JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

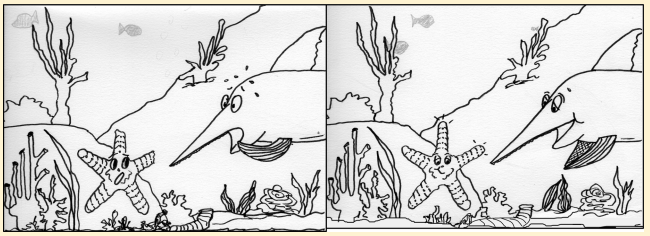
Methods for Addressing Missing Data with Applications from ACS Exams

Alexandra Brandriet and Thomas Holme*

Department of Chemistry, Iowa State University, Ames, Iowa 50011, United States

Supporting Information

ABSTRACT: As part of the ACS Examinations Institute (ACS-EI) national norming process, student performance data sets are collected from professors at colleges and universities from around the United States. Because the data sets are collected on a volunteer basis, the ACS-EI often receives data sets with only students' total scores and without the students' responses to individual exam questions. Nonetheless, several national norming statistics require students' item responses. This data return leads to missing data and potentially biased results when inferences are made based on that data set. This work uses student performance data sets from ACS-EI to consider how methods for replacing missing data, such as hot-deck imputation and simulating data, affect the nature of the analysis of quantitative data.



KEYWORDS: *First-Year Undergraduate/General, Second-Year Undergraduate, Testing/Assessment*

INTRODUCTION

Chemistry instructors often use different types of quantitative data to understand how their students are performing in the classroom. Needless to say, these judgments are difficult to make when there are few, or no, external standards that allow for comparison. Fortunately for chemistry instructors, the existence of ACS exams provides one means for obtaining national comparisons. In order to make this possible, instructors must voluntarily return student performance data, which makes the generalizability of this data an important concern for using ACS exams. Therefore, issues associated with sampling and data analyses warrant specific consideration, and in particular, concerns related to missing data should be incorporated into assessment analyses. One way to shed light on this concept is to examine it from the perspective of ACS exam national normative data sets, which are relevant to all ACS exam users and provide readily visualized challenges associated with data collection. While some of the data analyses herein may not be directly applicable to certain situations, the missing data discussion is relevant to anybody who draws conclusions from assessment data, including chemistry education researchers, chemistry instructors focused on departmental assessment efforts, and chemists who collect data in the laboratory.

Issues related to missing data are not unique to the ACS Examinations Institute (ACS-EI).[1,2] In contrast to data that have been observed and are directly available for analyses, missing data refers to data that are not observed or otherwise nonexistent in the data set of interest.[1,2] As an example, multiple students may skip a question on a test, and therefore, their responses are *missing*. Given the large number of ways that data can be collected in educational settings, it is not surprising that there are multiple contexts in which it can be missing. More importantly, missing data can lead to biased estimates of the construct measured, if it is not appropriately considered.[3] This problem is intensified by the fact that it is rare for the missing data issue to be discussed when studies are disseminated.[4,5] When missing data are discussed, the most frequent ways in which they are handled are through methods that ignore the missing responses or delete respondents with it.[3−5] However, these methods inherently assume that the missing data are randomly scattered throughout the data set, and unfortunately, this does not always hold true.[4] This manuscript has multiple objectives, which include (1) showing the limitations of deleting missing data when they are not missing completely at random, (2) showing how this assumption can influence the individual item results provided by the ACS-EI, and (3) discussing methods that the ACS-EI has used to address missing data. In this case, the idea is analogous to the starfish in the graphical abstract; measurements with missing data may be used to make inferences, but finding reliable ways to replace that data is helpful, just as the starfish is arguably better when it regrows an arm it has lost.

MISSING DATA THEORY

Two methods, listwise deletion (complete-case analysis) and pairwise deletion (available-case analysis), are arguably the most commonly used methods to handle missing data.[3−5] Listwise deletion involves deleting any case (e.g., person or student) in a data set that has any missing responses.[1−8] The key result of using listwise deletion is that all analyses conducted with that data set will have the same sample size.

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

In contrast, pairwise deletion involves deleting any case with missing data within a specific set of variables being analyzed. This method leads to more responses being preserved, but there may be different sample sizes for each analysis conducted. These methods are similar because any analyses conducted using these methods will only include observed responses. Deletion methods have the advantages of being easy to implement, being readily available in most statistical packages, and resulting in a complete data set. However, many argue that the limitations of these methods outweigh the advantages. Most importantly, removing participants with missing data can have a drastic impact on sample size, which reduces statistical power.[1−8] For this reason, deletion methods are recommended when only a small percentage of data are missing.[4] It is also assumed that when cases with missing responses are deleted, this will not have an influence on the overall results because the missing data are randomly distributed throughout the data set. Missing data of this nature is known as missing completely at random (MCAR).

