A peer reviewed, open-access electronic journal: ISSN 1531-7714

Increasing the Generalizability of Open-Ended Survey Item Responses

Emily Diaz, *Westat*
John H. Hitchcock, *University of Minnesota*
Greg Norman, *Westat*

Abstract: This article demonstrates a procedure for applying weights to data gathered from open-ended survey items. Open-ended survey items can provide detailed insights into phenomena in ways evaluators may not have anticipated when conceptualizing evaluands and research questions. However, these item types produce text data, which complicates analysts' ability to generalize related findings to a target population of interest. Such items also tend to increase respondent burden relative to closed-ended items because they require written responses instead of simply selecting from a list of options (e.g., using a Likert scale). Such increased burden can lead to higher rates of item non-response, which can further hinder evaluators' ability to generalize findings from a sample to a target population. These challenges may be partially addressed by applying survey weights to numerically coded open-ended responses; however, there is limited guidance on how to do so. This article therefore demonstrates how to efficiently code text data and apply survey weights to subsequent numerical codes. This demonstration is presented within the context of a study that employed stratified random sampling and experienced survey non-response. This article should help evaluators and researchers better generalize findings from numerically coded open-ended survey items to their target populations.

Keywords: Survey; Qualitative coding; Generalizability; Stratified random sampling

Introduction

Open-ended survey items represent one of two main question-response formats, the other being closed-ended options (Dillman et al., 2014; Fowler, 2013). There are two significant advantages to using open-ended items in research and evaluation: (a) respondents are able to express thoughts on topics that may not

otherwise be covered in the survey, and (b) these elicit more detailed responses than closed-ended options. However, there are also at least two key disadvantages to using open-ended survey items. The first is that these items typically require more effort from respondents, which can lead to higher non-response rates (Dillman et al., 2014). Many readers will recognize this phenomenon when reflecting on their own survey response behavior. For example, consider situations where you might have skipped an open-ended question (e.g., *What other thoughts do you have about your experience*?), response fatigue, time constraints, or the belief that you had already invested enough effort into the survey might have led to your decision not to answer the question. Such non-responses can hinder evaluators' ability to generalize findings from open-ended items to a larger population of interest. The second drawback of using open-ended items is that analyzing responses requires more effort than analyzing closed-ended options. This is because text responses must be read and coded. Additionally, there is limited guidance on how to analyze this type of data, particularly regarding the weighting of responses.

Article Purpose and Delimitations

To help address challenges with working with open-ended response data gathered through surveys, this article outlines an approach for straightforward coding of text (which many evaluators already use; Patton, 2015; Saldaña, 2021) and then applying sampling weights based on drawing a nationally representative sample from the population to these data (which appears to be a newer idea[1]). Combined, these techniques can help evaluators make fuller use of data drawn from open-ended survey items.

We demonstrate how we used this procedure in a national study, which examined the extent to which summer learning programs, often referred to as summer school, were implemented across U.S. school districts in 2021, during the COVID-19 pandemic. We focus this article on handling text data rather than the overall study findings because we assume PARE readers are primarily interested in study methods rather than topic-specific findings.[2] Furthermore, this article does not provide a detailed overview of how to construct survey weights. Establishing such weights to account for sampling strata and non-response is a highly idiosyncratic and technical process that is well-covered in the literature (e.g., Johnson & Elliott, 1998; Osborne, 2011); we, therefore, focus on applying weights to coded text data to inspire other researchers to make the most out of their open-ended response findings. We do, however, provide some study details to demonstrate the importance of weighting. Table 1 provides an advanced organizer covering the methodological steps and considerations needed to establish a weighting scheme for text data.

Details on the Merits of Open-Ended Survey Items

The merits of including open-ended items in a survey are straightforward. Consider a study focusing on nutrition and public health. A survey item might query respondents about their preference for certain types of fruit using a rank-choice approach (see Finch, 2022, for an overview of related analytic considerations).

---

[1] We conducted a brief search in Google Scholar using phrases like "weighting text data" and "applying survey weights to text data," in February 2026. Depending on the phrase we used, this search identified articles describing qualitative analyses, complex discussions of survey weights, or no matches at all. There might be methodologically oriented articles describing how to apply survey weights to coded text data, but they are not easily located.

[2] Readers interested in learning more about the underlying study about summer programming can see (Crean Davis et al., 2021: National Call to Action for Summer Learning How Did School Districts Respond? | Wallace Foundation).

