10.1016/j.compedu.2020.103967

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, 31(1), 16-30. https://doi.org/10.1177/0273475308324086

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Hate speech detection with a machine learning approach. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 922-930. https://doi.org/10.1609/aaai.v31i1.10966

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 491-500. https://doi.org/10.1609/icwsm.v12i1.15077

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 1-13. https://doi.org/10.1177/2053951720943234

Heffernan, T. (2022). Abusive comments in student evaluations of courses and teaching: The experience of instructors. *Assessment & Evaluation in Higher Education*, 47(3), 409-420. https://doi.org/10.1080/02602938.2021.1910928

Kinman, G., & Jones, F. (2008). A life beyond work? Job demands, work-life balance, and wellbeing in UK academics. *Journal of Human Behavior in the Social Environment*, 17(1-2), 41-60. https://doi.org/10.1080/10911350802165478

Kokkinos, C. M., Stavropoulos, G., & Davazoglou, A. (2015). Development of an instrument measuring cyberbullying: Implications for intervention. *Psychology in the Schools*, 53(7), 756-770. https://doi.org/10.1002/pits.21948

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291-303. https://doi.org/10.1007/s10755-014-9313-4

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2), 535-566. https://doi.org/10.1093/jeea/jvx057

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation: A comprehensive guide for the scholarship of teaching. *Anker Publishing*.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1), 1-9. https://doi.org/10.1186/2046-4053-4-1

Nielsen, M. B., Notelaers, G., & Einarsen, S. (2012). Measuring exposure to workplace bullying. In *Bullying and harassment in the workplace* (pp. 149-174). CRC Press. https://doi.org/10.1201/EBK1439804896-9

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71

Peterson, D. A., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLoS One*, 14(5), e0216241. https://doi.org/10.1371/journal.pone.0216241

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education*, 3(3), 137-152. https://doi.org/10.1037/a0019865

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Schmidt, P. (2019). When student feedback turns personal and mean. *The Chronicle of Higher Education*, 65(22), A12-A15.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598-642. https://doi.org/10.3102/0034654313496870

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

Winefield, A. H., Gillespie, N., Stough, C., Dua, J., Hapuarachchi, J., & Boyd, C. (2003). Occupational stress in Australian university staff: Results from a national survey. *International Journal of Stress Management*, 10(1), 51-63. https://doi.org/10.1037/1072-5245.10.1.51

Zhang, S., Liu, X., & Chen, Y. (2018). Automatic detection of abusive language in online student evaluations using deep learning approaches. *Educational Technology Research and Development*, 66(4), 985-1002. https://doi.org/10.1007/s11423-018-9597-4