In contrast to MCAR, missing data can also be missing at random (MAR) or missing not at random (MNAR). The distinction between these mechanisms is subtle, but important. Missing data is MAR when the probability of missing data in a variable is related to another observed variable, but not to the variable itself.[1−8] This means that the reason data are missing is not related to the specific variable itself after controlling for another variable. A test known as Little's MCAR test can be used to establish if the data are MCAR,[9] but there are no definitive tests to determine whether the data are MAR or MNAR without knowing the values that are missing. Therefore, it may not be possible to know if the data are truly MAR or MNAR.[3,7,8] The purpose of labeling missing data as MCAR, MAR, or MNAR is to better understand what could have led to the missing data, so that appropriate considerations can be made to reduce the effects on the results. Because these concepts can be confusing, an example of a variable (i.e., chemistry test scores) with MCAR, MAR, and MNAR data is shown in Table 1. This example is analogous to one provided by Enders[3] except that Table 1 is applied to a chemistry education context, and it builds on the idea that many aspects of chemistry require mathematical acuity, so relationships between math and chemistry understanding may be observed in student performance on test items.

In Table 1, the first two columns on the left are hypothetical math (blue) and chemistry (yellow) test scores with no missing data. The level of shading is used to highlight the degree of performance using dark shades for low scores and light shades for high scores. In the remaining columns, the chemistry test score variable is altered three times to show examples of MCAR, MAR, and MNAR data. In these alterations, the student has missed the test (has no score), but for the purpose of the illustration, the score that the student would have had is shown in the column labeled *complete*. As a result, the blue or yellow shades are used to visually show the association (or nonassociation) that the missing scores have with the math or chemistry variables.

The MCAR example has missing data randomly scattered across the chemistry score variable so that it is not related to the math or the chemistry scores. This is depicted by using split blue/yellow shading in the cells showing missing data. Varying shades of yellow and blue signify that the missing data has no apparent association with the math or chemistry performances. It might be surmised that students were missing the chemistry test for nonacademic reasons (e.g., illness or family situations). In the MAR example, the missing test scores are from students with low math performances (indicated by several dark blue cells), but column 2 shows that they are not universally low chemistry scores. In this case, the missing test scores might appear to be completely random with the range of scores evident, but the relationship with the nonchemistry variable (math) is not particularly random, and the data cannot be classified as MCAR. Finally, in the MNAR example, the missing test scores appear for students having low chemistry understanding (indicated by several dark yellow cells) so the missing data is related to the variable itself. These examples show why it is difficult to determine MNAR data in real-world analyses without knowing the missing scores.[3,8] In this illustration, the actual test scores are provided, but this is not the case with real data sets. Further, these examples are highly simplified because there are many reasons why a student might miss a test, beyond poor chemistry and math ability, and in real world analyses, missing data will likely never be purely MCAR, MAR, or MNAR.[7] Therefore, understanding the missing data requires knowing as much as possible about the sample itself.[3]

Table 1. Hypothetical Student Math and Chemistry Scores as Examples To Show Various Patterns of Missing Chemistry Scores[a]

Math Scores	Chemistry Test Scores			
	Complete	MCAR	MAR	MNAR
55	65	65	*	*
56	67	67	*	*
60	91	*	*	91
61	71	71	*	71
63	80	*	*	80
64	75	75	*	75
64	73	73	73	73
65	64	64	64	*
67	59	*	59	*
68	79	79	79	79
70	81	81	81	81
73	86	*	86	86
73	67	67	67	*
74	94	94	94	94
77	89	89	89	89
79	75	75	75	75
83	86	86	86	86
85	70	*	70	70
87	56	56	56	*
88	75	75	75	75
88	93	*	93	93
91	97	97	97	97
96	100	100	100	100
98	89	89	89	89

[a]Missing data points are marked with an asterisk (*). Blue shading indicates the degree of math score from dark blue (low scores) to light blue (high scores). Yellow shading indicates the degree of chemistry score from dark yellow (low scores) to light yellow (high scores).

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

Students	Questions								Total Score
	1	2	3	4	5	6	*etc.*	70	
1	A	D	D	B	C	A	…	A	34
2	B	C	D	A	D	C	…	C	47
3	A	D	C	D	A	C	…	B	56
4	C	A	C	B	B	A	…	D	63
5	D	B	A	A	D	B	…	C	68
6	*	*	*	*	*	*	…	*	56
7	*	*	*	*	*	*	…	*	59
8	*	*	*	*	*	*	…	*	69

Observed or Response Data

Missing Data

Figure 1. Example of observed and missing data for a multiple-choice exam with question response choices A, B, C, and D. Missing data points are marked with an asterisk (*).

MISSING DATA AT THE ACS EXAMS INSTITUTE

The ACS-EI collects student data sets from colleges and universities from around the country for the purposes of nationally norming exams. Several item-level statistics are calculated as part of this process, including difficulty and discrimination indices and the percentage of A, B, C, and D response choices per question. Because the national data sets are collected on a volunteer basis, the ACS-EI often receives data sets that include students' total scores but not their item-level responses. This idea is illustrated in Figure 1, where Students 1−5 have data at both the item and total score levels, while Students 6−8 have total score data but the item data are missing. Such missing data adds complexity in characterizing ACS exams and the item statistics that afford instructors potential insights about their courses. Item-level data also allows the ACS-EI to estimate other characteristics of student performances and capture effects such as item order effects.[10] Therefore, an effort to identify ways to address this missing data becomes important.