Table 1. Summary of Design Considerations

Step	Design and Decisions	Considerations	Limitations
1	Design open-ended survey items	Determine the number of topics and gather associated emergent information	Only accounts for variables included in the survey and may lead to increased nonresponse bias
2	Administer survey	Need to understand the population of interest and have data that describe it to include in weighted analyses	Be mindful of the number of survey items to minimize fatigue
3	Code open-ended responses	Use a web-based platform or paper format	May have low responses; need to verify categories based on subject matter expertise and content knowledge; endeavor to make coding categories mutually exclusive and jointly exhaustive
4	Analyze open-ended	Conduct reliability and validity checks Consider whether to code each response to the one category that is most representative or multiple categories Could use AI, machine learning, or natural language processing to create categories	May need to adjust weights due to missing data
5	Apply survey weights to coded text data	Need statistical software and related technical expertise; recommend conducting quality control checks (we conducted analyses in Stata and verified using Westat's proprietary software "Wesvar.")	Depending on the proportion of responses relative to the total sample size, the weights may not be able to appropriately adjust so they are representative of the population of inference
6	Interpret findings	Need to understand the population of interest have data on descriptive variables to include in weighting	Can only account for variables included in the sampling frame and survey

The ranking item(s) might cover respondent preferences for various fruits, including apples, oranges, bananas, grapes, cherries, berries, peaches, pears, plums, and watermelons. Suppose this item does not, however, offer respondents the opportunity to rank mangoes and kiwis, which may be the top choices among the target population. Failure to list top preferences would represent a survey limitation. One could expand the list to include more types of fruit to address this limitation, but doing so increases the risk that a respondent will not concentrate when completing the more extended task. Respondents may even skip the item altogether if they find it overwhelming.[3] The evaluator might then encounter potentially serious biases, such as cultural biases, and miss opportunities to fully understand a group of respondents and their nutrition (Dillman et al., 2014; Kalkbrenner, 2021). To address this concern, one could simply present an open-ended response item, such as "*What fruits do you like?*"

[3] A brief Internet search shows there are more than 2,000 types of fruit.

Per this line of thinking, one can see that open-ended items elicit emergent information, allowing for a potentially more detailed understanding of phenomena (e.g., some subgroups might adopt healthier dietary habits if they can access culturally preferred foods or eat what is locally available, which could be too extensive to list in one item). Adding an open-ended item may also enhance the survey's cultural sensitivity, supporting the social validity of the findings (i.e., the extent to which respondents value the information derived from the study) and the implications of the findings for developing future interventions and policies (Hitchcock et al., 2015). There are, of course, a variety of open-ended questions with various trade-offs. An evaluator could, for example, elicit short factual statements in the context of a list or request longer-narrative responses. Longer responses are likely to offer helpful detail, but it may be difficult to summarize text into categories. In such instances, we recommend conducting an in-depth thematic analysis informed by evaluation goals. Saldaña (2021) demonstrates there is an extensive list of strategies that may be paired with the evaluation at hand. However, as we established at the outset of this article, one might expect a request for long, narrative responses to be undermined by low response rates, and analysts will need to deal with the resulting text data. Using these data in standard statistical analyses typically entails additional coding steps and applying a numerical value to the information (Sandelowski et al., 2009). This, in turn, requires establishing intercoder reliability (e.g., Cohen, 1960; McHugh, 2012) to ensure numerical coding is consistent and follows established rules (e.g., does one code a tomato as a fruit?). Once a reliable coding procedure is established, survey weights can be applied to numerically coded responses to enhance the generalizability of sample findings to a target population of interest (Johnson & Elliott, 1998; Osborne, 2011).

Generalizing Findings from Open-Ended Survey Responses

Evaluators often want to assess the generalizability of their findings (Bell et al., 2016; Fay & Olsen, n.d.; Olsen et al., 2013; Olsen & Orr, 2016; Tipton, 2013; Tipton & Olsen, 2018; Tipton & Olsen, 2022) even as they gather emergent information (Patton, 2015) that often comes in the form of text data. The term *emergent* refers to the idea that evaluations can be adjusted as new insights about the phenomena of interest are discovered, potentially leading to new lines of inquiry and changes to the evaluation design. If evaluators are working within a multiphase process (e.g., an evaluator might use a survey to identify challenges that respondents face when implementing an intervention and, in the process, uncover concerns that had not been previously considered), they might want to understand how prevalent the newly identified challenge is within a specific target population. However, assume the response rate for open-ended survey items is low. In such cases, one possible solution is to apply survey weights (Johnson & Elliott, 1998; Osborne, 2011) to the text survey data containing the emergent information. Furthermore, even in the absence of a multi-phase study, researchers and evaluators may wish to assess the extent to which open-ended responses from a sample generalize to the population of interest. Overall, there can be value in applying weights to coded text data.

