processes,. B but these procedures will always start with the instructor's review of the comments provided by 2.a.i.

- At a minimum, upon receiving the contextualized reflection, summary, or analysis, the supervisor will review the document along with the summary quantitative data from the SOISEI/SET.
  - If the supervisor finds the quantitative and qualitative summaries sufficient for this portion of the evaluation of teaching, they will complete their evaluation or report.
  - The supervisor may have additional questions or concerns which they feel require additional review of the text comments.
    - The supervisor should not make arbitrary decisions when reviewing full records of student feedback, however, so units should provide specific guidelines that help to establish when a supervisor should conduct more detailed review. These guidelines shall not limit the supervisor's ability to review an instructor's performance, but instead should prevent arbitrary patterns of decision making that are otherwise subject to conscious and unconscious bias.
      a. Unit charters should provide guidance advice concerning circumstances under which a supervisor may choose to access the full teaching evaluation record. These might include:
        i. Unexplained disagreement between the quantatitivequantitative scores and the summary of qualitative feedback.
        ii. Independent complaints or requests received by the supervisor that raise concerns about a course.
        iii. QuantatitiveQuantitative evaluations that fall below minimal standards (currently ranking average of 3.2 among the "seven dimensions score.")
        iv. A rotating system that prompts more detailed individual review every few years, at an interval not to exceed 4 years between reviews.
        v. Other criteria identified by the unit.

2. As part of the review process supervisors should document their rationale in accessing student feedback.

Units will identify a series of conditions under which a supervisor may access the full records for additional information. These may include:

- Unexplained disagreement between the quantatitive scores and the summary of qualitative feedback.
- Independent complaints or requests received by the supervisor that raise concerns about a course.

Quantatitive evaluations that fall below minimal standards (currently ranking average of 3.2 among the "seven dimensions score.")

A rotating system that prompts more detailed individual review every few years, at an interval not to exceed 4 years between reviews.

Other criteria identified by the unit.

In their request to review the full data, supervisors will document to CTL the reason for their data access from the list provided in the unit's charter.

If the supervisor has additional questions and undertakes more detailed review, they will then meet with the instructor to discuss the teaching evaluation full range of open response question data. This meeting will serve for the instructor and supervisor to address the questions of bias in qualitative data.

In completing their summary, the supervisor shall not "dip into" or "cherry-pick" the qualitative data to find narrative text to illustrate their summary. This is also true for any peer or mentor involved in these processes. In all cases, the initial contextualized examples will be used in summary.

The complete and raw data from student evaluations, both quantitative and qualitative, are part of the confidential personnel file of each instructor and will be handled accordingly.

Because the SOISEI/SET may not be used for more than 50% of the entire evaluation of teaching, each unit should clarify the proportional weight that will be given to the quantitative vs. qualitative data analyses and if/how they will be integrated by the supervisor.

The purpose of the evaluation of teaching is to support the improvement of pedagogical practice at the university. For this reason, efforts to study SOISEI/SET feedback may be drawn into both formative and summative assessments. The weighting system should reflect the emphasis on individual and systemic improvement.

When analyses of open response data are designed to be rapid reviews included only with the SOI/SET portion of the evaluation, analysis of open comments may be weighted no more than 10% of the entire evaluation.

When analyses of open response data are designed as intensive studies, and/or when these analyses are fully integrated with the unit's other mentoring or improvement processes, the qualitative data analysis may be incorporated within the entire evaluation of teaching effectiveness, such as including it in the peer-evaluation or peer-mentoring process defined in the charter. In this situation, the qualitative data shall not also be counted as part of the SOI/SET and no single source may be weighed more than 50% of the entire evaluation.

a.