The ACS-EI has recently developed the Anchoring Concepts Content Map (ACCM) that was designed to help instructors gain additional insight into student content knowledge based on their ACS exam item performances.[11−13] In response to instructor interests in analyses of student content knowledge associated with things like anchoring concepts, the difference between test scores among data sets with only scores versus all item responses has emerged as an important issue to address. In addition to these practical needs, the ACS-EI data sets illustrate missing data problems and solutions well, because the sample sizes tend to be large and collected from various schools from across the country. Therefore, these data sets provide a relevant template for a discussion of the impact of missing data and methods for addressing it.

MISSING DATA METHODS

Some missing data methods have been classified as single imputation methods because they include filling in or *imputing* individual missing values on a case-by-case basis.[5] Many of these methods require that data is MCAR or MAR for accurate estimates.[1−8,14,15] Two of the most basic single imputation techniques are mean and linear regression imputation. Mean imputation involves replacing missing values with the average of the observed data.[1−8,16] This method is common and easy to implement, but it concentrates the results toward the center of the distribution, and thus can reduce the variability of the data and reduce the potential for correlations among variables.[3] Linear regression imputation uses regression equations to predict missing responses. However, this method can also build artificial correlations into the data and, therefore, reduces the variability that would likely occur if the missing values were observed.[3] Many additional single imputation methods exist and have been discussed in detail elsewhere.[1−8,14,15]

In addition to single imputation, there are a few model-based procedures that have been described as "modern"[5] or "state of the art"[3] in missing data techniques. One method is multiple imputation (MI), which works similarly to the regression-based techniques described previously, except that a residual term is added to help restore a loss in the variability of the estimated data.[8] Multiple data sets with estimated missing data are created and analyzed, and the resulting parameters are pooled together.[8] The drawback of multiple imputation is that it requires specifying a model based on predictive variables where "When in doubt including more variables in the imputation model is better."[17] While this method has potential to be used with ACS exam data sets, the ACS-EI collects very little information about the students beyond their exam performances, and this lack of predictive variables limits the potential for robust regression-based techniques. As a result, this method will not be discussed in this manuscript. Additionally, another popular technique uses a full-information maximum likelihood (ML) method to estimate parameters that have the highest likelihood of producing the results based on both the missing and observed data.[4,8] Unlike the methods described previously, ML estimation is both theoretically complex and alone does not impute values into the missing responses. However, the technique is simple to implement and is readily available for many statistical procedures.[4,8]

Hot-Deck Imputation

Of particular interest to this study is a method known as hot-deck imputation. In this method, missing responses are filled in with values from other "similar" participants in the observed data.[1−3,14,15] Hot-deck procedures are well-known for being used by the U.S. Census Bureau for the Current Population Survey (CPS).[14,15] Hot-deck methods have the benefits of being conceptually simple to understand and easy to implement without creating complex models, and (like other imputation methods) they can provide a common complete data set for use by multiple analysts.[14] Additionally, because the imputed values are derived from actual participants, the values will never be outside the possible range of responses; for example, a value of a 73 will never be imputed for a test where only 70 points are possible.[15] Finally, hot-deck works with categorical missing

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

data,[14,15] such as students' responses to A, B, C, and D multiple-choice exams. However, the hot-deck method is limited by the fact that it creates duplicate responses and can create artificial correlations.[3,14] Also, the hot-deck method can establish an illusion of a complete data set when in fact it imputes *estimates* of actual student responses.[14]

■ ACS EXAMS

For the purpose of this study, two exams released by the ACS-EI have provided examples of missing data problems. The first is the 2012 First-Term General Chemistry Exam (GC12F), and the second is 2010 First-Term Organic Chemistry Exam (OR10F). The GC12F exam has two parallel versions where the content of the questions and multiple-choice responses are the same, but differ in the order in which they are presented; these will be described as the gray and yellow forms. The ACS-EI does not require that gray and yellow versions of the exams be identified for volunteer data return. However, the form version is readily determined when full student response data is returned, and therefore, item statistics can be identified for each form.

■ MISSING DATA PROBLEM

For the purposes of this study, the exam data sets can be thought of as divided into two segments. One segment includes the student data with both total scores and item-level responses and will be referred to as the *observed* or *response data* (illustrated in Figure 1). The other segment includes the student data with total scores but no item-level responses and will be referred to as the *missing data* (illustrated in Figure 1). These labels are based on whether the question responses were observed or missing, because total score information was observed across both segments. Further, missing data also existed within the observed question data because some students skipped specific questions or perhaps were unable to answer all of the questions within the time limits. Because knowing which questions students skipped may be valuable to instructors, these responses were not imputed for the purposes of this study. The work described here focuses exclusively on missing data from the perspective of completely missing question data; that is, no answers were provided at all.

The GC12F and OR10F exam data sets described can be visualized in Figures 2 and 3. The GC12F data set included 10,087 students' total scores from various doctoral, four-year institutions, and community colleges from across the country. A total of 2,331 students (23.1%) were missing question data, and this is represented in blue in Figure 2. In contrast, the OR10F exam had more missing data than observed data. In this case,

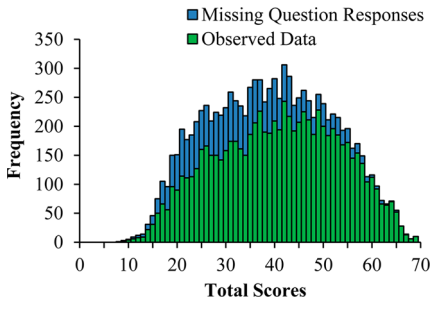


Figure 2. National sample of the students' total scores for the GC12F exam.

