Investigating the usage of Likert-style items within Computer Science Education Research Instruments

Brian M. McSkimming
Department of Engineering Education
University at Buffalo
Buffalo, NY, USA
bmm6@buffalo.edu

Sean Mackay
Department of Engineering Education
University at Buffalo
Buffalo, NY, USA
snmackay@buffalo.edu

Adrienne Decker
Department of Engineering Education
University at Buffalo
Buffalo, NY, USA
adrienne@buffalo.edu

***Abstract*— One of the most ubiquitous techniques for evaluating research, particularly in educational settings, has been Likert-style questionnaires. Having a participant rank their level of engagement, agreement, or interest on a scale can provide powerful insight into an individual's perspective and attitude. However, as computing education matures as a discipline it becomes important for us to examine our practices and ensure that we are employing the techniques of evaluation properly. Likert-style questionnaires can be prone to unintentional biases and noise which, from the perspective of the researcher, may affect the study in unknown, unexpected, and potentially undesirable ways.**

In this research we seek to aid the computing education researcher not only avoid biases and noise but also improve the reproducibility of their work. First, we establish best practices for these instruments by synthesizing recommendations from the original creator, Rensis Likert, the Center for Disease Control (CDC), the Association of American Medical Colleges (AAMC), and the American Association for Public Opinion Research (AAPOR). We then considered additional sources of unintentional biasing and noise resulting from the measurement scales/response options, specifically possible biasing resulting from how response options are presented to the study participant.

With these recommendations, we then examined 121 evaluation instruments used in computer science education research and curated in the csedresearch.org online database to see how often researchers unintentionally fall victim to the pitfalls of ambiguity, awkward phrasing, use of conjunctions, leading or biased statements, and double negatives. We found that the occurrence of at least one of these problematic statements to be in 82.6% of all instruments. We also examine demographic information for the intended study participants, number of response options, and how those options are presented. Overall, while we see many instrument authors falling victim to a few of these common pitfalls, we believe that increased awareness of potentially problematic statements/measurement scales and their impact on research bias and reproducibility will help insulate computing education researchers from avoidable complications and strengthen the discipline throughout.

Keywords—Likert, assessment, survey, survey creation, bias, fallacy, instruments, best practices

I. Introduction

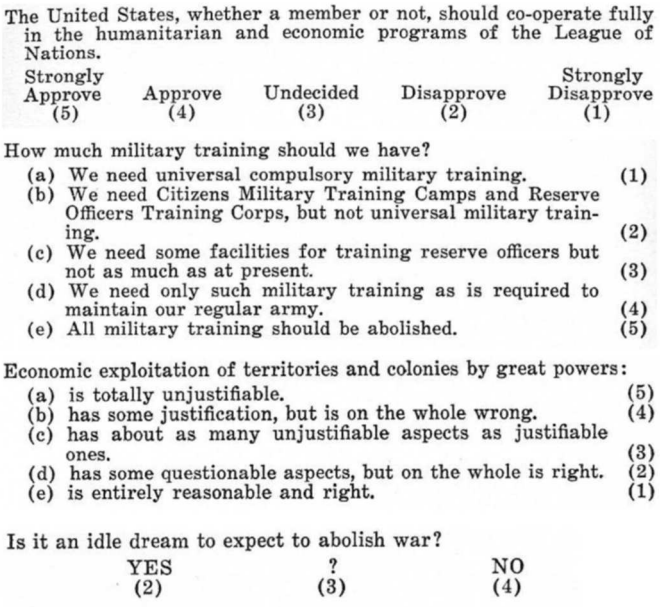
If you have ever rendered your "level of agreement with a statement", you have taken a Likert-style survey or scale. This technique for allowing participants to represent their attitudes about different statements has been widely adopted by many researchers to study phenomena in psychology, social science, marketing, and education, including computing education. However, what is important to realize about this style of questionnaire is that care must be taken when creating the prompts for the participants or the results of the survey can be exposed to unintended biases. As computing education continues to grow, the field is faced with the task of measuring and studying phenomena that lend themselves to Likert-style scales. Therefore, it is important to understand what characteristics prompts on these types of questionnaires should embody and what types of characteristics should be avoided.