An Example

We used open-ended survey items in a national survey of school districts to understand the summer programming offered in 2021. This year, U.S. schools implemented widespread, unplanned remote instruction in response to the COVID-19 pandemic. We aimed to understand the prevalence of summer school in this context. We therefore generated a nationally representative sample of traditional public and charter districts, also referred to as Local Education Agencies (LEAs), drawn from the U.S. Department of Education's Common Core of Data (CCD; National Center for Education Statistics, n.d.).

We used stratified random sampling. Strata were based on the variables (a) school district poverty and (b) district size. We categorized districts into two poverty levels based on CCD's estimated proportion of children living in poverty within each district. High-poverty districts were those in the quartile with the

highest poverty rate. The remaining districts were classified as low poverty. Our district size categorization was based on student enrollment, encompassing seven categories (i.e., from exceptionally large to small). We supplemented this approach by further stratifying on (a) district urbanicity, (b) whether the district was a charter district, and (c) the proportions of students in different race and ethnicity categories. We selected 550 school districts from the 17,218 districts in the sample frame. To ensure that high-poverty communities were included in the sample, we oversampled LEAs that served such communities, resulting in 220 (40%) of these districts in our sample. In sum, this design yielded a nationally representative sample of school districts, with an oversample of high-poverty districts; furthermore, the sampling approach explicitly controlled for district poverty and size.

Survey Topics

Our closed-ended survey items addressed various aspects of LEA staff's summer learning programming and whether they had begun planning summer activities for the following year. Using open-ended response items, the survey asked questions about:

- the summer programming activities LEA staff would continue to deploy in future years,
- what LEA staff would change about their 2021 programming,
- experiences that would be helpful for summer program staff to consider when planning or delivering future summer programming and
- any information that would be important to share with other LEAs to impart a sense of lessons learned.

Statistical Weighting

Our use of stratified random sampling ensured that the sample had different proportions of LEAs (again, based on factors such as poverty level and urbanicity) than national population. As a result, the basic descriptive statistics derived from the sample would be over-representative of high-poverty districts and would not accurately reflect the nation as a whole. We, therefore, incorporated sampling weights when analyzing survey data (e.g., Lee & Forthofer, 2005; Osborne, 2011). Weighting involved applying statistical corrections to account for sample characteristics considered in stratified sampling and to address survey non-response, ensuring the findings accurately represent the larger population of interest. This is a standard analysis technique in survey work, designed to yield representative findings, as weighting adjusts standard errors to support statistical inference (Johnson & Elliott, 1998; Osborne, 2011). For example, standard errors can appear smaller when conducting unweighted analyses, leading to the incorrect rejection of a null hypothesis and resulting in a Type I error (Johnson & Elliott, 1998; Osborne, 2011).

In our study, weighting reallocated the proportion and distribution of LEAs in the sample to represent the population from which it was drawn. We produced district weights that represented the school districts in the sampling frame by creating base weights that incorporated each sampled district's chance of being selected for the study. Full sample weights for responding districts were created by adjusting their base weights to account for nonresponding districts. This was done using a method called *raking* (Wolter, 2007). Raking uses characteristics from the population of interest to adjust respondent base weights for key variables (control variables) so that the sum of their final weights is equal to the known population totals (control totals).[4] Raking is an iterative process and proceeds sequentially from one control variable to the next. After the last of the variables is controlled, the prior variables may not match their control totals established at the beginning of the process. Thus, raking consists of repeating this process until the control totals for all control variables match. In our study, district characteristics found in the sample frame were

[4] Post-stratification is a special case of raking, where there is only one control variable.

used to create the full sample weights. Jackknife replicates were also created to facilitate variance estimation and produce confidence intervals that account for the sampling design. A detailed discussion of replication and the jackknife procedure can be found in Appendix A of the WesVar Users Guide.[5]

Base Weights

Mathematically, base weights are the reciprocal of a sampled unit's probability of selection. In this study, the district-level base weight was equal to the inverse of the districts' probability of selection. The sum of these base weights yielded an unbiased estimate of the total number of districts. Replicates were subsamples of the full sample used to estimate variance. Finite population corrections were incorporated into the replicate weights following a procedure developed for the National Assessment of Education Progress (NAEP); see Rizzo and Rust (2011).