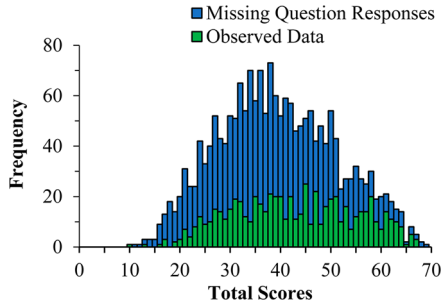


Figure 3. National sample of the students' total scores for the OR10F exam.

the data set included 1,933 students and 1,309 students had missing data. Figure 3 shows the missing data rate of 67.7%, which is extreme. These two exams were chosen because they provide examples of two very different missing data scenarios. For illustration purposes, three different ways to handle missing data were used for these two exam data sets: listwise deletion, hot-deck imputation, and a simulation method assiduously designed for ACS exam data sets. R version 3.1.2[18] and SAS 9.4[19] were used to analyze the data.

■ RESULTS

Listwise Deletion

In both Figures 2 and 3 more data is missing at the lower end of the total score distribution than at the higher end. This is shown by a greater portion of the blue distribution (missing data) represented near the lower end of the total score range, while a larger portion of the green distribution (observed data) is evident at the higher end. Recall that all data was from voluntary returns of student performances, so one possible explanation for the differences lies in the idea that an instructor who finds out that their students have fared poorly may be less inclined to submit their student responses.

An example of listwise deletion was provided by calculating descriptive statistics and percentiles[20] based *only* on the data with observed question responses, and therefore, the students with missing responses were deleted. These results for the descriptive statistics are summarized in Table 2. With this method, a significant portion of the GC12F (23.1%) and OR10F (67.7%) data was missing and deleted. The exam averages increased as a result of the listwise deletion because the missing data was skewed toward the lower end of the total score distribution (shown in Figures 2 and 3).

Table 2. Overall Test Results for the GC12F and OR10F Exams for Data Sets with and without the Missing Question Responses

	GC12F		OR10F	
	Full Data Set (with Missing Data)	Data Set with Listwise Deletion (Observed Data)	Full Data Set (with Missing Data)	Data Set with Listwise Deletion (Observed Data)
N	10,087	7,756	1,933	624
Average	39.2	40.8	39.4	42.6
SD	12.6	12.6	11.7	12.1
Min	8.0	8.0	10.0	10.0
Median	39.0	41.0	39.0	42.0
Max	69.0	69.0	69.0	67.0

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

Figure 4 displays the differences between the percentiles calculated using all of the total score data (including the students with missing question responses) and the listwise deleted (i.e., observed) data sets. Removing the students with missing responses had a very large impact on the percentiles. The differences ranged from 0% to 6% for the GC12F exam and 0% to 11% for the OR10F exam. If students were to be graded based on national percentiles, having differences as large as 6% and 11% would have a noteworthy effect on grades predicated on percentile rankings. In part because data used in the norm process for the ACS-EI is clearly not MCAR, listwise deletion methods tend to have unacceptably large impacts on the results.

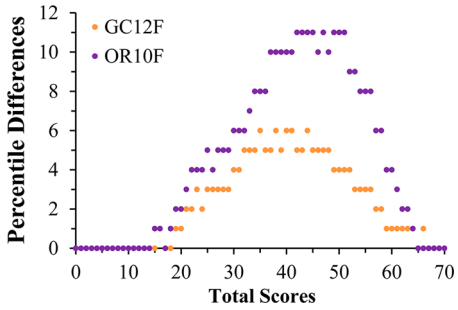


Figure 4. Differences in the percentiles for the full and listwise deleted data sets for the GC12F and OR10F exams.

Hot-Deck Imputation

Another way missing data was addressed was by using a traditional form of hot-deck imputation. In this scenario, the missing values were replaced with the question responses from similar students from the observed data. Similar students were identified by using the students' total scores since total scores were not missing in either the observed or missing data segments. As an example from Figure 1, Students 3 and 6 could be matched because they have the same total scores, and the observed responses from Student 3 could be used to fill in the missing responses for Student 6. The missing values were imputed *without replacement*; this meant that once a string of responses was sampled, they could not be sampled a second time. The purpose of this was to reduce the number of duplicated answer strings within the data set. Unlike with listwise deletion, the distribution of total scores in the final data set stayed the same after the missing item responses were imputed. This factor has the important benefit of allowing exam averages and percentile rankings to stay consistent. Additionally, the sample size was not reduced like it was with listwise deletion.