- Units may choose to supplement SEI/SET data or entirely replace qualitative studies of open-ended question responses, choosing instead to have instructors incorporate additional use an alternate methods of quantitative and qualitative analysis. in placeinstead of the open response questions on the SETSOI/SET. This For example, units might may be done by gathering a subset of students and additional data by conducting a workshop or focus group, semi-structured interviews, or another qualitative research method. These studies These alternate assessments may be run by the instructor, s or by a peer mentor, or another outside facilitator as prescribed in the charter.
    - Units are encouraged to collaborate with the Jackson Center for Teaching and Learning on the development, design, and implementation of alternate evaluation plans.
    2. When alternate methods are undertaken, the open response questions will be removed from the end of term SOI/SET. Students will not be asked to answer survey questions when there is no plan to analyze their responses.
- When the narrative responses to the SETSOI/SET's open-ended questions are analyzed or considered, the analysis will be done by the instructor or administrator in a systematic manner with attention to the identified and potential biases in response and the context of the class/term. Such systematic consideration helps to prevent "cherry-picking" and impressionistic, cursory reviews which can give disproportionate weight to response outliers among the responses.
- When combining the results of multiple modes of evaluating teaching effectiveness, instructors and their supervisors should beware of biases, and particularly confirmation bias, during their integration of assessments. *It is particularly inappropriate to make a short review of raw qualitative data from open responses to find illustrations of a supposition or conclusion already drawn from quantitative data.* All types of qualitative review, including thematic reviews of SOI/SET open comments, peer evaluations, mentor evaluations, workshops, focus groups, and so on, must be used within their context. "Cherry picking" of illustrations is discouraged.

Implementation and Assessment

- In Fall 2022, Units will prepare their summary of bias document and the Charter modifications needed to add qualitative analysis of open response questions.
    - Units will forward these to the Senate and Provost by the 12th week of Fall 2022, as per Senate policy 2-22.
    - Senate and Administration will complete review and approval of plans before January 1st, 2023 immediately and approved plans will be implemented in reviewing Fall 2022 SOT/SEI .
    - Units unable to complete this process will adopt the practices defined above for Spring 2023Fall 2022 and will follow these procedures until charter revisions are approved.

i. As described above, Tthese practices will include: instructional units will maintain complete archives of responses to open response questions, along with numeric SOT/SOI data,supervisors will have access to all SET/SEI data for each instructor and course; instructors will prepare summaries of written comments for their supervisor at the end of each semester, examining both positive and negative comments and identifying potential course improvements; supervisors will review this summaries and when appropriate, review complete files of original feedback; and supervisors will use integrated quantitative and qualitative data in evaluative processes as otherwise outlined in the unit's charter and in compliance with guidelines about weighting.

- During AY 2022-2023, the University Senate's Academic and Instructional Policy Committee will develop metrics by which they can evaluate the effectiveness and impacts of this policy. Within two years, they will propose to the senate a method by which they will periodically review Michigan Tech's SOISEI-based evaluations of teaching.
- In no more than four years, the University Senate's Academic and Instructional Policy Committee will undertake a review of these procedures. They will gather critical feedback from their constituents, administrators, and Undergraduate and Graduate Student Governments. The committee will also conduct an updated review of peer-reviewed literature on the effectiveness of MTU's SOISEI/SET system.

Appendix A: Additional Discussion of SEI/SETs and ~~Discussion of~~ Recommendations Concerning Baseline Standards

I. Additional Discussion of SEI/SETs

SEI/SETs are the most common form of data collection in the United States for the evaluation of student satisfaction with their courses. Common surveys, such as the Student Evaluation of

Educational Quality (SEEQ) are in wide use around the world, collecting both rank data from closed questions (such as Likert even-point scale) and open-response questions that solicit narrative responses. These are not necessarily the best method for collecting data, but these surveys remain very common because they provide easily quantifiable rank data suitable for cost-effective analysis.