Calibration Adjustments

In addition to developing weights, raking provides a type of calibration (Deville & Sarndal, 1992) method used to adjust weights. Replication methods work by dividing the sample into subsample replicates that mirror the sample design. A weight is calculated for each replicate using the same procedures as for the full-sample weight. The district-level base weights were calibrated to known control totals according to the raking procedure. The auxiliary control totals were generated from the sampling frame, which contained information about all school districts. The full sample and replicate base weights were raked to control totals for several dimensions. These dimensions were created separately for two groups of districts: (a) high-poverty districts and (b) all other districts. Within these two district types, the raking dimensions were:

- Percent minority (quartiles within census regions; 1—lowest minority quartile, 4—highest minority quartile)
- Urbanicity (four cells: 1—Central City; 2—Urban Fringe; 3—Town; 4—Rural)
- Charter school status (two cells; only charter schools, all other districts)
- District size class (seven cells)
- Census Region (four cells).

The raking process was done separately for the full sample and the replicates, using the same control totals. The final district-level (raked) weights were calculated by applying raking adjustment factors to the full sample and replicate base weights.

Demonstrating the Use of Weights with Standard Numerical Data

Before we describe how we used weights with text data, we present a standard statistical analysis to demonstrate the importance of weighted analyses. In our survey study, we asked LEAs about the number of students they served during Summer 2021. Table 2 presents the related descriptive statistics (unweighted and weighted).

As shown in Table 2, the number of students served varies by a factor of at least 3, depending on the data analysis technique used. A key point of the table is to demonstrate that the weighted average should be used; had the unweighted average been used, a vastly different conclusion about the average number of students served would have been made. This example addresses one important type of validity in research, statistical conclusion validity (Shadish et al., 2002), because policy decisions or implications based on unweighted data would not be warranted.

[5] https://www.westat.com/our-work/information-systems/wesvar-support/wesvar-documentation

Table 2. Weighted and unweighted descriptives of the number of students served with Summer 2021 programming

Data Analysis Method	N	Mean	Standard Error	Wilson 95% Confidence Interval around the Mean	
Unweighted	193	1,680.26	237.06	1,212.68	2,147.85
Weighted	193	501.83	48.23	421.68	619.63

Numerical Coding of Text Data and Establishing Intercoder Agreement

We used district-level weights to analyze all numerical data (i.e., data derived from closed-ended survey items) and applied these weights to our analyses of text data. Doing so allowed us to consider the weighted proportion of districts that provided a given response type when responding to an open-ended item. This provided the simultaneous advantage of gathering emergent, exploratory information about summer school programming while deepening our understanding of the proportion of U.S. districts that might have responded with similar answers. We applied a simple numerical coding scheme, in which we quantified open-ended responses and then subjected the coded text data to weighted analyses. To generate numerical codes, two coders reviewed all open-ended survey responses to each item and coded all text responses into a single category (e.g., the "x" response pertains to summer experiences that would be helpful to consider when planning or delivering future programming). The coders then assigned a numeric code to each response category (i.e., Category 1, 2, 3…); 46 codes captured the entirety of the text data.

The two coders initially demonstrated high agreement, likely facilitated by the fact that open-ended items yielded similar responses, as the items asked respondents to comment on a narrow topic (e.g., *what information about summer school programming would be important to share with other LEAs*?). When an initial disagreement arose between the two coders, they discussed their coding rationale until they reached a consensus (i.e., consensus coding; Saldaña, 2021). Disagreements were sometimes solved by creating a new coding category. By implication, the coding process yielded mutually exclusive codes (i.e., a text response was not coded into more than one category) and jointly exhaustive (i.e., the researchers accounted for all usable[6] text data), with perfect agreement rates, all of which is characteristic of a high-quality coding approach (e.g., Patton, 2015).

