Working Paper

LOUIS T. MARIANO, ANDREA PHILLIPS, KEVIN ESTES, AND M. REBECCA KILBURN

Should Survey Likert Scales Include Neutral Response Categories?

Evidence from a Randomized School Climate Survey

For more information on this publication, visit **www.rand.org/t/WRA3135-2**.

About RAND

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

Limited Print and Electronic Distribution Rights

About This Working Paper

Likert scales often feature survey response categories ranging from "Strongly Disagree" to "Strongly Agree." While survey scales of this form are commonly used, including in areas such as research and quality improvement, the choice of whether to offer a neutral response category is not well informed by the available literature. Related to a broader effort using such scales in examinations of school climate, this study embedded an experiment into a survey of teachers, randomly offering the same survey scales with and without a neutral option.

This working paper was developed as a guide for practitioners and consumers of Likert scale surveys. Summarized results of the survey randomized experiment are presented. The manuscript then provides recommendations for the inclusion of neutral categories in both the adaptation of pre-validated survey scales and the development of new scales. Recommendations for reporting results from degree-of-agreement response scales are also discussed.

This working paper derives from a companion working paper, "Examining the Inclusion of Neutral Response Categories Using an Item Response Theory Approach: Analysis of a Randomized Survey of Teachers" (Mariano, et al., 2024). The companion manuscript presents a novel approach for examining the utilization of Likert scale neutral categories, provides a detailed discussion of the methods used to analyze the survey data from this randomized experiment, and provides an in-depth examination of the experiment results.

Justice Policy Program

RAND Social and Economic Well-Being is a division of the RAND Corporation that seeks to actively improve the health and social and economic well-being of populations and communities throughout the world. This research was conducted in the Justice Policy Program within RAND Social and Economic Well-Being. The program focuses on such topics as access to justice, policing, corrections, drug policy, and court system reform, as well as other policy concerns pertaining to public safety and criminal and civil justice. For more information, email justicepolicy@rand.org.

Acknowledgments

This project was supported by Award No. 2017-CK-BX-0020, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice. The authors wish to thank the RAND American Educator Panel team, including Claude Setodji, Daniel Ibarrola, Roulin Lu, and Michelle Bongard, for their expert assistance.

Contents

Figures and Tables

Figures

Tables

Chapter 1. Introduction

Survey Likert scales are common in research and other applications. Such scales use a set of questions with a common response scale to inform an underlying latent construct, also called a latent variable. For example, a scale informing school leadership might ask teachers about the culture among adults in their school, such as "There is an atmosphere of trust and mutual respect in this school" (North Carolina Department of Public Instruction, 2023). These scales often feature ordinal degree-of-agreement response categories ranging from "Strongly Disagree" to "Strongly Agree." Commonly, the responses to the individual items within the scale are assigned integer scores (e.g., values ranging one through five for five ordinal response categories), and then the item scores are summed or averaged to produce a scale score.

When assessing degree of agreement, there is currently conflicting guidance on whether to use a neutral response category. Prior studies that generally focus on the psychometric properties of degree of agreement survey scales have established that greater than three categories yield improved reliability and validity (Preston and Colman, 2000; Weng, 2004; Lee and Paek, 2014; Finn, Ben-Porath, and Tellegen, 2015; Alwin, Baumgartner, and Beattie, 2018; Simms, et al., 2019), with the benefits of additional categories tailing off after seven response options are present (Lozano, Gargia-Cueto, & Muniz, 2008). In practice, a 5-category version—"Strongly Disagree," "Disagree," "Neutral," "Agree," "Strongly Agree"—and a 4-category option that excludes "Neutral" are common.[1]

To enhance the available guidance on whether to include a neutral option, we conducted a randomized experiment with a survey of teachers on the general topic of school climate. As described below, our design allowed us to examine responses to 4-category and 5-category versions of the same scale by similar groups of respondents. Our goal is to understand how response data may inform the value of including a neutral option. While scale design needs extend beyond what the data can tell us and we only examine six scales, the experimental design allows us to gain keen insight into features of the responding population that may be critical to consider.

Advocates of including "Neutral" seek to provide a viable option for respondents whose sincere belief falls between "Disagree" and "Agree." Opponents question whether the meaning of the neutral response may be distinguished among valid responses or is a "dumping ground" of other reasons for the choice, such as lack of information or competence to reply, inability to decide, or indifference (Kulas, Stachowski, & Haynes 2008). The value of the neutral option may also be context specific. For example, the fact of whether teachers are provided time for

[1] Alternative phrasing of the neutral category, e.g., "Neither Agree nor Disagree," is also common.

observation in their school may make it easy to provide a non-neutral response to the item "Teachers take time to observe each other teaching." In contrast, when asked whether "Students at this school follow rules of conduct," the presence of different sets of students with different behavioral patterns or some rules that are followed while others are not may motivate the teacher to desire a sincere response between disagreement and agreement.