II. Background

In 1932, Rensis Likert published a description of a technique for the measurement of attitudes and demonstrated how a scaled scoring system could provide a simple, fast, and perhaps most importantly statistically analyzable measurement [8]. Likert's method of attitude measurement was presented as an alternative to the technique reported by Louis Thurstone. Thurstone's method, derived based upon his law of comparative judgement, also involves assessing an attitude based upon a series of statements. However only two options (agree & disagree) are presented, and a "judge" weights each statement based upon its strength with regard to the attitude being evaluated. [13]

In response, Likert proposed a technique for the measurement of a general attitude on a topic utilizing a series of statements as well as open-ended and multiple-choice style questions which when scored by a participant on an interval scale could implement common statistical techniques in the analysis [8]. Now, nearly 90 years later, Likert and Likert-style (Likert-type) surveys and questions are still used to probe another person's attitude or perception on virtually any topic. Examples of various Likert-style prompts from the original paper are presented in Fig. 1.

This material is based upon work supported by the U.S. National Science Foundation under Grant Nos. 1933671 & 1625335.



The United States, whether a member or not, should co-operate fully in the humanitarian and economic programs of the League of Nations.

Strongly Approve (5)	Approve (4)	Undecided (3)	Disapprove (2)	Strongly Disapprove (1)

How much military training should we have?

(a) We need universal compulsory military training. (1)
(b) We need Citizens Military Training Camps and Reserve Officers Training Corps, but not universal military training. (2)
(c) We need some facilities for training reserve officers but not as much as at present. (3)
(d) We need only such military training as is required to maintain our regular army. (4)
(e) All military training should be abolished. (5)

Economic exploitation of territories and colonies by great powers:

(a) is totally unjustifiable. (5)
(b) has some justification, but is on the whole wrong. (4)
(c) has about as many unjustifiable aspects as justifiable ones. (3)
(d) has some questionable aspects, but on the whole is right. (2)
(e) is entirely reasonable and right. (1)

Is it an idle dream to expect to abolish war?

YES (2)	? (3)	NO (4)

Fig. 1. Example prompts from Likert's Original Paper [8]

In a series of papers and studies surrounding the heuristics used by people in decisions involving uncertainty, Tversky and Kahneman describe some of the systematic errors which result from intuitive inferences [14, 15]. One of the errors described has been coined as the conjunctive fallacy, and results from the tendency for people to allow their intuitive judgments to violate extensional laws of probability. The iconic presentation of the conjunctive fallacy involves a description of an individual and a series of statements which study participants were asked to rank based upon their likelihood. Included in that list of statements are items which strongly do not agree with the description as well as a statement which conjoins the unlikely statement with one which is highly likely. According to probabilistic laws, the probability of a conjunction must be less than or equal to the probability of either of its parts, i.e. $P(A\&B) \leqslant P(A)$ and $P(A\&B) \leqslant P(B)$. Among the participants of their study, 85% violated the conjunction rule's probability regardless of their training in probability and statistics, with naïve participants exceeding 90% violation [14].

Within computing education, there are a number of studies that have been conducted on evaluation and assessment instruments within the field [7, 9, 10]. However, each of these studies were concerned with holistic characteristics of the instrument (i.e. constructs measured, validity, types of analysis employed). None of them look at the instrument contents in aggregate to look for issues of soundness within the questions.

As the computer science education community strives to embrace and adopt rigorous research practices the ability to reproduce, replicate, compare, and conduct meta-analyses rely upon high fidelity and low-noise representative data. Better data enables better studies which catalyzes future research endeavors and understandings reliant upon today's explorations and discoveries. It is with this consideration in mind that this study was undertaken to provide researchers additional insight into possibly unexpected biasing present in Likert-style items.

This study focuses on identifying some of the features within commonly used evaluation instruments which have the potential to skew results due to innately human characteristics, specifically the tendency to rely upon heuristics when faced with uncertainty. This study is intended to highlight some of the ways responses to Likert-style items may be skewed.

The overarching research question guiding this work was: What biases evaluation instruments in unintentional or unexpected ways? From this, we developed two research questions that are the focus of this study:

- **RQ1: What are the commonly recommended best practices for composing Likert-style statements that are important to computing education research?**
- **RQ2: To what extent have the instruments used in computing education research previously violated these best practices?**

III. Best Practices & Pitfalls

To answer RQ1: *What are the commonly recommended best practices for composing Likert-style statements that are important to computing education research?*, we identified a set of recommendations from medicine: the Centers for Disease Control and Prevention (CDC) [6], medical education: the Association of American Medical Colleges (AAMC) [3], and the leading professional organization of public opinion and survey research in the United States: the American Association for Public Opinion Research (AAPOR) [2]. Combining these recommendations with the original suggestions from Likert [8], we focused upon the considerations which would be most relevant to computing education research.