Surveys include numerical evaluations of student replies on issues presumed to be significant to their experiences, but because those general questions necessarily lack course-specific nuance, most survey instruments also include open comments where students can provide detailed information (see discussions in Harvey 2011). Researchers often find student responses to open questions contrast to the generally satisfactory evaluation in closed questions (such as Likert scale rankings) on the same survey responses. Students use open form comments to give specific suggestions for course changes or to identify issues they feel the closed questions failed to adequately address. Researchers attribute students' tendency to emphasize negative feedback in written comments to students' feeling that the survey design neglected to consider their perspectives on appropriate improvements. As a consequence, university students sometimes feel indifferent toward the SEI/SET or consider the process to lack legitimacy. Student feelings about the SET instrument have demonstrated impact upon the rankings and evaluations they provide through them (Suárez, Gómez Suárez, & Paredes 2022; Johnson 2012)

SEI/SET data is also complicated because both the instrument and the data produced through it can be subject to various types of response and non-response biases, as detailed below. A great deal of research literature shows the impact of various types of biases and generally advises on how to avoid or minimize their harm. Among the most significant biases of concern for SEI/SET at Michigan Tech are "prestige and stereotype response biases" and "threat or hostility biases," both of which produce data reflecting rankings that favor men over women; white and native-born persons over persons of color and those who speak with ESL accents, and other identities.

Open response questions provide more complicated challenges for interpreting biases. Students sometimes misunderstand the purpose of the SEI/SET instrument, adding open response comments they expect to be read by other students considering a particular class, by their classmates, by university administrators, or for whom the audience is otherwise unclear. As with any forum built around anonymous messaging, inappropriate comments also occur.

Scholars have written a great deal about SEI/SET instruments and their uses and are increasingly critical of their validity for any summative assessment (Esarey and Valdes 2020). SEI/SET instruments are found to be useful for formative assessment, as the instrument can provide information instructors can use to improve classroom design and pedagogy. Researchers are increasingly critical of SEI/SET use in summative assessments. Scholars increasingly find more value in measurements of demonstrated learning outcomes.

~~Units use data from SETs for several purposes beyond the formative and improvement-oriented flow of information from student-to-instructor. These uses are also defined in each unit's charter. SET results are currently used as part of *summative* assessment processes for instructor~~

~~promotion, tenure, and reappointment, for the allocation of merit raises, as evidence of classroom innovation, and other areas of professional activity.~~

II. Additional Discussion of Baseline Standards

1. Regarding the identification of potential bias in student evaluation of teaching:

a. When the narrative responses to the SEI/SET's open-ended questions are analyzed or considered, the analysis ~~will~~should be done by the instructor or administrator in a systematic manner with attention to the identified and potential biases in response and the context of the class/term. Such systematic consideration helps to prevent "cherry-picking" and impressionistic, cursory reviews which give disproportionate weight to response outliers.

b. When combining the results of multiple modes of evaluating teaching effectiveness, instructors and their supervisors should beware of biases, and particularly confirmation bias, during their integration of assessments. It is particularly inappropriate to make a short review of raw qualitative data from open responses to find illustrations of a supposition or conclusion already drawn from quantitative data. All types of qualitative review, including thematic reviews of SEI/SET open comments, peer evaluations, mentor evaluations, workshops, focus groups, and so on, must be used within their context.

c. In completing their summary, the supervisor ~~shall~~should not "dip into" or "cherry-pick" the qualitative data to find narrative text to illustrate their summary. This is also true for any peer or mentor involved in these processes. In all cases, the initial contextualized examples ~~will~~should be used in summary.

d. Units may want to examine how conscious and unconscious biases are likely to influence SEI/SOT data in the discipline(s) and communities of practice among their members. Toward this end, units might develop a written summary of these findings that unit members, including administrators, would consult whenever measures of teaching effectiveness are under consideration (e.g., for promotion, tenure, and reappointment; merit raises; peer mentoring, etc.).

2. Advice ~~R~~regarding the weighting of components of teaching evaluations:

a. When analyses of open response data are designed to be rapid reviews included only with the SEI/SET portion of the evaluation, you might consider weighting analysis of open comments at a lower value, e.g., no more than say 10% of the entire evaluation.

b. When analyses of open response data are ~~designed~~included as part of intensive studies, and/or ~~———~~when these analyses are fully integrated with the unit's other mentoring or improvement processes, the qualitative data analysis ~~may~~should be ~~incorporated~~considered ~~within the entire evaluation of teaching effectiveness, such as including it in the peer-evaluation or peer-mentoring process defined in the charter. In this situation, the qualitative data would be counted~~ as part of the SEI/SET portion of the evaluation of teaching, and no single source may be weighed more than 50% of the entire evaluation.