Examining a single "Neutral" response, it is difficult to identify it as a sincere opinion or a "dumping ground" choice. Herein, we take an Item response Theory (IRT) approach, using a Graded Response Model (GRM; Samejima, 1969) to map values of the latent variable to the probability or responding in each category, allowing us to parse the meaning of an individual's response. Consider, for example, Figure 1.1. where response probabilities to the 5-category version of the item "The school rules are fair" are displayed. This item is one of eight present on a scale informing fairness and rule clarity. The x-axis contains the range of the latent variable, the teacher's assessment of the fairness and clarity of school rules, which is scaled to follow and Standard Normal distribution. The y-axis displays the probability of responding in each category. Notice that each response category is utilized, meaning that there is a segment of the latent variable for which the probability of responding that that category is greatest. The neutral category is preferred by individuals with latent variable values on the segment (-1.72, -1.10).

Alternatively consider Figure 1.2. which displays responses to the item "Teachers have opportunities for dialogue and planning across grades and subjects." This item is part of a scale informing teacher collaboration. Here the neutral category is not utilized; there is no segment on the range of latent variable values where "Neutral" is preferred over the other choices. As these two examples illustrate, by examining probabilities of responding in each category relative to the latent variable value we can determine whether a neutral response is a sincere reflection of opinion between disagreement and agreement.

Figure 1.1. An Example of Item Category Response Probabilities with the Neutral Category Utilization

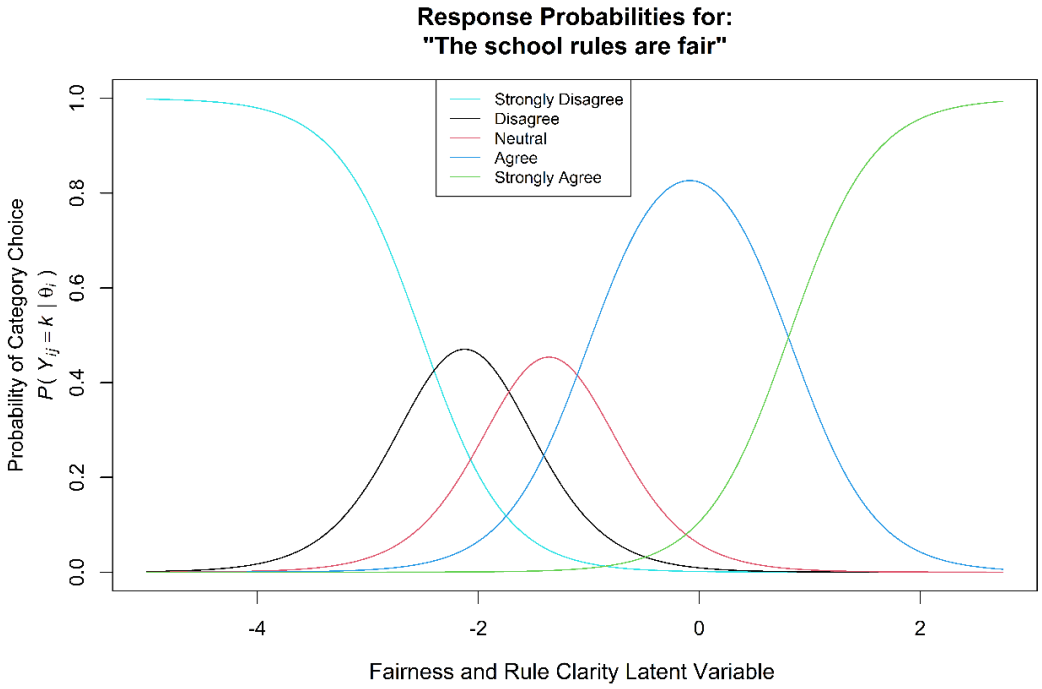
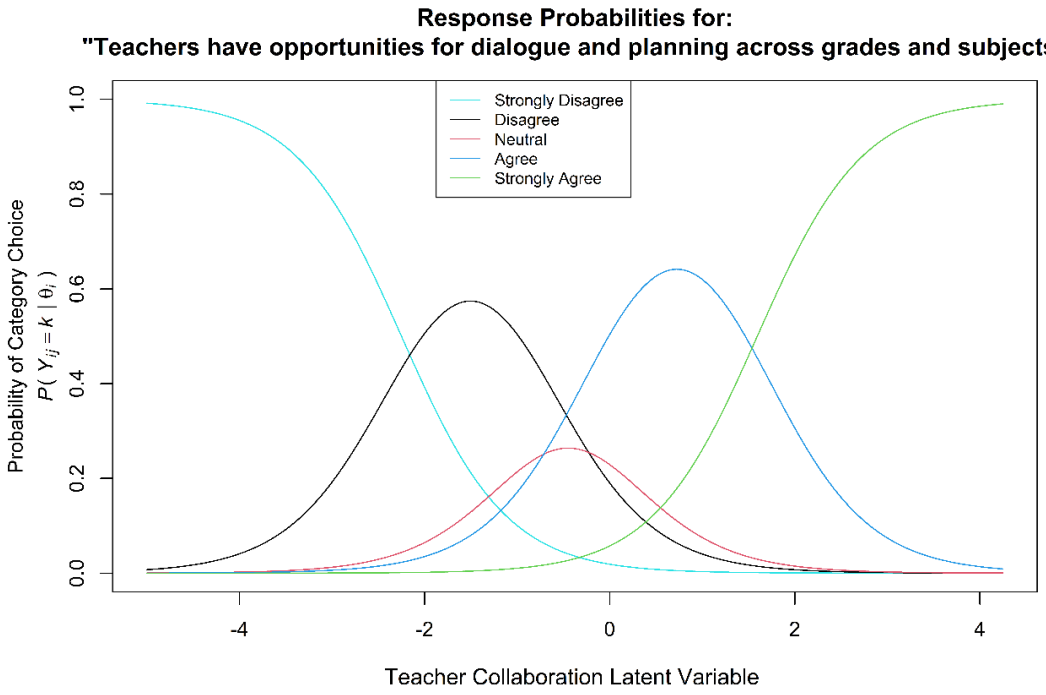