Figure 5 shows the differences in the difficulty indices calculated for the listwise deleted data set (i.e., observed data) and the hot-deck imputed data set for the gray and yellow versions of the GC12F exam. Difficulty indices are item-level statistics often used with Classical Test Theory to evaluate the data produced by test questions. They are calculated based on the fraction of students who answered a question correctly, and their values can range from 0.0 to 1.0.[21,22] As a result, a question with a high difficulty index has strong student performances and is arguably an easy question, and a question with a low difficulty index is arguably a difficult question. Figure 5 displays a slight decrease in the difficulty indices for the hot-deck imputed data relative to the observed data (or positive differences in difficulty indices). It should be noted that the *y*-axis in Figure 5 is magnified to 0.0 to 0.05 for clarity; however, the differences ultimately could range from −1.0 to 1.0. As a result, the differences appear large, but are actually quite minimal. This minimal decrease is desired because the missing data that was imputed was more commonly found for students in the lower end of the total score distribution, and since lower performing students were better represented in the hot-deck imputation, the difficulty indices decreased.

While the hot-deck imputed data seems to produce reasonable estimates of item statistics, there are several limitations. Perhaps the biggest concern is that this method cannot be easily implemented for data sets with higher percentages of missing data, such as the OR10F data set. When considering the OR10F data set, a couple of issues arose. First, because of the large number of missing values, not all of the students with missing item responses could be matched to students in the observed data set. As an example, Figure 3 shows that there were several students with missing question responses that had a total score of 18, but there were no students with observed question responses that had an 18. Second, the hot-deck method has the limitation of multiple duplicated responses, which would become much more pronounced for the OR10F data set where 67.7% of the sample would be duplicated. The duplicated data would be especially problematic at the tail ends of the distributions where there were so few students with observed response data.

Simulating Missing Data

The final method used incorporated *simulating* or generating, item-by-item, student responses using response profiles (patterns of answers) from the observed data, rather than filling in entire strings of missing responses (i.e., hot-deck imputation). One key advantage of the simulation was that

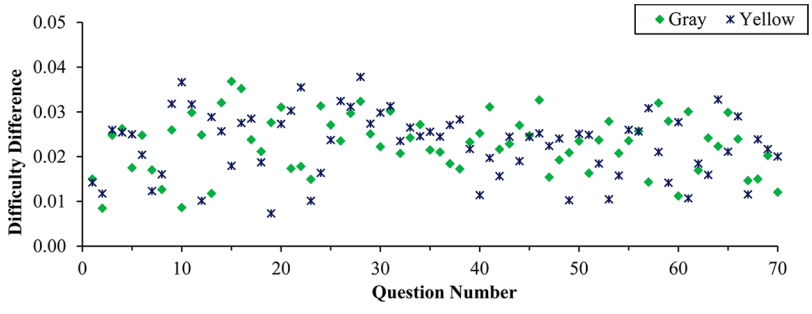


Figure 5. Differences in the difficulty indices calculated for the observed and the hot-deck imputed data sets for the gray and yellow versions of the GC12F exam. The items have been reordered to match content equivalent items.

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

there was no need to match students using a single raw total score. Instead, groups were defined with ranges of scores to simulate student responses. Further, since the data was generated item-by-item, it was unlikely that entire strings of response choices would be duplicated in the data set. As an example, "group 1" could potentially be defined as having the lowest performing students with total scores from 10 to 24. The pattern of answers for Question 1 for the observed student data in the previously defined "group 1" might be 23% "A", 15% "B", 37% "C", 24% "D", and 1% left the question blank. The computer would then randomly select an "A", "B", "C", "D", or blank response from a distribution based on the 23%, 15%, 37%, 24%, and 1% distribution and impute that response for a student with a total score between 10 and 24 that is missing a response for Question 1. This same process would be done for each of the 70 items on the test, thereby simulating the question answer string and filling in the missing responses. The same simulation would then be done for the next higher proficiency group, until all missing data is imputed via this method.

Both the hot-deck and simulation methods filled in missing data based on the response patterns of the observed student data. The differences between the two methods are illustrated in Figure 6. The top arrow in this figure represents the hot-deck

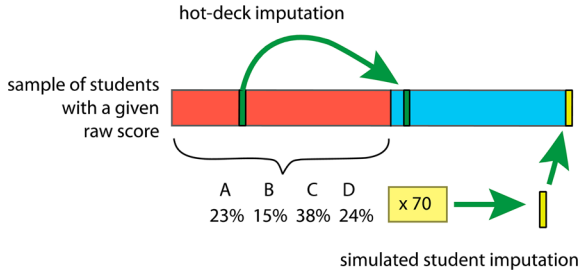


Figure 6. Illustration depicting hot-deck imputation vs the simulating data method. Observed data is shown in red, and missing data is shown in blue.

method where an existing string of student responses (in red) was exactly duplicated to fill in a string of missing responses. The missing response data is shown in blue, and the duplicated response is shown in green. The bottom arrows show the simulation method where the response patterns for all 70 items were used to estimate (item-by-item) simulated responses for students with roughly the same proficiencies (in yellow). Because the question responses for all 70 items were generated based on "A", "B", "C", "D", and blank distributions for each item, the new total scores for the students with missing data did not perfectly match the total scores initially observed. Nevertheless, with careful choices made for the total score ranges used, it was possible to keep the differences small.