Appendix B: Alternate Qqualitative Aassessment Mmodels:

All units must utilize the baseline guidance provided by the Senate for incorporating student comments in the evaluation of teaching. However, uUnits may additionally adopt more formalized or detailed standards of analyses for the qualitative responses to open ended questions, that go going beyond the initial a review and contextualized summary. Such reviews require more commitment of time and resources, but if designed judiciously, can provide useful information. Examples of acceptable methods include:

a. Thematic Analysis of open response questions from SOISEI/SET:
   i. Reviewing the comments to identify themes for analysis. These themes might arise from the responses or could be predetermined (accessibility/communication, preparedness/organization, topical/subject mastery, assignment designs, inclusive/welcoming, etc.…).
      i.1. Inappropriate or biased comments should be removed from analysis at this stage in the process, before detailed or semi-quantitative analyses are undertaken.
   ii. Coding With identified themes, instructors can each response to eachthen count each mention of that theme question as positive, neutral, or negative. , orInstructors could use more detailed ordinal scales as appropriate for their study. with a more detailed ordinal scale.
      iii. Then identifying themes for the analysis from the responses or using predetermined themes (accessibility/communication, preparedness/organization, topical/subject mastery, assignment designs, inclusive/welcoming,…).
      iv. Following the process of identifying themes, each comment can then also be counted when the reviewer's comment includes information on a theme.
   iii. These results can then be compiled to reveal patterns in the comments:
      1. What issues concerned the majority of students in this class?
      2. , and these may be further brokenHow do responses fall for each issue as out by overall positive/negative/neutral classifications?.
      3. Do these patterns indicate improvement or benefits from patterns observed in previous semesters?
      4. Do these patterns vary by other demographic or other factors? (Such as gender, major, class standing, etc., assuming the additional anonymized data is collected)

i.5. *This For larger classes*, such analyses will can be helpful to identify meaningful patterns where many critical students identify similar common themes vs. positive students emphasize others. This can be used to provide additional context for understanding comments and numerical data from the SOISEI/SET. Examples include: Variations in response patterns for first year vs. second year students, responses among students affiliated with different colleges, or those who identify in different demographic communities.

iv. Other eExamples can be found in Zakrajsek (2019), *Analyzing Student End of Course Written Comments*. include link, link, link.

ii.

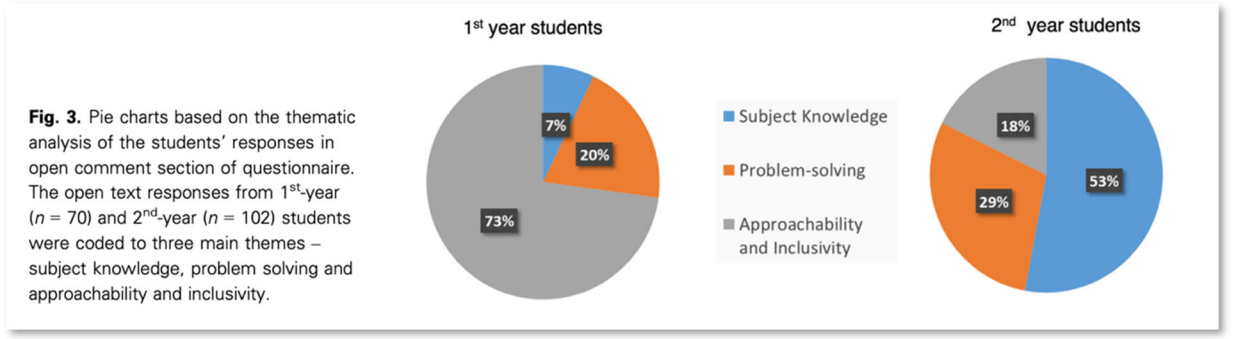

Fig. 3. Pie charts based on the thematic analysis of the students' responses in open comment section of questionnaire. The open text responses from 1st-year (n = 70) and 2nd-year (n = 102) students were coded to three main themes – subject knowledge, problem solving and approachability and inclusivity.