Another Approach to Intercoder Agreement

Although not used in the example study presented in this article, an alternative approach to assessing coding reliability across multiple individuals is to establish interrater reliability (Cohen, 1960; McHugh, 2012). Cohen's kappa can be used to assess such interrater reliability (Cohen, 1960). Cohen developed this statistic to address what we saw as a shortcoming of prior methods, which relied on percent agreement among raters. He contended that raters might make guesses, necessitating correction. Kappa, therefore, indicates how often different raters assign the same score to the same variable, accounts for guessing, and ranges from -1 to 1 (similar to a correlation coefficient), where 0 indicates agreement due to random chance and 1 indicates perfect agreement. Cohen suggested Kappa results be interpreted as follows: values ≤ 0 indicate "no agreement," values between 0.01–0.20 as "none to slight," 0.21–0.40 as "fair," 0.41– 0.60 as "moderate," 0.61–0.80 as "substantial," and 0.81–1.00 as "almost perfect agreement" (Cohen, 1960; McHugh, 2012). For a more detailed understanding of the nuances and mechanics of interrater reliability, see Cohen (1960), Rudner (1991), and Stemler (2004).

[6] Some text data were incomprehensible. For more information, see the subsection below on coding challenges.

Coding Results in Our Study

Table 3 presents examples of open-ended responses and their categorization to illustrate this process for the three most frequently coded categories from the text data derived from one open-ended item.

Table 3. A sample of open-ended responses and their categorization to the item: *If you could keep only one thing about Summer 2021 programming, what would it be*?

Code	Sample Open-Ended Response	Category
1	Quality hands on problem-based instruction.	Quality
2	Continue to offer well-rounded opportunities for students to engage or re-engage them in learning.	
3	Small group instruction based on data.	
4	The field trip experiences.	Experience/ Program Offerings
5	The STEM and Reading summer camps were very successful.	
6	The option to participate virtually or in person.	
7	Offering the summer intervention program at all our school sites.	Expanded Reach
8	The programs and activities that were implemented with a new vision.	
9	The use of technology that the teachers used to reach students that [sic] were participating remotely.	

Note: This sample illustrates our method for coding the top three most frequently endorsed options.

As displayed in Table 3, when asked *If you could keep only one thing about Summer 2021 programming, what would it be?* One response was, "The STEM and Reading summer camps were very successful." We categorized this text under "Experience/Program Offerings." Additional text responses coded into this category were "The field trip experiences" and "The option to participate virtually or in person." Another set of responses to this item was numerically coded into a separate category of "Program quality." This included responses such as:

- "[sic] Providing quality, hands on problem-based instruction,[7]"
- "Continue to offer well rounded opportunities for students to engage or re-engage them in learning," and
- "Continue using the state's curriculum which supported acceleration with just in time interventions."

Coding Challenges

We found our coding and subsequent weighted analyses to be straightforward and efficient. We did, however, experience challenges when summarizing long responses into a single category. For example, we received some detailed responses to the item: *Based on your district's experience in Summer 2021, what should other LEAs or charter schools consider when it comes to planning or delivering programming in Summer 2022?* Two examples follow:

[7] For context, problem-based instruction, also known as problem-based learning (PBL), is a student-centered teaching method where students learn about a subject by working collaboratively to solve open-ended, real-world problems (e.g., Reed et al., 2021); they solve real-world problems together.

- "Offer a variety of programming, some that is designed to meet the needs of specific students and others that are open to all students. It is important to have qualified staff, who know the curriculum and can provide effective instruction, deliver the programming."
- "Summer learning days should be full days (8:00 am - 3:00 pm) and include both academic and enrichment opportunities for EVERY student. The summer schedule should be Monday - Friday. The length of the program should be 4-5 weeks. Teacher development is critical. Professional development prior to the beginning of the program, followed by ongoing coaching and support maximizes teacher effectiveness and increases student outcomes. Hands-on, project-based learning keeps students engaged, increases retention of concepts, promotes social-emotional development, and decreases disruptive student behavior issues."

We coded the first response as "Offerings" because, although the respondent also mentioned qualified staff, the text primarily focused on program offerings. We coded the second response as "Planning." This presented a challenge because the respondent also mentioned professional development and tangible aspects of the program, such as hands-on and project-based learning. However, most of this text described summer programming planning, so we made a judgment call, facilitated by consensus coding. We assume evaluators using this general technique might also be challenged when coding long text responses. On the other end of the spectrum, some responses were too pithy or otherwise incomprehensible. We did not have the necessary resources to gather follow-up information or further clarify the respondent's meaning. We therefore did not use these responses in our analyses.