Grouping Students by Total Scores. For the simulation method an attempt was made to keep the number of students in each of the groups as close to a constant value as possible, but slight deviations did exist because the number of students that received each total score was not evenly distributed. Therefore, the groups that included the students on the tail ends of the distributions had a larger range of total scores than the students in the center of the distribution, because fewer people existed at tail ends (as shown in Figures 2 and 3). The ranges of total scores and the number of students with observed responses within each group are shown in the Supporting Information. Table 3 provides the descriptive results of three simulated data sets for the OR10F exam in which the students

Table 3. Overall Test Results of the Simulated Data Based on Different Numbers of Groupings for the OR10F Exam

	Original Data Set	3 Groups	6 Groups	12 Groups
N	1,933	1,933	1,933	1,933
Average	39.4	39.9	39.6	39.4
SD	11.7	11.4	11.8	12.0
Min	10.0	10.0	10.0	10.0
Median	39.0	39.0	39.0	39.0
Max	69.0	67.0	68.0	67.0

were arranged into 3, 6, and 12 groups. The overall descriptive statistics for each simulation were similar to the initial total score data.

There is, however, a dependence on the specified ranges of total scores or, in other words, the number of groups of students with different ranges of total scores. Figure 7a shows

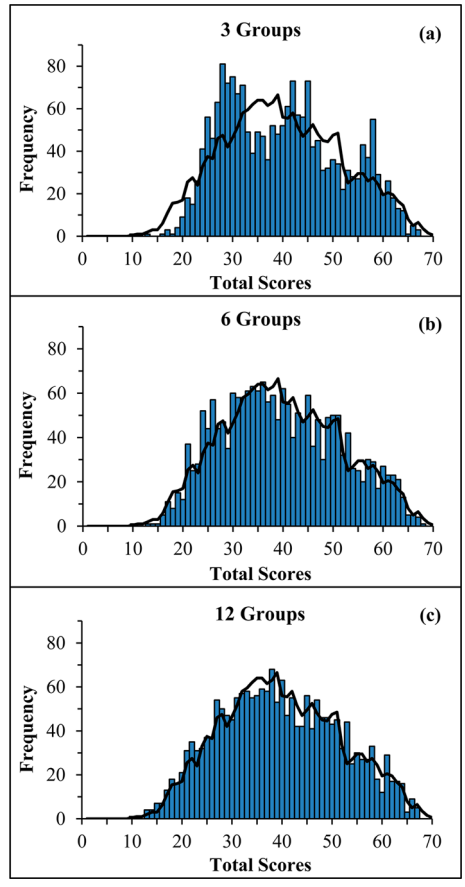


Figure 7. Histograms of simulated data based on different numbers of groupings for the OR10F exam with an overlaid curve of the initial data from Figure 3 which has been smoothed by averaging adjacent pairs of score frequencies.

the results for only three groupings with a corresponding wide range of raw scores, and the simulation resulted in an inaccurate trimodal distribution. However, using more groups with smaller ranges of raw scores produced more comparable distributions to that shown Figure 3. This is illustrated in Figure 7b for 6 groups and Figure 7c for 12 groups. In Figure 7a–c, the black line shows the raw score distribution from Figure 3 that has

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

been smoothed using averaging adjacent pairs of score frequencies.

Because the total scores for the students with imputed missing responses often changed when using the simulation method, the calculated percentiles[20] for student performances also varied. This information was used to adjudicate the quality of the estimates of the student performances that were produced. Figure 8a–c shows the differences in the percentiles

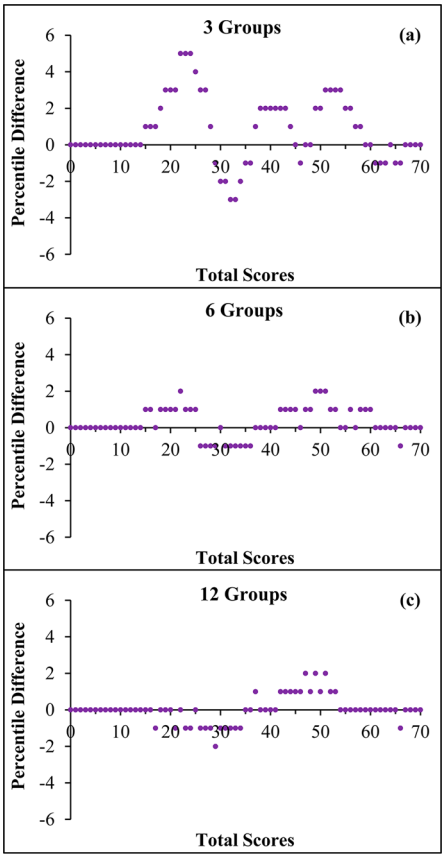


Figure 8. Differences in percentiles between the original data set and simulated data sets for the OR10F exam.

between the initial raw total score data and the values obtained in the three simulations. Again, the 3 group simulation poorly represented the actual data with the differences being as great as 5% (Figure 8a). However, it should be noted that this is still better than the listwise deletion scenario where the percentile differences were as great as 11%. The simulations with 6 and 12 groups performed better with percentiles varying within ±2% (Figures 8b and 8c).