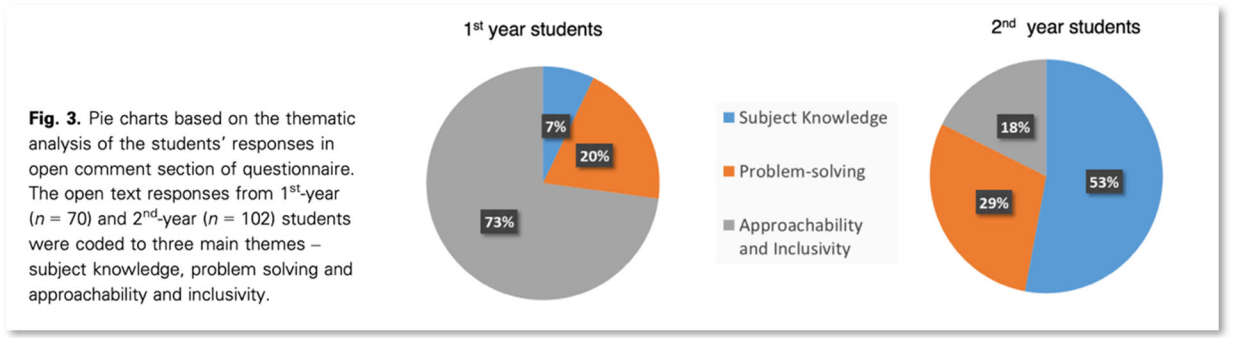

Fig. 3. Pie charts based on the thematic analysis of the students' responses in open comment section of questionnaire. The open text responses from 1st-year (n = 70) and 2nd-year (n = 102) students were coded to three main themes – subject knowledge, problem solving and approachability and inclusivity.

Figure 1. Sample of pie charts with basic descriptive summaries of themes identified in student responses to open questions from a bioscience course, examining the difference in first- and second-year students in their concerns about subject knowledge, problem-solving, and approachability. Such charts could be developed by instructors, and examined along different axes, such as overall positive vs. overall negative comments. Such charts illustrate how coding themes can be used to identify important topics among open comments. Charts from Awais and Stollar (2021:2897).

Sample of pie charts with basic descriptive summaries of themes identified in student responses to open questions from a bioscience course, examining the difference in first- and second-year students in their concerns about subject knowledge, problem-solving, and approachability. Such charts could be developed by instructors, and examined along different axes, such as overall positive vs. overall negative comments. Such charts illustrate how coding themes can be used to identify important topics among open comments. Charts from Awais and Stollar (2021:2897).

~~. Responses that are personal and/or unkind should be set aside from analysis. This should include responses that remark on an instructors physical characteristics or cultural background or other attributes of identity.~~

~~5. When combining the results of multiple modes of evaluating teaching effectiveness, instructors and their supervisors should beware of biases, and particularly confirmation bias, during their integration of assessments. It is particularly inappropriate to make a short review of raw qualitative data from open responses to find illustrations of a conclusion drawn from quantitative data from SET results. All types of qualitative review, including thematic reviews of SET open comments, peer evaluations, mentor evaluations, workshops, focus groups, and so on, must be used within their context. "Cherry picking" of illustrations is discouraged.~~

~~6.~~

~~1)~~ **Appendix C~~B~~: ~~5) :~~ Explanation of T~~t~~ypes of B~~b~~ias I~~i~~dentified by R~~r~~esearchers in ~~SO~~SEI/SET L~~l~~iterature~~.~~**