Our approach generally worked well, as the responses were reflective of a single category. In other instances, more complex, in-depth responses might entail using multiple categories when coding. Practically speaking, a strong coding approach can support both the use of rich open-text data and analytic generalizability. As we mentioned earlier in this article, one way to judge a coding scheme is to consider:

- the extent to which codes are mutually exclusive
- the extent to which the codes are jointly exhaustive
- and if all codes may be used in weighted analyses, in part by considering the extent to which a numerical code summarizes text data.

These countervailing considerations may necessitate trade-offs during coding. At an extreme, assigning a unique code to each meaningful segment of text data would yield a set of codes that are mutually exclusive and jointly exhaustive, but all this would achieve is translating data from one form (text) to another (numerical codes) without any aggregation. As a thought experiment, a parallel approach in quantitative work would be to conduct a factor analysis that yields as many factors as there are survey items, which would not be useful. At the other extreme, overly broad codes might account for all the data, but the codes would overlap (e.g., phrases might focus on summer programming quality and expanding reach to students, two meaningfully different ideas, into one superordinate code that would be difficult to interpret). Analysts will need to strike a good balance, which we think is simpler when dealing with short segments of text and more complex as responses grow longer. Fortunately, there is longstanding and extensive guidance in qualitative analysis sources (e.g., Saldaña, 2021; Sandelowski et al., 2009) that can inform analysts' consideration of their research purpose, the richness of their data, and the best use of the study findings.

Applying Weights to the Numerically Coded Data

Once text data are numerically coded, survey weights may be applied. The aforementioned sampling strata and survey non-response informed our development of weights. Another complexity is that non-response differed by survey item, and for some items, it exceeded 60%. With that in mind, the

appropriateness of the methods we propose here is governed by the missing data literature (e.g., Little & Rubin, 2019; Schafer, 1997; van Buuren, 2018), in which we assess whether data are missing completely at random. If data are missing completely at random, the weighting methods are appropriate without using additional techniques. The penalty of missing information simply manifests as wider confidence intervals. However, if missingness is related to some measured variable (e.g., district size), then imputation approaches can be used to account for item nonresponse, but we found it fair to assume data were missing completely at random. This left another concern on the table: in cases where there are more non-respondents than respondents, weighting can be problematic because those with novel responses might be over- or under-represented. However, we argue that applying a method to account for nonresponse is better than ignoring the issue and not using open-ended survey data.

We conducted analyses in Stata and administered quality control checks using Westat's proprietary software "Wesvar." Following this step, we analyzed the categorical variables using district weights in the same way we analyzed data from the closed-response items. Our focus was on determining the proportion of districts responding to an open-ended item that reflected a specific code. Table 4 provides the weighted estimates for the top three most frequently endorsed response categories shown in Table 3. The weighted analysis yielded a confidence interval, which helped demonstrate our confidence in the point estimate.

Table 4. Weighted frequencies for the top three most frequently endorsed categories responding to the item: *If you could keep only one thing about Summer 2021 programming, what would it be?*

Category	Proportion	Jackknife Standard Error	Wilson 95% Confidence Interval	
Quality	0.269	0.083	0.140	0.453
Experience/Program Offerings	0.226	0.046	0.149	0.327
Expanded Reach	0.132	0.042	0.069	0.237

Consider, for example, that, when answering the item: *If you could keep only one thing about Summer 2021 programming, what would it be?* about 27% of responding districts (Wilson 95% Confidence Interval 0.140 to 0.453) provided a response focusing on program quality. In other words, about a quarter of responding districts described elements of summer programming quality as a feature to maintain in future years. Critically, our use of weights gives us some confidence that we can generalize emergent text findings about program quality to the population of school districts, as is the case for the other three themes in Table 3.