Figure 1.2. An Example of Item Category Response Probabilities Lacking Neutral Category Utilization



Chapter 2. Design and Methods

To better inform the question of whether to include a neutral category, we embedded an experiment into a 6-scale survey of school climate. Four of the six scales are primarily fielded with 4-category response scales; two of these also include a "Don't Know" option. The other two are primarily fielded with 5-category response scales. For each scale, we also created an alternate version that included or excluded the neutral category as opposite the primary version ("Don't Know" was retained in the alternative version when present in the primary). We created four different survey forms, each of which contained either a primary or alternate version of each scale. We then randomly assigned four samples of public elementary school teachers to the four forms, yielding approximately 1,000 responses to each scale in each of the primary and alternate versions. The scales and randomization pattern are detailed in Tables 2.1 and 2.2. The individual scale items and additional source information for each scale are presented in the Appendix A. The samples were drawn, and the survey administered online via the American Teacher Panel (ATP; Robbins and Grant, 2020).

For each survey scale, we use a GRM to model the probabilities of responding in each category on each item. The 5-category responses and 4-category responses are modeled separately. We examine the fitted 5-category model to determine, for each individual item, whether there is evidence that the neutral category is utilized. To determine statistical significance, we construct a 95% confidence interval for the length of the segment of the latent variable where neutral is preferred; items with confidence intervals strictly above zero demonstrate significant neutral category utilization.

We provide complimentary standard (discussed below) analyses to address several additional questions, including: (i) whether "Don't Know" may serve as a proxy for "Neutral" when the latter is not offered (ii) What are the implications for Likert Scale scores when including/excluding "Neutral" (iii) Does including/excluding "Neutral" impact aggregations such as percent who "agree or strongly agree" (iv) Do response times differ when "Neutral" is included/excluded.

Table 2.1. Survey Scales and Versions Fielded

Scale	Primary Version	Alternate Version	Source	Items
1. Fairness and Rule Clarity	(1P) 4-category	(1A) 5-category	California School Staff Survey	8
2. Safety-Bullying Subscale	(2P) 4-category	(2A) 5-category	ED School Climate Survey	8
3. Evidence Based Practices	(3P) 5-category	(3A) 4-category	Adaptation by Lyon, et al., (2018) from the Implementation Climate Scale	10
4. Teacher Collaboration	(4P) 5-category	(4A) 4-category	The Culture Survey	6
5. Managing Student Conduct	(5P) 4-category + Don't Know	(5A) 5-category + Don't Know	North Carolina Teacher Working Conditions	7
6. School Leadership	(6P) 4-category + Don't Know	(6A) 5-category + Don't Know	North Carolina Teacher Working Conditions	9

Note: The 4-category response versions includes "Strongly Disagree," "Disagree," "Agree," and "Strongly Agree." The 5-category version adds a "Neutral" middle response category.