Discussion of t~~T~~ypes of bias in ~~student teaching evaluation~~SEI/SET responses ~~(often names Student Evaluations of Teaching (SET)~~ are common in published research. These biases can be identified in both numerical and "free response" textual responses. There are two main types:

a. **Non-Response Bias**: These biases arise when sets of the survey audience do not respond to (or engage with) the survey. In student evaluation situations, this might be caused among populations of students that are very busy in the final weeks of the semester and feel they cannot spare the time to respond, those who feel the survey has no real purpose or will not be used to constructive effect, or those who feel their voice is not valued by those who review the data. Non-response bias may be different for closed question rankings vs. open-ended survey prompts that require thoughtful narrative replies.

b. **Response Bias**: Researchers have shown many different ways that surveys can bias the responses of those surveyed, above and beyond the codification or amplification of existing social prejudices. This can vary from subtle effects that skew numbers toward more positive or negative rankings to survey participants consciously or unconsciously providing false information (or half-truths) which inject results with bad information that leads to bad conclusions.

   As detailed below, these biases are not necessarily intentional lies, nor does their existence mean that survey responses essentially have no value. Some of these biases are the result of poorly phrased questions, the survey format or situation during data collection, fatigue or boredom among respondents, or many other issues. The evaluations of instruction at Michigan Tech are designed and instituted by the staff at CTL using instruments and processes in a manner intended to collect good data. The processes are periodically reviewed by the University Senate and administrative offices including Michigan Tech's Office of Diversity and Inclusion. Our community seeks to identify sources of bias in our instruments, eliminate them, and mitigate those that cannot otherwise be resolved.

Biases cannot be eliminated entirely from any survey-based study. Those who use survey data, including instructors and their supervisors, must understand that biases cannot be entirely eliminated by design. Interpretation of teaching evaluation data, both numerical and "free-form" text responses must be evaluated with critical eye toward the context of the survey.

Subtypes of response biases (sometimes also called cCognitive bBiases):

a. **Acquiescence or Agreeability Bias**: conscious or unconscious effort to be polite and/or likable, so they agree with the survey questions. This bias results in inflated ranks in closed question evaluations.
b. **Demand Characteristics Bias**: subconscious or conscious adjustment of responses to fit perception of purpose of the experiment. This bias plays a role because students do not have a clear understanding who (if anyone) reads the numerical data or the written comments that they contribute. See Sponsorship Bias below.
c. **Extreme Responses:** survey responses by those people who give answers that are either extremely positive or negative. While it is possible for students to decide that everything about a class was uniformly "excellent" or "poor," it is also possible that these reviewers have not provided thoughtful critique.
d. **Neutral Responses**: Some survey respondents give only "middle of the road" responses, choosing the center of the Likert scale, for example. As with extreme responses, these may be the result of a reviewer simply filling in bubbles instead of providing meaningful feedback.
e. **Social Desirability or Conformity Bias:** People tend to give answers that they think the readers of the survey will find useful and important and that the surveyor will then think of the respondent as a reasonable and "desirable" member of the community. This is an unlikely bias for respondents to surveys that are conducted anonymously.
f. **Question Order Bias**: The order in which questions are asked can lead to answers that have more disparate results (contrasting effect) or similar results (assimilation effects). As a relevant example, asking people about their satisfaction with specific services *first* and then about overall satisfaction results in higher overall satisfaction results (Thau et al. 2020). Using the reverse order in a survey, by contrast, yields results showing lower rates of overall satisfaction when the only change in the survey was shifting the order of questions in the study population. The effect is clear because there should have been no difference in the overall average rates of satisfaction, but changing the order of questions produced that effect.
g. **Mindset/Carry-over Effects Bias**: Survey respondents can carry negative or positive feelings evoked from one question into their response to the question that follows. This could be particularly sensitive in the transition from the closed questions to the open-question sections of the SET.
h. **Prestige Bias**: Respondents will modify their responses to a survey to "round up" or "round down" their assessments based upon the "prestige" of the subject of the

survey. As examples, survey respondents will round up when estimating the income of male- vs. female-presenting persons. This can be realized in different ways in academic reviews, along lines of gender, race, ethnicity, ability, sexuality, and other intersectional identities, as well as by disciplinary lines.