Alternative Methods

Although we accounted for item non-response through weighting, we could have employed other strategies for handling missing data, such as multiple imputation (Little & Rubin, 2019; Rubright et al., 2014). Other researchers should consider using these techniques to assess the extent to which their preferred method for handling missing data yields findings that are consistent with those from other approaches. However, in the approach we present in this article, we focused on open-ended item responses, which inherently yield text data, and there was a higher non-response rate across these items relative to closed-ended items. Importantly, had we received more responses to open-ended items, we might have generated a new response category not previously identified. Moreover, we did not use any technique to simulate a response when it was missing, as our purpose was to analyze the data in the respondent's voice and from their perspective. Any imputed missing data would likely result in researcher-induced error or bias, as we would be speculating about the response. Clearly, this is a limitation to our methods. On the other hand, we

propose applying survey weights to coded text data to maximize their utility without resorting to imputing emergent information. In the end, this method can take analyses of text data a step further than what we expect most analysts pursue, and we recommend being transparent regarding the proportion of responses when applying this approach.

Discussion and Future Directions

This article presents a procedure that applies survey weights to open-ended survey items, accounting for stratified random sampling and missing data, to yield more statistically generalizable findings than presenting open-ended findings without weighting. The procedure we presented here enabled us to gather emergent information that we did not consider when using closed-ended items, while still allowing us to generalize findings from text data more effectively to our population of interest. We hope this demonstration will inspire other survey methodologists and researchers to adopt a similar approach, to maximize the potential of text data collected through open-ended items. We do not see our method as a panacea for balancing the need to gather open-ended responses with generalizing findings. After all, without replicating the survey, we have no way to assess whether respondents consistently provide the same information when answering open-ended items. Rather, we view the methods presented here as a means to enhance the utilization of existing text data.

The survey we deployed included both closed and open-response items to identify overall themes and gain more nuanced insights into specific subtopics. The open-response format allowed respondents to provide detailed feedback that was not captured in the closed-response items. These survey findings add context and depth, enabling respondents to share information. It would have been easy to treat the text data as strictly supplemental and descriptive without any attempt to ascertain the extent to which they represented our target population of interest—the nation's school districts. However, by numerically coding the text, we could leverage district weights (accounting for both stratified sampling and survey non-response), thereby increasing their usefulness. Weighting helps adjust for unequal chances of selection and differential nonresponse, but it can also increase variance and reduce statistical precision—especially with high item nonresponse. In situations with high missingness, there are important research considerations as the sample may be too small or too selective, which can undermine both statistical power and representativeness. In these cases, researchers must weigh the tradeoffs between producing unbiased but highly variable weighted estimates versus restricting analyses, using imputation strategies, or focusing on outcomes with sufficiently complete data to preserve reliability. In sum, readers should now see pathways toward applying weights to numerically coded text data gathered via surveys.

We summarize scenarios in which we can envision evaluators and researchers applying this technique.

1. **Enhance representativeness of text data**. This approach is the one we demonstrated in this article, and it can enhance the generalization of text-based findings.
2. **Improve policy-relevant insights in national or statewide surveys**. Weighted survey findings based on open-ended responses can be used to answer questions such as

a. "How prevalent is this need?"

b. "What types of districts are most affected?"

c. "Are these sentiments more common for certain subgroups (e.g., regional or rural)?"

The answers to these questions will provide valuable insights into the frequency of occurrences and their impact across different groups.

3. **Support rapid response research**. By design, one feature of open-ended items is to allow researchers to explore emerging issues not anticipated during item construction. By incorporating weights into the analysis of categorized qualitative data, researchers may be made aware of new challenges or innovations, identify areas of need, and prioritize follow-up interventions and research. This will be helpful for research in fast-changing environments (e.g., technology adoption, teacher workforce shortages, medical and pharmaceutical research).
4. **Strengthen evaluations of programs and interventions.** Evaluators use open-ended items to capture a variety of information, ranging from the benefits of program participation to implementation challenges. The use of weights allows evaluators to estimate the prevalence of specific challenges in a population, compare the benefits across the treatment and comparison groups, and determine how findings are reported by subgroups. These answers provide crucial context, enhance the usefulness of program planning, and support a continuous improvement mindset.
5. **Enable more robust subgroup analyses.** To ensure that high-poverty communities were included in the sample, we oversampled LEAs that served such communities, resulting in 220 (40%) districts in our sample. We then needed to weight our analyses to correct for this sample imbalance. Oversampling high-poverty communities helped ensure we had this important subgroup in our sample and enabled us to assess summer programming in these districts.
6. **Encourage More Strategic Use of Open-Ended Items.** Researchers can include more open-ended items in surveys, knowing their findings can be weighted and reflect the population of inference, rather than those who responded. Incorporating these data-driven insights into the design phase will allow a blend of exploratory and confirmatory design options when creating surveys.