In this section, we summarize the possible pitfalls as recognized by the recommendation documents and briefly consider how uncertainty heuristics could affect participant responses to statements which succumb to each pitfall.

One of the challenges and difficulties encountered when preparing Likert-style prompts is interpretation. When a study participant encounters statements which cause them to experience uncertainty, they may rely upon their set of personal heuristics for interpretation. [14] As such, minor variations in statements have the potential to change not only their direct meaning but may cause the survey-taker to provide responses to unintended prompts. In an attempt to avoid these situations, Likert emphasized "the necessity of stating each proposition in clear, concise, straight-forward statements." [8]

As such, to achieve intended research results when using surveys containing Likert-style statements the recommendations from the CDC, AAMC, and AAPOR emphasize avoiding ambiguous, awkward, double-barreled, and leading statements.

A. Ambiguous Statements

Likert specifically cautions against statements which contain any manner of ambiguity, recommending that each proposition is stated in such a way that "persons of less understanding than any member of the group for which the test is being constructed will understand and be able to respond to the statements." [8] Any ambiguity in a Likert-style prompt needs to be avoided to minimize confusion and maximize the possibility of accurate

assessment of the participant's attitude and opinions on the subject being tested.

The concern with ambiguity extends beyond the possibility of two participants having different interpretations of the ambiguous statements. When the respondent is faced with attempting to provide a single answer to a question with multiple interpretations, there is a tendency to rely upon heuristics and priming for their responses. [14] When a participant responds in this way, they may not be providing answers which are an accurate measurement of their attitude or beliefs.

B. Awkward Statements

Awkward statements are those which are difficult to understand often due to a malformed sentence and/or poor sentence structure. While the prompt's meaning may not be ambiguous, the statement itself is awkward for the participant to understand and answer.

Similar to ambiguous statements, awkward statements force the survey participant to attempt to interpret what the prompt is asking and therefore may be uncertain and reliant upon unintentional heuristics.

C. Conjunctive (Double-barreled) Statements

In contrast to ambiguous statements which suggest multiple interpretations, double-barreled statements are composed of multiple things being asked in a single prompt. These may take the form of an explicit conjunction using words such as "and" or they may be more subtly implied. According to the recommendations from the CDC, a double-barreled question "is a question that touches upon more than one issue, yet allows for only one answer." [6] The AAMC suggestions recommend that "survey items should address one idea at a time." [3]

This type of statement was the direct focus of Tversky's & Kahneman's work on the conjunctive fallacy [15], made evident through exploration of an individual's inability to properly assess the probability of a situation involving two conjoined sets.

D. Leading/Biased Statements

Leading statements are those in which the statements themselves introduce bias which may influence the way a respondent answers the question. [6] As noted within the AAPOR's recommendations leading statements also include phrases which lead "the respondent by suggesting the position or stance of an authority with which it might be difficult for the respondent to disagree." [2]

IV. METHODS/METHODOLOGY

To answer RQ2: *To what extent have the instruments used in computing education research previously violated these best practices?*, we initiated a systematic inspection of the evaluation instruments curated in the https://csedresearch.org database. For each instrument with Likert-style questions, we collected the following data:

- Target Demographic (as reported in the database)
- Total number of Likert-style statements/prompts
- The individual statements from the surveys (to be coded for failure to use best practices)
- Number of Likert-style options given for each statement/prompt
- Ranking direction of Likert-style options (Negative to Positive, or Positive to Negative)

Not all instruments which were listed in the database had their questionnaires locally stored on https://csedresearch.org. For those instruments, internet search engines were employed using the instrument title as listed in the database. When found, demographic information was verified to ensure that the instrument found was the same as the one referenced by csedresearch.org.