Table 2.2. Survey Forms

Form A	Form B	Form C	Form D
1P	1P	4A	3A
4A	3A	2P	2P
3P	4P	1A	1A
2A	2A	3P	4P
5P	6P	6P	5P
6A	5A	5A	6A
491 responses	526 responses	511 responses	492 responses

Note: Form components are identified in Table 2.1. "P" indicates the primary version of the scale. "A" indicates the alternate version of the scale.

Chapter 3. Results

Neutral Category Utilization

Table 3.1. summarizes neutral category utilization for each of the six scales. Across scales, 31 of 48 items (65 percent) contained a significant segment of the population that preferred the neutral response option. In a given scale, the percentage of items with neutral utilization ranged for 100 percent for the Evidence Based Practices scale, down to 25 percent for the Safety-Bullying scale. The results suggest that, when presented, "Neutral" is a meaningful expression of sentiment and commonplace, at least in the school climate domain. Two of the four primarily 4-category scales exhibited a high percentage of items with neutral utilization present when the 5-category version was offered. Conversely, one of the two primarily 5-category scales exhibited neutral category utilization for only half the scale's items. This dichotomy suggests that survey developers may not have the best sense of when a neutral preference may be preferred.

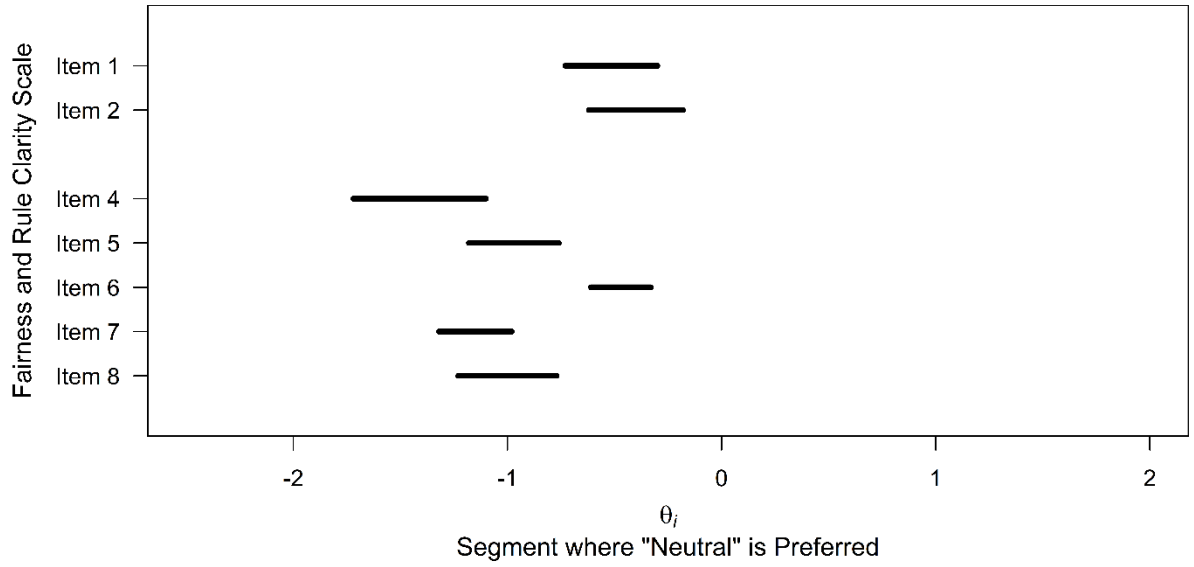
Across respondents, it was rare for a respondent to select "Neutral" on a large number of items. For each scale, we further examined the range of the latent variable where "Neutral" was preferred (i.e., the neutral utilization range) for each item. Figure 3.1. displays these results for the Fairness and Rule Clarity Scale. We found considerable variation in the range of neutral utilization within each scale, further indicating choice of "Neutral" as a sincere expression of belief.

Table 3.1. Summary of Items Demonstrating "Neutral" Category Utilization

Scale	Primary Version Number of Categories	Total Items	Items with Significant Utilization of Neutral	Average Latent Variable Range Preferring Neutral	Average Percentage of Population Preferring Neutral
Fairness and Rule Clarity	4-category	8	7	0.43	11.27
Safety-Bullying	4-category	8	2	0.25	3.78
Evidence Based Practices	5-category	10	10	0.94	26.83
Teacher Collaboration	5-category	6	3	0.56	21.35
Managing Student Conduct	4-category + Don't Know	7	3	0.30	10.41
School Leadership	4-category + Don't Know	9	6	0.57	16.80

Note: Average range and average percentage preferring neutral are calculated only among items with significant utilization. Average range is presented in standard deviations of the θ scale.