i. **Threat and Hostility Bias:** When respondents are thinking about unpleasant things, feeling hostile, or recalling difficult or bad experiences, they will consequently emphasize negative rankings or feedback.
j. **Sponsorship Bias:** Survey respondents will shift their evaluations based upon the persons or organizations sponsoring a survey. This is perhaps most relevant to the SET process because students question the usefulness and purpose of teaching evaluations, both numerical and—very specifically—written feedback. While not many published articles engage this question in SET processes, one recent study showed that students who believe that SETs are valued are more likely to respond to surveys and more likely to provide higher evaluation scores. The authors speculated that their perception of professors teaching competence may also be influenced by their perception of their own role as evaluators of university professors. This study further reported that students in this study generally doubted that professors use students' open response suggestions in course improvement and that their opinions varied on whether or not SET results (written or numeric) should be included in ~~professors~~professor's promotion, reappointment, and tenure decisions (Spooren and Christiaens 2017). Notably, this study relied upon a survey of student opinions that did not include or examine open response questions.
k. **Stereotype Biases:** asking about biographical information, such as gender, race, technical ability/major, or other questions of identify can prime respondents to shift their evaluative rankings or shape comments in different directions. This bias effect seems to be true, no matter the identities of the student completing the SET or of the instructor being evaluated, although the directionality of the bias's effect on rankings is difficult to predict.
l. **Motivated Forgetting Bias:** Because memories as very malleable, people tend to shape memories to fit their current beliefs, contexts, or feelings. They may recall events happening more recently or longer ago than reality, or they may confuse the order of events (Kjellsson, Clarke, & Gerdtham 2014). The implication here is that the regular cycle of events during a semester can have the same type of bias effect as major historical or cultural events at the end of the academic semester.

c. **Confirmation Bias**: This bias occurs post-survey, in analysis rather than in survey design, and is a major concern in the misuse of survey data. This bias occurs when a researcher or evaluator seeks to illustrate or prove a point that they believe to be true. As an example, if a supervisor were to form expectations of teaching performance based upon numerical rank data from a group of SET results, then make a quick review of open response question answers to find illustrations of those problems (or successes). Such an action ~~would be cherry picking~~would-be cherry-picking information to confirm an expectation, while neglecting examine the context of responses. Open response question answers have been shown to differ substantively in tone and enthusiasm from the ranking

reviews of Likert-scale closed questions. Such casual review of qualitative data is to be avoided as bad practice.

~~Study examples and methods:~~

1. ~~One example provides a detailed examination of both Likert-scale style responses to numerical survey and thematic analysis of open question text responses in a bioscience setting. The open text replies were coded to examine three main themes: subject knowledge, problem solving, and accessibility/inclusivity. Within these categories, each response was coded as positive or negative (and presumably not coded if neutral or absent). Study quantified percentage of students that provided feedback in on each of the three theme areas, both positive and negative, while listing popular examples from the examples comments (Accessibility/Inclusivity>>"Was approachable"). The patterns among positive and negative comments were then examined by student cohort (1st vs. 2nd year students in the same program). This study shows detailed analysis that joined both student teaching evaluation and instructor and TA self-evaluations used together to assess learning experiences. Data analysis is presented as pie charts to show proportion of responses concerning different themes (to represent student priorities in response).~~

Awais, R., & Stollar, E. (2021). Demonstrator training needs to be active and focused on personalized student learning in bioscience teaching laboratories. *FEBS Open bio*, *11*(11), 2888-2901. *FEBS Open Bio*. Awais, https://febs.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/2211-5463.13299

Bargh, John A., Mark Chen, and Lara Burrows. (1996). "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action." *Journal of personality and social psychology* 71(2):230. https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Social_Cognition/Bargh_et_al_1996_Automaticity_of_social_behavior.pdf

Clayson, D. E. (2020). *A Comprehensive Critique of Student Evaluation of Teaching: Critical Perspectives on Validity, Reliability, and Impartiality*. Routledge.