In section 3, we described the best practices we were most interested in investigating for computing education research instruments. To determine to what extent scales violated each of the best practices, each statement from each instrument was considered independently of all the other statements. Specifically, each statement was pasted into a row of a spreadsheet, considered by a researcher, and coded as having or not having:

- Ambiguous wording
- Awkward wording
- Biased (or leading) wording
- Conjunctions
- Double negative statements

Three coders worked on the coding of the data. The coders had varied experience and backgrounds, with one being a first-year graduate student, one being a postdoctoral researcher five years removed from their graduate work, and the last serving as faculty advisor to the other two coders. The graduate student's background is Computer Science, the postdoctoral researcher's background is in Engineering, and the faculty coder holds an appointment in Engineering Education with twenty years' experience in Computing Education. There was a formal training and discussion session held where the definitions of the types of pitfalls documented by the CDC, AAMC, AAPOR, and Likert [1-3, 5] were read and discussed until all parties had the same working definition of the kinds of statements that would fall into this pitfall category. A sample (10) from the set of statements to be coded were also looked at together to determine if they contained any of the pitfalls and allowed for further discussion of how to identify the pitfalls in future statements. Much of the discussion amongst the coders focused on considerations of what could be coded as ambiguous statements and how to code such statements. After that point, statements were independently coded by the three coders.

In an effort to establish interrater reliability, 410 randomly chosen statements (every 10th statement in the dataset) was checked by both coders. The agreement for the coding of these

statements is presented in Fig. 2, indicating a sufficient level of interrater reliability in the coding. It may be interesting to note that the greatest variation in coding occurred with statements considered to be ambiguous, accounting for more than 55% of the disagreeing statements.

	Ambiguous	Awkward	Conjunctive	Leading	Double-Negative
Number of Coding Disagreements	49	13	25	2	0
Percentage of Coding Agreements	88.1%	96.8%	93.9%	99.5%	100.0%

Fig. 2. Interrater coding reliability counts and percentages

V. RESULTS

Of the 190 instruments listed in the https://csedresearch.org database (as of January 2021), 130 were identified as being composed of and/or including Likert-style questions. Of those 130 instruments listed, 121 were found, examined, and comprised the data set for this study. The remaining nine instruments were not able to be found through internet searches or were behind paywalls and not accessible. Within these instruments, there were 4098 total statements analyzed.

A. Target Demographic Results

The demographic breakdown of the instruments as reported in the online database with statement count is presented in Fig. 3. Due to some of the instruments reporting that they were intended for varied audiences the total number of instruments and statements exceeds what was reported in this study.

Demographic (U.S. Grades)	Instruments	Statements
Elementary (K-5)	11	228
Middle (6-8)	33	1156
High (9-12)	32	1185
Undergrad	38	1035
Grad	8	183
Teacher/Adult	33	1045

Fig. 3. Demographic Breakdown of Instruments Studied

B. Ambiguous Statement Results

The total number of ambiguous statements coded across all the evaluation instruments was 62, accounting for 1.5% of the 4098 total Likert-style statements. Of the 121 instruments evaluated, 91 (75.2%) contained 0 conjunctive statements, and no instrument contained more than 5 statements. Fig. 4 shows the breakdown of the number of conjunctive statements across all 121 instruments.

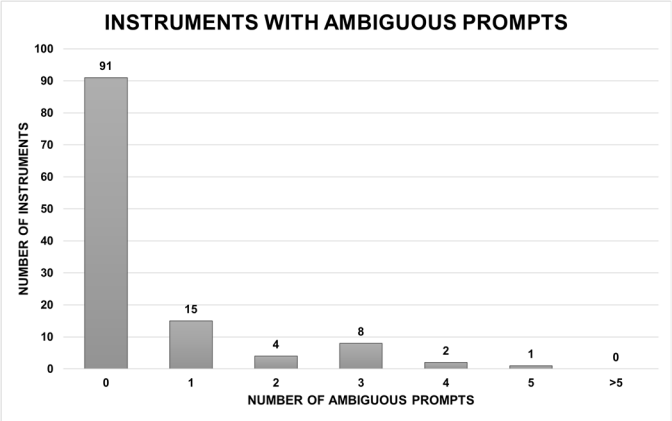


Fig. 4. Distribution of Ambiguous Statements Across Instruments Studied

C. Awkward Statement Results

The total number of awkward statements coded across all the evaluation instruments was 53, accounting for 1.3% of the 4098 total Likert-style statements. Of the 121 instruments evaluated, 97 (80.2%) contained 0 awkward statements, and only 2 (1.7%) contained more than 5 awkward statements. Fig. 5 shows the breakdown of the number of awkward statements across all 121 instruments.