Figure 3.1. Segments of the range of the latent variable where the neutral category has the greatest probability of response: Fairness and Rule Clarity Scale



Don't Know" versus "Neutral"

We next investigated the relationship between "Don't Know" and neutral responses. Across the two scales that offered a "Don't Know" option, selection of "Don't Know" was rare—below one percent of the time when "Neutral" was offered and two to three percent of the responses otherwise. For 11 of these 15 items, the lower frequency of "Don't Know" selection in the presence of "Neutral" was statistically significant; however, this choice of "Don't Know" is just a small fraction of the portion of the population preferring "Neutral," suggesting that "Don't Know" is not serving as a proxy for "Neutral" when the latter is not offered.

Recognizing that individuals with clear preferences of agreement or disagreement will not select a "Neutral" or "Don't Know" response option, we conducted a closer examination focusing only on those individuals with latent variable estimates in the in the range where "Neutral" is the highest probability choice when offered (and sub-setting to the items where "Neutral" had evidence of utilization). Results are displayed in Table 3.2. In this focused examination, none of the items on the Managing Student Conduct scale and only two items on the School Leadership scale had a significant difference in "Don't Know" selections with and without "Neutral" offered. Across the two items that did show a significant difference in "Don't Know" endorsement, the average difference was about seven percent of those who were most likely to choose "Neutral." These results imply that lack of information or competence to reply is an unlikely explanation for "Neutral" selection for these two scales.

Table 3.2. Summary of "Don't Know" Selection

Scale	4-Category Percentage "Don't Know" Average	5-Category Percentage "Don't Know" Average	Total Items	Items with Significant "Don't Know" Difference	Average "Don't Know" Significant Difference Percentage
Among Respondents with Latent Variable Estimates in the Neutral Utilization Range					
Managing Student Conduct	2.06	0.93	3	0	--
School Leadership	4.64	0.64	5	2	7.23

Note: One outlier School Leadership item was omitted; see the Appendix A for additional details. Average difference calculated only among those items with a significant difference in "Don't Know" selection. The neutral utilization range is the segment on the latent variable upon which the neutral category is preferred; averages over the neutral utilization range is taken over all items with significant neutral utilization.

The Impact of a Neutral Option on Scale Scores

For each scale, we compared the difference in average scale scores when the neutral response category was and was not offered. A standard Likert average score was created for the 5-category versions of the scales, with "Strongly Disagree" to "Strongly Agree" mapped to the integers from one to five. For the 4-category versions, we shifted the scores assigned to "Disagree" and "Agree" to keep the scoring range consistent, as shown in Table 3.3.[2]

Table 3.3. Equating Likert Scoring Between 4-Category and 5-Category Versions

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
4-Category	1	2.33	N/A	3.67	5
5-Category	1	2	3	4	5

Results of tests of a difference in 4-category and 5-category scale means are displayed in Table 3.4. Although two of the six scales displayed significant mean differences, these differences were under one-tenth of a point in magnitude on a 4-point scale. Such a small difference is unlikely to materially impact conclusions.

We also implemented a Kolmogorov-Smirnov test to examine whether the distributions of Likert scores were the same across versions. The results imply that distributional differences are present among 4-category and 5-category versions of the Likert scores. Further examination (details omitted for brevity) revealed that for nearly half of all survey items a higher value on the latent variable was needed before a respondent would prefer "Strongly Agree" over "Agree,"

[2] We conducted a sensitivity analysis where the 4-category "Disagree" and "Agree" responses were mapped to 2 and 4 respectively. This alternate equating led to similar conclusions as those reported here.

suggesting a respondent's perceptions of what it means to "Strongly Agree" may shift between the 4-category and 5-category versions.

Table 3.4. Difference in Scale Scores when "Neutral" is Present or Absent

Scale	4-Category Mean Scale Score	5-Category Mean Scale Score	Difference in Mean Scale Scores	Test of Equal Distributions P-value
Fairness and Rule Clarity	3.57	3.66	-0.09**	<0.001
Safety-Bullying	3.91	3.95	-0.04	<0.001
Evidence Based Practices	3.45	3.42	0.03	<0.001
Teacher Collaboration	3.04	3.02	0.02	0.002
Managing Student Conduct	3.56	3.60	-0.04	<0.001
School Leadership	3.51	3.58	-0.07*	<0.001

Note: 4-catergory items exclude "Neutral." 4-category scale score calculated as indicated in Table 3.3. Test of equal distributions is conducted via a Kolmogorov-Smirnov test. *Significant at α=0.05; **Significant at α=0.01; ***Significant at α=0.001

Collapsing Adjacent Categories

It is frequent practice to aggregate the "Agree" and "Strongly Agree" responses and report the percentage at least agreeing with an item. For each of the 48 items across the six scales, we calculated percentage at least agreeing under both the 4-category and 5-category response format and tested whether significant differences were present across formats. Table 3.5 summarizes the results be scale. All but two of the 48 items demonstrated a significantly higher percentage who at least agree when responding under the 4-category version. Differences typically ranged about nine to eleven percentage points higher. Results for the Evidence Based Practices scale were higher, averaging 22 percent higher when the 4-category response scale was used. Results for reporting a combined percentage who "Disagree" or "Strongly Disagree" were similar, although slightly smaller in magnitude (details omitted for brevity). Results suggest that collapsing adjacent categories materially changes findings.