Received: 4/20/2025. **Accepted:** 3/22/2026. **Published:** 4/21/2026.

Citation: Diaz, E., Hitchcock, J., & Norman, G. (2026). Increasing the generalizability of open-ended survey item responses. *Practical Assessment, Research, & Evaluation*, *31*(1)(9). Available online: https://doi.org/10.7275/pare.3068

Corresponding Author: Emily Diaz, Westat. Email: emilydiaz@westat.com

References

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis, 38*(3), 318-335. https://doi.org/10.3102/0162373715617549

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Crean Davis, A., Hitchcock, J., Tek, B.-A., Diaz, E., & Hershey-Arista, M. (2022). National call to action for summer learning: How did school districts respond? Wallace Foundation. https://wallacefoundation.org/report/national-call-action-summer-learning-how-did-school-districts-respond-how-did-school

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*(376), 376-382. https://doi.org/10.1080/01621459.1992.10475217

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Fay, R. E., & Olsen, R. B. (n.d.). Comparing balanced and probability site selection methods for randomized trials in education. Manuscript submitted for publication.

Finch, H. (2022). An introduction to the analysis of ranked response data. *Practical Assessment, Research, and Evaluation, 27*(1), 7. https://doi.org/10.7275/tgkh-qk47

Fowler, F. J. (2013). *Survey research methods*. Sage.

Hitchcock, J. H., Onwuegbuzie, A. J., & Khoshaim, H. B. (2015). Examining the consequential validity of standardized examinations via public perceptions: a review of mixed methods survey design considerations. *International Journal of Multiple Research Approaches, 9*(1), 24–39. https://doi.org/10.1080/18340806.2015.1076757

Johnson, D. R., & Elliott, L. A. (1998). Sampling design effects: Do they affect the analyses of data from the National Survey of Families and Households? *Journal of Marriage and Family, 60*(4), 993-1001. https://doi.org/10.2307/353640

Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The MEASURE Approach. *Practical Assessment, Research, and Evaluation, 26*(1), 1. https://doi.org/10.7275/svg4-e671

Lee, E. S., & Forthofer, R. N. (2005). *Analyzing complex survey data*. Sage.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

National Center for Education Statistics. (n.d.). Common core of data: America's public schools. Retrieved from https://nces.ed.gov/ccd/ccddata.asp

Olsen, R. B., & Orr, L. L. (2016). On the "where" of social experiments: Selecting more representative samples to inform policy. *New Directions for Evaluation, 2016*, 61-71. https://doi.org/10.1002/pam.21660

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32*(1), 107-121. https://doi.org/10.1002/ev.20207

Osborne, J. (2011). Best practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Practical Assessment, Research & Evaluation, 16*(12). http://pareonline.net/getvn.asp?v=16&n=12

Patton, M. Q. (2015). *Qualitative research and evaluation methods*. Sage.

Reed, S. S., Mullen, C. A., & Boyles, E. T. (2021). *Problem-based learning in elementary school*. Springer International Publishing.

Rizzo, L., & Rust, K. (2011). Development of a first-stage finite population correction in the variance estimator for a two-stage PPS design. *Proceedings of the Section on Survey Methods, American Statistical Association*.

Rubright, J. D., Nandakumar, R., & Gluttin, J. J. (2014). A simulation study of missing data with multiple missing x's. *Practical Assessment, Research & Evaluation, 19*(10). Available online: http://pareonline.net/getvn.asp?v=19&n=10

Rudner, L. M. (1991). Reducing errors due to the use of judges. *Practical Assessment, Research, and Evaluation, 3*(1), 1-3. https://doi.org/10.7275/w4a1-cb66

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research, 3*(3), 208-222. https://doi.org/10.1177/1558689809334210

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*(1), 1-11. https://doi.org/10.7275/96jp-xz07

Tipton, E. (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review, 37*(2), 109-139. https://doi.org/10.1177/0193841X13516324

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher, 47*(8), 516-524. https://doi.org/10.3102/0013189X18781522

Tipton, E., & Olsen, R. B. (2022). Enhancing the generalizability of impact studies in education. *Toolkit. NCEE 2022-003*. National Center for Education Evaluation and Regional Assistance.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC.

Wolter, K. M. (2007). *Introduction to variance estimation*. Springer.