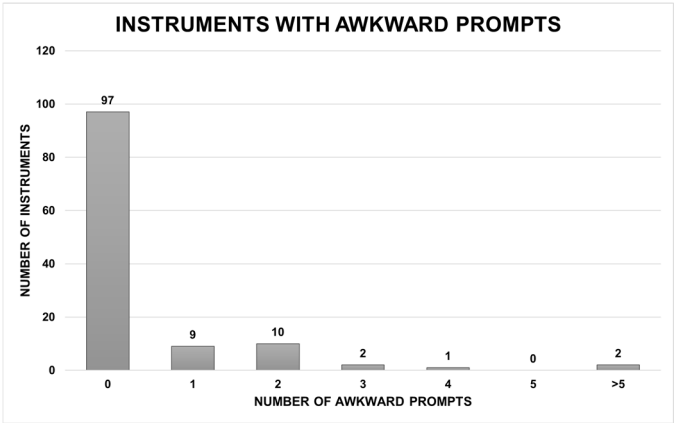


Fig. 5. Distribution of Awkward Statements Across Instruments Studied

D. Conjunctive Statement Results

The total number of conjunctive statements across all the evaluation instruments was 588, accounting for 14.3% of the 4098 total Likert-style statements. Of the 121 instruments evaluated, just 24 (19.8%) contained 0 conjunctive statements, and 31 (25.6%) contained more than 5 statements. Fig. 6 shows the breakdown of the number of conjunctive statements across all 121 instruments.

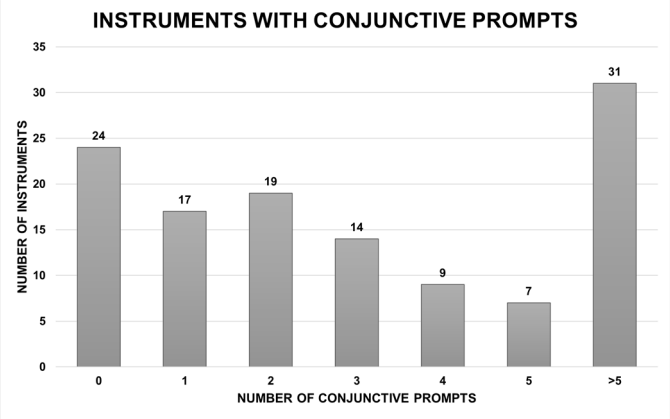


Fig. 6. Distribution of Conjunctive Statements Across Instruments Studied

It may be of interest to note that one of the evaluation instruments, the Relevance of Science Education (ROSE) Questionnaire [12], composed of more than 240 Likert-style statements was responsible for 114 of the conjunctions. When removing this instrument from the overall count, the percentage of conjunctive statements dropped to 11.6% and totaled 474 of 4098 statements.

E. Leading Statement Results

The total number of statements coded as being leading across all the evaluation instruments was 9, accounting for 0.2% of the 4098 total Likert-style statements. Of the 121 instruments evaluated, 115 (95.0%) contained 0 conjunctive statements, and no instrument contained more than 3 leading statements. Fig. 7 shows the breakdown of the number of leading statements across all 121 instruments.

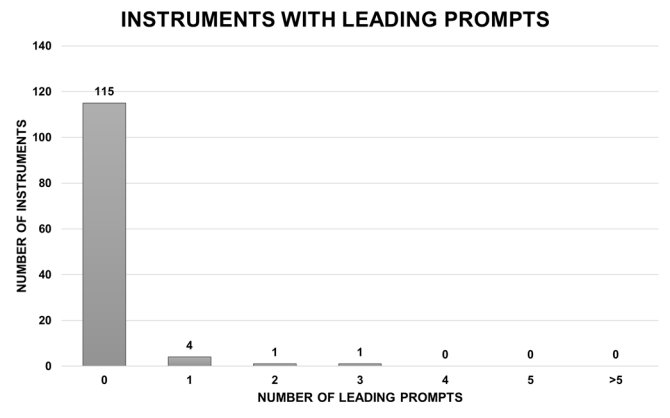


Fig. 7. Distribution of Leading Statements Across Instruments Studied

F. Double-Negative Statement Results

Throughout the coding of the 4098 statements across the 121 instruments, zero (0) statements were coded as containing a double-negative.

G. Number of Options Results

The quantity of options or response levels presented along with a statement for consideration in a Likert-style question has been previously reported to possibly bias the responses without improving the reliability or validity [11]. Specifically, there is the possibility of cognitive load concerns when participants are presented with more than 7 options.

The breakdown of the number of options is presented in Fig. 8. Considering that only 1 of the 121 evaluation instruments used more than 7 options, this study did not further consider the implications resulting from the quantity of response levels.