Table 3.5. Summary of Difference in Percentage Who "At Least Agree" when "Neutral" is Present or Absent

Scale	Total Items	Items with Significant Difference in "At Least Agree"	Average Percentage Point Significant Difference in "At Least Agree"
Fairness and Rule Clarity	8	6	8.80
Safety-Bullying	8	8	8.64
Evidence Based Practices	10	10	22.03
Teacher Collaboration	6	6	13.34
Managing Student Conduct	7	7	9.19
School Leadership	9	9	10.51

Note: Average percentage significant difference is calculated on significant items only. Positive values indicate the percentage is higher when "Neutral" is absent. "At Least Agree" includes "Agree" and "Strongly Agree."

Differences in Response Times

If selection of the neutral category is caused by indecision about the item, as opposed to a sincere neutral stance, a response scale that does not offer a neutral option may result in longer time to contemplate and select a response. To investigate the potential for indecision-motivated "Neutral" selection we collected the elapsed time for completion of each of the six scales and tested whether a significant difference in response times were present between the 4-category and 5-category response versions. Similar to our examination of "Don't Know" responses, we note that individuals with strong positive or negative sentiment toward an item will not consider using the neutral category. Including such individuals in examining response times could mask differences present among those who may use the neutral choice. Here, we subset the data for each scale to only examine response times for those with latent variable values within the range which "Neutral" is the preferred option.[3] Because response time is a scale-level value, we include the range of the latent scale where respondents may prefer "Neutral" for any item on the scale. For example, considering ranges of neutral preference for the Fairness and Rule Clarity Scale illustrated in Figure 3.1., we include individuals with estimated latent values between -1.72 (the low end of the lowest range, which is item 4) and -0.18 (the high end of the highest range, which is item 2).

Response time differences are displayed in Table 3.6. In both the 4-category and 5-category response versions, the respondents move quickly through the items, typically averaging between 6 and 9 seconds per item response. Only one scale, Fairness and Rule Clarity, had a significant

[3] We conducted a sensitivity analysis using all respondents that led to the same general conclusions about response times discussed here.

response time difference, with 4-category responses taking about a mere six-tenths of a second longer per item. Point estimate for the other five scales were not significant and varied between longer and shorter for the 4-category version. These results are inconsistent with the idea that neutral responses reflect respondent indecision.

Table 3.6. Difference in Mean Response Time when "Neutral" is Present or Absent

Scale	Total Items	4-Category Mean Response Time (Seconds)	5-Category Mean Response Time (Seconds)	Difference in Mean Response Time (Seconds)	Per Item Difference (Seconds)
Among Respondents with Latent Variable Estimates in the Neutral Utilization Range					
Fairness and Rule Clarity	8	53.32	48.53	4.79*	0.60
Safety-Bullying	8	62.52	54.06	8.46	1.06
Evidence Based Practices	10	91.69	90.62	1.07	0.11
Teacher Collaboration	6	47.90	45.00	2.90	0.48
Managing Student Conduct	7	49.02	50.64	-1.61	-0.23
School Leadership	9	55.55	56.00	-0.45	-0.05

Note: 4-catergory items exclude "Neutral." The neutral utilization range is the union of the segments on the latent variable upon which the neutral category is preferred for all items with significant neutral utilization. *Significant at α=0.05; **Significant at α=0.01; ***Significant at α=0.001.

Chapter 4. Conclusions

Using the unconventional approach of Item Response Theory to link Likert scale item responses to the underlying latent trait they inform, we demonstrate that the neutral category is the preferred choice for a well-defined segment of the population for nearly two-thirds of the survey items examined. We also find variation in which segment of the population prefers the neutral response option across items within the same scale. These results support the present of valid expressions of sentiment between agreement and disagreement, as opposed to choosing "Neutral" due to indecision, indifference, or lack of information. Further, we do not find that a "Don't Know" option serves as a proxy neutral response when the latter is not offered, nor do we find meaningful differences in response times when the neutral option is present or absent.

While the Likert scale averages did not appear to be impacted in a practical way by whether or not "Neutral" was offered, we find important differences when adjacent categories are aggregated, such as the percentage who "Agree or Strongly Agree," *implying that one may reach meaningfully different conclusions depending on whether or not a neutral category is offered*.