Finally, Figure 9 shows the results of the differences in the difficulty indices between the values from the observed and the three simulated data sets. It should again be noted that the *y*-axis in Figure 9 is magnified to −0.01 to 0.10 for clarity; however, the differences ultimately could range from −1.0 to 1.0. As a result, the differences appear large, but are actually quite minimal. Because the observed data set was weighted toward the students who scored better on the exam, as is shown in Figure 3, it is not surprising that Figure 9 shows a positive difference in difficulty indices for the items. This observation is in line with expectations and is desired because more of the lower performing students were represented in the simulations, but not in the observed data set. The results provided in this section are for the OR10F exam, but similar results are also shown for the GC12F exam in the Supporting Information.

CONCLUSIONS AND IMPLICATIONS

Users of ACS exams have long valued the ability to make comparisons between their own students and those in a national database. Historically, these comparisons have been emphasized on a whole-test basis using percentile performances. Increasingly, departments are being asked to contribute to institutional assessments in ways that go beyond these overall comparison methods.[23,24] Recent reports by the Committee on Professional Training have recommended content guidelines for ACS program approval,[25,26] and ACS exams are excellent tools for assessment of such guidelines. Organizing content knowledge using the ACCM[11−13] represents one way that item-level statistics may help departments in their assessment reporting. Because these methods rely on item-level data, missing data mitigation becomes important to consider. ACS exams are secure tests,[27] so using groups of items, like those categorized within the "big ideas" on the ACCM, allows instructors and researchers to investigate content-based inquires while maintaining the security of individual exam items. As a result, security becomes a key motivation for using improved statistical methods.

Choosing a method to manage missing data is a significant decision that takes consideration. Two techniques, MI and ML, are often recommended,[4,8] and should be considered by data analysts when they are applicable and the data analytic software is available. The limited availability of predictor variables and the need to have a complete data set to work with made these methods less attractive for these purposes. The hot-deck imputation and simulation methods both performed better than listwise deletion, even though deletion methods are very

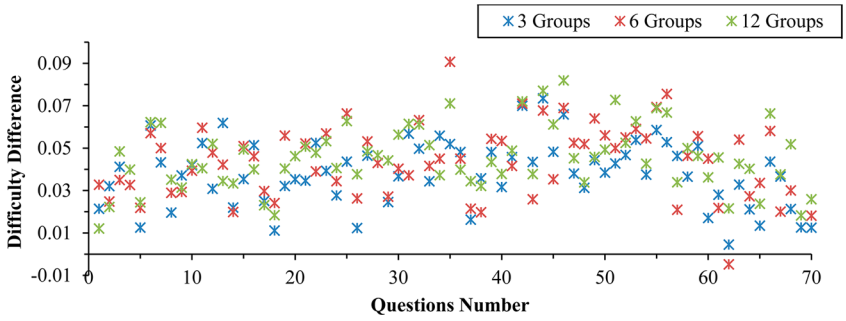


Figure 9. Differences in the difficulty indices calculated for the listwise deleted dataset and the simulated dataset for the OR10F exam.

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX

commonly used in research.[3−5] The methods described are not directly applicable to all situations, but the discussion of missing data is relevant to any reader who draws conclusions from assessment data. For both data sets generated by the ACS-EI, the simulation method performed best, especially when the missing data was imputed from groups of student performances with relatively narrow ranges of total scores. However, a limitation of both of these methods is that they do not account for the potential variability that would have been present if the missing data had been collected, and resulting standard error estimates would likely be attenuated. This simulation was uniquely developed for the missing data scenario experienced by the ACS-EI, but it may also have applications for other data sets. The ACS-EI is developing several data analytic tools for chemistry instructors at this time, and future work may focus on developing workshops or webinars to help train faculty in their use.

In addition to normative data that is partially missing, it is also important to consider the data that the ACS-EI does not receive at all because of the voluntary nature of the data return. For example, it is likely that the ACS-EI may not receive *any* data from some poorly performing classes. This means that there is potentially a subset of the national performance sample where data is missing, and this missingness is directly related to the exam scores themselves (MNAR). Unfortunately, without knowing what data is missing, it is difficult to know if the data is MNAR or how to best handle it.[3] With real-world data, "pure MCAR", "pure MAR", and "pure MNAR" data will probably never exist.[7] This possibility motivates the efforts by ACS-EI to encourage instructors to submit students' scores on newly released exams and, importantly, individual student responses. The ACS-EI has a policy that student item response data will be accepted in any format that can be provided, and instructors who have such data for any recently released ACS exams are routinely encouraged to contact the ACS-EI.

Some may question whether these techniques are just fabricating data. The goal of most quantitative analyses is to use the sample data to *estimate* population results. Filling in missing data based on sound methodological decisions has the potential to better represent the population in comparison to deleting data. Therefore, it is recommended that data analysts make every effort to collect information that may be a good predictor of missing data when it arises. Listwise and pairwise deletion are convenient ways to handle missing responses, but this study has shown that it can appreciably bias results. Because there are inherent limitations in the collection of data for norm calculations at the ACS-EI, any effort to represent missing data will always be an estimate and subsequent usage of the augmented data set should be made with full acknowledgment of this limitation. Ultimately, this consideration is important to anyone who uses ACS exams.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.5b00180.