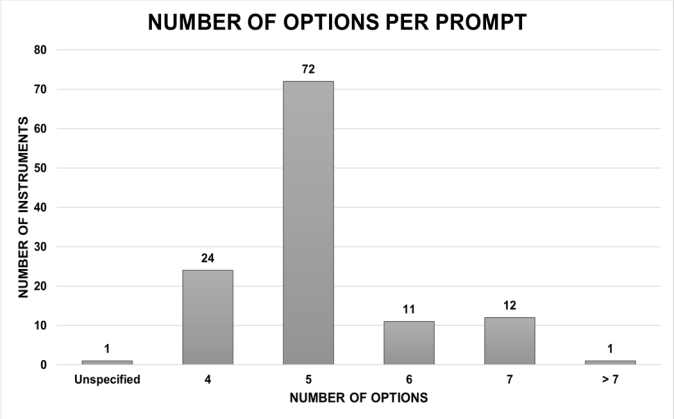


Fig. 8. Distribution of Number of Options per Likert-Style Prompt Across Instruments Studied

H. Ranking Direction Results

The ranking direction of the response options, (i.e. from strongly disagree (1) to strongly agree (5) or from strongly agree (1) to strongly disagree (5)), presented with the statement for consideration in a Likert-style question has also been reported to strongly bias the responses [1, 4, 5]. Regardless of the numerical values associated with the rankings, for Likert-style questions which presented the options from strongly disagree to strongly agree were tabulated as having a negative ranking direction and questions which presented options from strongly agree to strongly disagree were tabulated as having a positive ranking direction. Fig. 9 shows the breakdown of the option directions for the analyzed evaluation instruments.

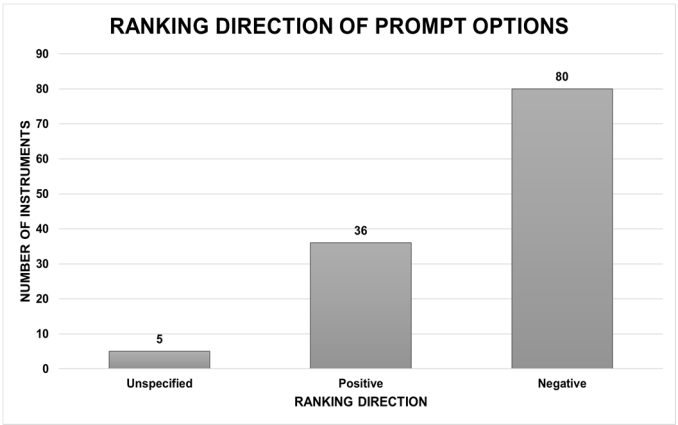


Fig. 9. Distribution of Ranking Direction of Prompt Options Across Instruments Studied

Of the 121 instruments analyzed, 5 left the ranking direction unspecified in the instructions. It is assumed that researchers using the evaluation instrument were free to assign their own order to the responses. Of the remaining tools, 80 specified a negative (strongly disagree to strongly agree) ranking, leaving 36 tools specifying a positive (strongly agree to strongly disagree) ordering; a ratio of 2.2 negatively ranked instruments for every positively ranked instrument.

When taking into consideration the demographic information referenced in the previous section, a tendency for instruments directed at younger participants to be more positively ranked than average was apparent. Instruments reported as being intended for K-12 students tended to be more positively directed than average (1.3 negative instruments for each positive) while instruments reported as being for post-secondary and older persons (undergraduates, graduate students, teachers, and adults) were more often negatively directed (2.8 negative instruments for each positive) than average. The breakdown of this result by demographic is presented Fig. 10. It is not readily apparent in the reports from the researchers who created the evaluation instruments whether the ranking direction for these instruments was an intentional design decision.

Demographic (U.S. Grades)	Total Instruments	Number Positive	Number Negative	Negative:Positive Ratio	Unspecified
Elementary (K-5)	11	6	5	0.8	0
Middle (6-8)	33	13	18	1.4	1
High (9-12)	32	12	19	1.6	1
Undergrad	38	8	25	3.1	4
Grad	8	2	6	3.0	0
Teacher/Adult	33	9	23	2.6	0

Fig. 10. Demographic Breakdown of Ranking Direction of Prompt Options Across Instruments Studied

VI. DISCUSSION

Rather unexpectedly, there were few instruments curated at csedresearch.org (21 out of 121; ~17.4%) which were completely free from the pitfalls recognized by the CDC [6], AAMC [3], AAPOR [2], or Likert's original paper [8]. The two pitfalls which statements were most susceptible to were ambiguity and conjunctions.