While there are broader considerations for scale development than what response data can inform and this examination is limited to six scale examples, our results suggest including a neutral response option when a new scale is developed. Respondents may meaningfully utilize the neutral option such that exclusion could result in inaccurate conclusions, particularly for item-level inference. We recommend including the neutral option and then reconsidering the offering if a sufficient sample size demonstrates that the neutral category is not being utilized.

Often, validated scales are used for additional purposes, as there is value in having public results against which to compare findings and it avoids the expense of creating and validating a new scale. If the existing scale does not offer a neutral response option, one may wish to investigate if it would be used if offered, particularly if adjacent category aggregations are planned. Importantly, we strongly recommend that in all cases, results are reported concurrently with detail of whether a neutral response was offered to respondents, so that the reader may place the results in better context.

The intended use of a pre-existing scale plays an important role when considering adding a neutral option to an existing 4-category scale. If benchmarking to prior results is a priority, modification of the response scale may be undesirable. Conversely, if the scale informs quality improvement or progress monitoring, the presence of a neutral option may better capture changes in perception. Improving the accuracy of respondent sentiment could improve results for such applications.

Appendix A. Survey Scale Sourcing and Item Details

In this analysis we included six subscales from widely used teacher instruments in the education research community. Scales or subscales consisting of at least six items, previous evidence of validity and reliability in its primary form and offering either four or five response options were all considered for inclusion. We also included scales or subscales from multiple instruments to ensure findings were not limited to a single instrument. From the California School Staff Survey, the Fairness and Rule Clarity subscale (Mahecha & Hanson, 2020) was included. The California School Staff Survey is part of the California School Climate, Health, and Learning Survey (Cal-SCHLS) System and supported by the California Department of Education (California Department of Education, 2023). The Safety-Bullying Subscale was included from the ED School Climate Survey developed for the U.S. Department of Education and supported by the National Center on Safe Supportive Learning Environments (National Center on Safe Supportive Learning Environments, 2023). From an adaptation for teachers of the Implementation Climate Scale the Evidence Based Practices scale was selected for inclusion (Lyon, *et al.,* 2018). From the Culture Survey (Gruenart, 1998) the Teacher Collaboration scale was included. Finally, the Managing Student Conduct and School Leadership scales were included from the North Carolina Teacher Working Conditions instrument developed by the North Carolina Department of Public Instruction and presently replicated in more than 20 states (North Carolina Department of Public Instruction, 2023). Table A.1 presents the item stem wording for each of the 48 items included in these six scales.

Table A.1. Survey Scales

Item #	Item
	Fairness and Rule Clarity
1	This school clearly communicates to students the consequences of breaking rules.
2	This school handles discipline problems fairly.
3	Adults at this school treat all students with respect.
4	The school rules are fair.
5	Rules in this school are made clear to students.
6	This school clearly informs students what will happen if they break school rules.
7	Students know what the rules are.
8	This school makes it clear how students are expected to act.

Continued

Item #	Item
	Safety-Bullying
1	I think that bullying is a frequent problem at this school.[r]
2	I think that cyberbullying is a frequent problem among students at this school.[r]
3	Students at this school would feel comfortable reporting a bullying incident to a teacher or other staff.
4	Staff at this school always stop bullying when they see it.
5	Staff at this school are teased or picked on about their race or ethnicity.[r]
6	Staff at this school are teased or picked on about their cultural background or religion.[r]
7	Staff at this school are teased or picked on about their physical or mental disability.[r]
8	Staff at this school are teased or picked on about their sexuality.[r]
	Evidence Based Practices
1	The practice is easy to implement.
2	The practice is easy to fit into my day given other demands on my schedule.
3	The practice takes time away from instruction.[r]
4	The practice fits my personal teaching style.
5	The practice is useful in managing my students' classroom behavior.
6	The practice aligns with my school's behavior management system.
7	The practice is developmentally appropriate for my students.
8	The practice is culturally appropriate for my students.
9	My students need a practice like this.
10	The amount of time, resources, and effort needed to implement the practice is reasonable.
	Teacher Collaboration
1	Teachers have opportunities for dialogue and planning across grades and subjects.
2	Teachers spend considerable time planning together.
3	Teachers take time to observe each other teaching.
4	Teachers are generally aware of what other teachers are teaching.
5	Teachers work together to develop and evaluate programs and projects
6	Teaching practice disagreements are voiced openly and discussed.
	Managing Student Conduct
1	Students at this school understand expectations for their conduct.
2	Students at this school follow rules of conduct.
3	Policies and procedures about student conduct are clearly understood by the faculty.
4	School administrators consistently enforce rules for student conduct.
5	School administrators support teachers' efforts to maintain discipline in the classroom.
6	Teachers consistently enforce rules for student conduct.
7	The faculty work in a school environment that is safe.