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, *45*(8), 1106-1120. http://justinesarey.com/teacher-evaluation-decisions.pdf

Johnson, D. M. (2011). Teaching effectiveness as measured by student evaluation of teaching: an empirical study. *International Journal of Information and Operations Management Education, 4*(3/4), 212-228. http://doi.org/10.1504/IJIOME.2011.044564

Kjellsson, G., Clarke, P., & Gerdtham, U. G. (2014). Forgetting to remember or remembering to forget: a study of the recall period length in health care survey questions. *Journal of health economics*, *35*, 34-46. https://www.sciencedirect.com/science/article/pii/S0167629614000083

Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*, 1-12.

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535-566. https://www.econstor.eu/bitstream/10419/170984/1/dp11000.pdf

Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in educational evaluation*, *54*, 43-49. https://www.sciencedirect.com/science/article/pii/S0191491X16300256?casa_token=orndmqDbEQoAAAAA:zWm_pCMQE1XHV-FhMlt59kbwm6TZTxQo_jBbjisNDqR78sYBjSZqhNFsNviH_G43q2nw50c

Spooren, Pieter, Frederic Vandermoeren, Raf Vanderstraeten, and Koen Pepermans. 2017. "Exploring High Impact Scholarship in Research on Student's Evaluation of Teaching (SET)." Educational Research Review 22: 129-41.

Stroebe, Wolfgan. 2020. "Student Evaluation of Teaching Encourages Poor Teaching and Contributes the Grade Inflation: A Theoretical and Empirical Analysis." Basic and Applied Social Psychology 42(4): 276-94.

Suárez Monzón, N., Gómez Suárez, V., & Lara Paredes, D. G. (2022). Is my opinion important in evaluating lecturers? Students' perceptions of student evaluations of teaching (SET) and their relationship to SET scores. *Educational Research and Evaluation*, *27*(1-2), 117-140.

Thau, M., Mikkelsen, M. F., Hjortskov, M., & Pedersen, M. J. (2021). Question order bias revisited: A split-ballot experiment on satisfaction with public services among experienced and professional users. *Public Administration*, *99*(1), 189-204. https://onlinelibrary.wiley.com/doi/abs/10.1111/padm.12688

Uttl, Bob, Carmela A. White, Daniela Wong Gonzalez. 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.

Uttl, Bob, Kelsey Cnudde, and Carmela A. White. 2019. "Conflict of Interest Explains the Size of Student Evaluation of Teaching and Learning The many harms of SETs in higher education 309 Correlations in Multisection Studies: A Meta-Analysis." PeerJ 7: e7225. http://doi.org/10.7717/peerj.7225.

Zakrajsek, Todd. June 26, 2019. "Analyzing Student End of Course Written Comments." https://www.scholarlyteacher.com/post/analyzing-student-end-of-course-written-comments

Awais, R., & Stollar, E. (2021). Demonstrator training needs to be active and focused on personalized student learning in bioscience teaching laboratories. *FEBS Open Bio*. Awais,

Thau, M., Mikkelsen, M. F., Hjortskov, M., & Pedersen, M. J. (2021). Question order bias revisited: A split-ballot experiment on satisfaction with public services among experienced and professional users. *Public Administration*, *99*(1), 189-204. https://onlinelibrary.wiley.com/doi/abs/10.1111/padm.12688

Kjellsson, G., Clarke, P., & Gerdtham, U. G. (2014). Forgetting to remember or remembering to forget: a study of the recall period length in health care survey questions. *Journal of health economics*, *35*, 34-46.

Bargh, John A., Mark Chen, and Lara Burrows. (1996). "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action." *Journal of personality and social psychology* 71(2):230.

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535-566.

Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in educational evaluation*, *54*, 43-49.

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, *45*(8), 1106-1120. http://justinesarey.com/teacher-evaluation-decisions.pdf