Simulated data for the GC12F exam (PDF, DOCX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: taholme@iastate.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by NSF DUE #1323288; any opinions, findings, conclusions and/or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

■ REFERENCES

(1) Little, R. J. A.; Rubin, D. B. *Statistical Analysis with Missing Data*, 1st ed.; John Wiley & Sons: New York, NY, 1987.
(2) Little, R. J. A.; Rubin, D. B. *Statistical Analysis with Missing Data*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, 2002.
(3) Enders, C. K. *Applied Missing Data Analysis*; Guilford Press: New York, NY, 2010.
(4) Pigott, T. D. A Review of Methods for Missing Data. *Educ. Res. Eval.* **2001**, *7* (4), 353−383.
(5) Peugh, J. L.; Enders, C. K. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Rev. Educ. Res.* **2004**, *74* (4), 525−556.
(6) Widaman, K. F. Missing Data: What to do With or Without Them. *Monogr. Soc. Res. Child. Dev.* **2006**, *71* (3), 42−64.
(7) Graham, J. W. Missing Data Analysis: Making It Work in the Real World. *Annu. Rev. Psychol.* **2009**, *60*, 549−576.
(8) Baraldi, A. N.; Enders, C. K. An Introduction to Modern Missing Data Analyses. *J. School. Psychol.* **2010**, *48* (1), 5−37.
(9) Little, R. J. A. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J. Am. Stat. Assoc.* **1988**, *83* (404), 1198−1202.
(10) Schroeder, J.; Murphy, K. L.; Holme, T. A. Investigating Factors that Influence Item Performance on ACS Exams. *J. Chem. Educ.* **2012**, *89* (3), 346−350.
(11) Holme, T.; Murphy, K. The ACS Exams Institute Undergraduate Chemistry Anchoring Concepts Content Map I: General Chemistry. *J. Chem. Educ.* **2012**, *89* (6), 721−723.
(12) Raker, J.; Holme, T.; Murphy, K. The ACS Exams Institute Undergraduate Chemistry Anchoring Concepts Content Map II: Organic Chemistry. *J. Chem. Educ.* **2013**, *90* (11), 1443−1445.
(13) Holme, T.; Luxford, C.; Murphy, K. Updating the General Chemistry Anchoring Concepts Content Map. *J. Chem. Educ.* **2015**, *92* (6), 1115−1116.
(14) Ford, B. L. An Overview of Hot-Deck Procedures. In *Incomplete Data in Sample Surveys*; Madow, W. G., Olkin, I., Rubin, D. B., Eds.; Panel on Incomplete Data; Academic Press: New York, NY, 1983; Vol. 2, pp 185−207.
(15) Andridge, R. R.; Little, R. J. A. A Review of Hot-deck Imputation for Survey Non-response. *Int. Stat. Rev.* **2010**, *78* (1), 40−64.
(16) Wilks, S. S. Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples. *Ann. Math. Stat.* **1932**, *3* (3), 163−195.
(17) Berglund, P.; Heeringa, S. G. *Multiple Imputation of Missing Data Using SAS*; SAS Institute: Cary, NC, 2014.
(18) R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org (accessed April, 2015).
(19) SAS Institute Inc. *SAS 9.4 Software*. http://www.sas.com/en_us/software/sas9.html (accessed April, 2015).
(20) Each percentile was calculated using the following formula: $P = [(\sum(N_l) + (1/2)N_p)/N_T] \times 100$, where P is the percentile for a specific total score, N_l is the frequency of students with a total score less than the specific total score, N_p is the frequency of students with the specific total score, and N_T is the total number of students. P was rounded to the nearest whole number.

(21) Ding, L.; Beichner, R. Approaches to Data Analysis of Multiple-Choice Questions. *Phys. Rev. ST—Phys. Educ. Res.* **2009**, *5* (2), 020103.
(22) Adams, W. K.; Wieman, C. E. Development and Validation of Instruments to Measure Learning of Expert-like Thinking. *Int. J. Sci. Educ.* **2011**, *33* (9), 1289−1312.
(23) Towns, M. H. Developing Learning Objectives and Assessment Plans at a Variety of Institutions: Examples and Case Studies. *J. Chem. Educ.* **2010**, *87* (1), 91−96.
(24) Emenike, M. E.; Schroeder, J.; Murphy, K.; Holme, T. Results from a National Needs Assessment Survey: A View of Assessment Efforts within Chemistry Departments. *J. Chem. Educ.* **2013**, *90* (5), 561−567.
(25) Wenzel, T. J.; McCoy, A. B.; Landis, C. R. An Overview of the Changes in the 2015 ACS Guidelines for Bachelor's Degree Programs. *J. Chem. Educ.* **2015**, *92* (6), 965−968.
(26) *Undergraduate Professional Education in Chemistry: ACS Guidelines and Evaluation Procedures for Bachelor's Degree Programs.* http://www.acs.org/content/dam/acsorg/about/governance/committees/training/2015-acs-guidelines-for-bachelors-degree-programs.pdf (accessed June, 2015).
(27) Holme, T. Assessment and Quality Control in Chemistry Education. *J. Chem. Educ.* **2003**, *80* (6), 594−596.

DOI: 10.1021/acs.jchemed.5b00180
J. Chem. Educ. XXXX, XXX, XXX−XXX