Continued

Item #	Item
	School Leadership
1	There is an atmosphere of trust and mutual respect in this school.
2	Teachers feel comfortable raising issues and concerns that are important to them.
3	The school leadership consistently supports teachers.
4	Teacher performance is assessed objectively.
5	Teachers receive feedback that can help them improve teaching.
6	The faculty and staff have a shared vision.
7	The procedures for teacher evaluation are consistent.
8	The school improvement team provides effective leadership at this school.
9	The faculty are recognized for accomplishments.

Note: ˥ = item was reverse coded. For Managing Student Conduct Item 5 and School Leadership item 2, a mouse-over definition of “Teachers” was provided as: “Teachers means a majority of teachers in your school.” For School Leadership Item 3, a mouse-over definition was provided for “School leadership” as: “School leadership is an individual, group of individuals or team within the school that focuses on managing a complex operation. This may include scheduling; ensuring a safe school environment; reporting on students' academic, social, and behavioral performance; using resources to provide the textbooks and instructional materials necessary for teaching and learning; overseeing the care and maintenance of the physical plant; or developing and implementing the school budget.”

References

Alwin, D. F., Baumgartner, E. M., & Beattie, B. A. "Number of Response Categories and Reliability In Attitude Measurement." *Journal of Survey Statistics and Methodology*, Vol. 6, No. 2, 2018, pp. 212-239.

California Department of Education (2023), California School Staff Survey. As of September 28, 2023: https://www.cde.ca.gov/ls/he/at/cscs.asp

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. "Dichotomous Versus Polytomous Response Options in Psychopathology Assessment: Method or Meaningful Variance?" *Psychological Assessment*, Vol. 27, No. 1, 2015, pp. 184-193.

Gruenert, S. W. (1998). *Development of a School Culture Survey*. Pd.D. Thesis. Available from ProQuest Dissertations & Theses Global. (304438877). As of September 28, 2023: https://www.proquest.com/docview/304438877

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). "Middle Response Functioning in Likert-Responses to Personality Items." Journal of Business and Psychology, 22(3), 251-259.

Lee, J., & Paek, I. "In Search of the Optimal Number of Response Categories in a Rating Scale." *Journal of Psychoeducational Assessment*, Vol. 32, No. 7, 2014, pp. 663-673.

Lozano, L. M., García-Cueto, E., & Muñiz, J. "Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, Vol. 4, No. 2, 2008, pp. 73-79.

Lyon, A. R., Cook, C. R., Brown, E. C., Locke, J., Davis, C., Ehrhart, M., & Aarons, G. A. "Assessing Organizational Implementation Context in the Education Sector: Confirmatory Factor Analysis of Measures of Implementation Leadership, Climate, and Citizenship." *Implementation Science*, Vol. 13, 2018, pp. 1-14.

Mahecha, J., & Hanson, T. (2020). "Measurement Structure of the California School Climate, Health, and Learning Surveys: Student, Staff, and Parent Surveys." San Francisco, CA: WestEd. As of November 11, 2023: https://calschls.org/docs/measurementstructurecalschls_final.pdf

Mariano, L.T., Phillips, A., Estes, K., & Kilburn, M.R. "Examining the Inclusion of Neutral Response Categories Using an Item Response Theory Approach: Analysis of a Randomized Survey of Teachers." Santa Monica, CA: RAND Corporation, 2024. As of January 9, 2024: https://www.rand.org/pubs/working_papers/WRA3135-1.html

National Center on Safe Supportive Learning Environments (2023), ED School Climate Surveys (EDSCLS). As of September 28, 2023: https://safesupportivelearning.ed.gov/edscls/administration

North Carolina Department of Public Instruction (2023), North Carolina Teacher Working Conditions Survey. As of September 28, 2023: https://nctwcs.org/index.html

Preston, C. C., & Colman, A. M. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica*, Vol. 104, No. 1, 2000, pp. 1-15.

Robbins, M. and Grant, D. *RAND American Educator Panels Technical Description*. RAND Corporation, RR-3104, 2020. As of December 26, 2023: https://www.rand.org/pubs/research_reports/RR3104.html

Samejima, F. "Estimation of Latent Trait Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph*, No. 17, 1969.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. "Does the Number of Response Options Matter? Psychometric Perspectives Using Personality Questionnaire Data." *Psychological Assessment*, Vol. 31, No. 4, 2019, 557-566.

Weng, L. J. "Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-retest Reliability." *Educational and Psychological Measurement*, Vol. 64, No. 6, 2004, pp. 956-972.