Regarding ambiguous statements, there were two observations which resonated. First, the interrater reliability was lowest when coding ambiguous statements (~88.1%). While exploring this result was beyond the scope of this study, one possible reason for this was the large difference in age and experience between the coders. Likert strongly emphasized the necessity to simplify statements, recommending that the statements are written directed at "persons of less understanding than any member of the group for which the test is being constructed." [8] Considering that the curated instruments were composed for specific studies on specific age groups, terms within the statements which are unambiguous for certain groups may have been coded as being ambiguous as a result of differing experiences of the coders from the intended study participants, and between themselves. Because the coding of the statements was done separate from the context, it is also the case that coded ambiguities may not be ambiguous if there was a way to understand the context of the instrument and target demographic. Therefore, instrument creators should be mindful not only of the ambiguity of their statements in isolation, but also in the context of the instrument.

The relatively large number of statements coded as having conjunctions (588 out of 4098; ~14.3%) resulted in observations and concerns surrounding common colloquialisms and word groupings which are part of accepted vernacular. For example, the conjunction "up and running" (e.g., "the computer program is up and running") may be a commonly accepted singular term easily understood by one group of participants but considered to be two different items by another group and thus a source of uncertainty. These concerns become amplified when considering different cultural groups (North American vs. European) or languages (English vs. Spanish), especially when instruments need to be translated. Additionally, translation between languages may result in awkward statements.

As recognized by Tversky and Kahneman [14], when a person is met with a situation in which they need to make a complex judgement shrouded in uncertainty that person will often simplify the decision through a limited number of heuristics. Two of these heuristics are representativeness and availability. Representativeness is activated when a person is asked to judge the relatedness of two or more items and results in that person making their judgement based upon the degree to which one represents the other. Availability is the propensity for a person to judge the probability of an event based upon the ease which they can bring an example forward in their thinking [15].

As a demonstration of these heuristics at work within the current study and to suggest how conjunctions may be a source of biasing and noise, consider as an example the conjunction of 'designing and building'. An example Likert-style statement fabricated for the current study (and not found in the analyzed evaluation instruments) using this conjunction could ask a respondent to evaluate their level of agreement to 'I like designing and building computers in my free time.'

Considering representativeness and its implication with this conjunction, if a person identifies and believes they are representative of one of the conjoined options but not the other, their response may be influenced dependent upon the strength of that identification. Availability would be activated if the respondent can easily recall an experience involving either of the conjoined options and not necessarily both.

Regardless of the heuristic involved, the response and thus the research may be unintentionally biased solely because of presence of the conjunction. Ambiguous and awkward statements which result in participant uncertainty are susceptible to similar unintentional biasing.

A. Limitations

There are a few limitations of this study to consider. The first is that the instruments collected were analyzed by hand and thus human error can cause erroneous coding for any particular statements/prompts. Further, the data studied was housed in a public database (csedresearch.org) and subject to possible curation errors. However, the data curation process for csedresearch.org is documented and involves multiple levels of

review which would mitigate the associated risks of possible curation errors.

VII. CONCLUSIONS

In this study, we examined instruments that used Likert-type prompts and statements to discover whether they succumbed to certain pitfalls that could impact their efficacy and ability to measure attitudes accurately. We discovered that only 17.4% of the instruments housed in csedresearch.org were completely free from the pitfalls of ambiguity, awkwardness, conjunctions, and direct biasing. While most (~82.8%) of the problematic statements were coded for conjunctions, 30 instruments had at least 1 ambiguous statement, 24 had at least 1 awkward statement, and 6 instruments had leading statements.

Overall, these results point to reasonably designed Likert-style surveys and questionnaires, but with the few areas of concern noted. As such, it is important to remind survey designers of the key points of good design for Likert-style prompts and questionnaires.

Finally, presented in Fig. 11 is a summary of recommendations for survey statement/prompt creation to avoid the pitfalls described throughout this report. In addition, there are examples of statements recognized as succumbing to these pitfalls and possible revisions to those statements. We encourage instrument creators, researchers, and evaluators to be mindful of these pitfalls when designing their instruments and to work to ensure that the most commonly identified errors (ambiguity and conjunction) have been removed from the final scales.

	DESCRIPTION	EXAMPLE FROM CSEd RESEARCH	REVISED VERSION
Ambiguous Statements	Statements which are capable of being understood and/or interpreted in multiple ways.	"I like to scan computer journals."	"I like to read articles about computing topics that are not assigned as coursework."
Awkward Wording	Awkward statements are usually the result of malformed and/or poor sentence structure.	"I am pleased with the instructor's evaluations of my work compared to how well I think I have done."	"The work I have done has been evaluated fairly by my instructor."
Conjunctive Prompts (Double-barreled)	Statements which are composed of multiple things being asked in a single prompt; consideration of items from multiple sets.	"Computers make me feel uneasy and confused."	Separate into two statements: 1. "Computers make me feel uneasy." 2. "Computers make me feel confused."
Leading Statements	Leading statements are those in which the statements themselves introduce bias, whether through ideas or suggestions of an authority figure.	"Learning to operate computers is like learning any new skill – the more you practice, the better you become".	"You get better at operating computers the more you practice using them."
Balanced Questions & Responses	Including the necessary range and/or number of response options for all respondents to be able to answer the prompt.	"How often each day do you worry about debugging a computer program?" • Once • Twice • Always	"How often each day do you worry about debugging a computer program?" • Zero times • 1-3 times • More than 4 times

Fig. 11. Recommendations and Examples

REFERENCES

[1] Albanese, M., Prucha, C., Barnet, J., and Gjerde, C., 1997. The Effect of Right or Left Placement of the Positive Response on Likert-type Scales Used by Medical Students for Rating Instruction. *Academic Medicine* 72, 627-630.

[2] American Association for Public Opinion Research [AAPOR], 2007. Question Wording. Retrieved from: https://www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Question-Wording.aspx

[3] Artino, A.R., Gehlbach, H., and Durning, S.J. 2011. AM Last Page: Avoiding Five Common Pitfalls of Survey Design. Academic Medicine, 86(10), 1327. DOI: 10.1097/ACM.0b013e31822f77cc

[4] Atkins-Burnett, S., 2016. Assessing Children and Adolescents in Large Scale Surveys. Mathematica Policy Research commissioned by National Academies of Sciences, Engineering, and Medicine. Retrieved from: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_171794.pdf

[5] Carp, F., 1974. Position Effects on Interview Responses. *Journal of Gerontology* 29, 581-587.

[6] Centers for Disease Control and Prevention [CDC], 2011. "Constructing Survey Questions", Retrieved from: https://www.cdc.gov/dhdsp/docs/constructing_survey_questions_tip_sheet.pdf

[7] Decker, A., McGill, M. M., 2019. A Topical Review of Evaluation Instruments for Computing Education. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). Association for Computing Machinery, New York, NY, USA, 558–564. DOI: https://doi.org/10.1145/3287324.3287393

[8] Likert, R., 1932. "A Technique for the Measurement of Attitudes," Archives of Psychology 22, 140, 5-55.

[9] Margulieux, L., Ayer Ketenci, T., and Decker A. 2019. Review of Measurements Used in Computing Education Research and Suggestions for Increasing Standardization. Computer Science Education. 29:1, 49-78, DOI: 10.1080/08993408.2018.1562145

[10] McGill, M. M., Decker, A., McKlin, T., and Haynie K. 2019. A Gap Analysis of Noncognitive Constructs in Evaluation Instruments Designed for Computing Education. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). Association for Computing Machinery, New York, NY, USA, 706–712. DOI:https://doi.org/10.1145/3287324.3287362

[11] Revilla, M., Saris, W., and Krosnick, J., 2014. Choosing the Number of Categories in Agree-Disagree Scales. *Sociological Methods and Research* 43, 73-97.

[12] Schreiner, C. and Sjoberg, S., 2004. "Sowing the seeds of ROSE: Background, rationale, questionnaire development and data collection for ROSE (The Relevance of Science Education): A comparative study of students' views of science and science education," Acta didactica. Retrieved from: https://roseproject.no/key-documents/key-docs/ad0404-sowing-rose.pdf

[13] Thurstone, L. L. (1928). "Attitudes can be measured," American Journal of Sociology, 33, 529-54

[14] Tversky, A. and Kahneman, D., 1974. "Judgment under Uncertainty: Heuristics and Biases," Science 185, 1124-1131.

[15] Tversky, A. and Kahneman, D., 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," Psychological Review 90, 293-315.