

**CHARACTERIZING MULTIPLE-CHOICE ASSESSMENT PRACTICES
IN UNDERGRADUATE GENERAL CHEMISTRY**

by

Jared B. Breakall

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Department of Chemistry

West Lafayette, Indiana

December 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Roy Tasker, Co-Chair

Department of Chemistry

Dr. Marcy Towns, Co-Chair

Department of Chemistry

Dr. Anne Traynor

Department of Educational Studies

Dr. Franziska Lang

Department of Chemistry

Approved by:

Dr. Christine Hrycyna

Head of the Graduate Program

This work is dedicated to my loving wife and children.

ACKNOWLEDGMENTS

Sir Isaac Newton once said of his accomplishments, “If I have seen further than others, it is by standing on the shoulders of giants.” Much like Sir Isaac Newton, I feel that anything I have accomplished with this work has only been possible because of the giants in my life who have let me stand on their metaphorical shoulders.

First and foremost, I would like to acknowledge my wife Sierra for her unending love, support, and encouragement. You have held our little family together while I have been working, buoyed me up when I have felt overwhelmed, and have always been a source of inspired advice and wisdom.

Additionally, I would like to acknowledge my amazing children, Emma and Ethan. You truly light up my world and bring the biggest smile to my face. Your unconditional love, trust, and Christlike examples have inspired me to do and be better every day.

Next, I’d like to acknowledge my Mom and Dad. What can you say to those who gave you everything and expected nothing in return? Thank you. Thank you for your support, encouragement, and belief in me no matter what dreams I have been pursuing. You taught me so much as I was growing up and have shaped me into who I am today.

Furthermore, I would like to acknowledge my grandparents and in-laws. Grandma and Grandpa Breakall and Lords, your love and support has been so valuable to me throughout this graduate school journey. You have supported me mentally and even financially. The monetary support you have given to my family and I has helped us through these years at Purdue University. Eric and Marylin, thank you for letting me marry into your family and for being amazing in-laws. You are some of the kindest and most supportive people I have ever met. You are such an inspiration to me, and I am so grateful to be a part of your family.

In addition to my family, I would like to acknowledge past and present members of the Tasker Research Group for their help and support. Specifically, I would like to acknowledge Roy Tasker, Chris Randles, Hannah Sturtevant, Alex Kararo, Ashton Hjerstedt, Elijah Roth, and Kevin Wee. Firstly, I would like to thank Roy for your encouragement to pursue my own research interests and for your supportive nature, which has helped shaped me into the researcher that I am. Secondly, I would like to acknowledge Chris Randles for your mentorship and advice throughout this journey. You have become a great friend to Sierra and I and hopefully a lifelong colleague. Thirdly, Hannah and Alex, you both helped me to see my first glimpse of what chemical education research is all about. Thank you for that. Fourth, Ashton, you are such a giving person who has given me great advice and help along the way. Fifth, Elijah, I can't think of anyone better to have started this journey with. Working by your side in this research group has helped me to develop as a person and as a researcher. Lastly, Kevin, working with you in this research group and in Chem 111/112 has helped me to develop as a leader. Although we are different, those differences have facilitated my growth as a leader and researcher. Thank you.

In summary, I am grateful to all the giants in my life who have allowed me to stand on their shoulders in order to reach new heights. Life, learning, and the progression of knowledge is best accomplished with the assistance of those around us. I have certainly received that assistance in this journey. Thank you to all of those who have helped me along this path of learning and progression.

TABLE OF CONTENTS

LIST OF TABLES	11
LIST OF FIGURES	12
ABSTRACT	14
CHAPTER 1. OVERARCHING INTRODUCTION	16
1.1 Purposes of this Study.....	18
1.2 Significance of this Study	18
1.3 Research Questions.....	19
1.4 Project Overview and Organization.....	20
CHAPTER 2. LITERATURE REVIEW	21
2.1 Basics of Educational Assessment.....	21
2.2 Test Development	22
2.3 Assessment Literacy	24
2.3.1 Studies of assessment literacy	27
2.3.2 Assessment literacy of higher education chemistry instructors.....	28
CHAPTER 3. METHODOLOGY.....	30
3.1 Guiding Research Question	30
3.2 Theoretical Framework.....	30
3.3 Methodological Framework.....	32
3.4 Participants and setting	33
3.5 Recruitment and sampling	34
3.6 Data collection	34
3.6.1 Demographic Data.....	35
3.6.2 Instructor interviews	35
3.7 Data Analysis	39
3.7.1 Demographic data.....	39
3.7.2 Transcription.....	39
3.7.3 Coding Process	39
3.8 Establishing Trustworthiness	40
3.8.1 Role of the Researcher.....	40

3.8.2	Potential Biases.....	41
3.8.3	Inter-Rater Reliability	41
3.8.4	Limitations.....	41
CHAPTER 4. RESULTS AND DISCUSSION		43
4.1	Views of Learning.....	43
4.2	Assessment Values and Principles.....	43
4.2.1	Obviously Implausible Distractors	44
4.2.1.1	In Support of Implausible Distractors	44
4.2.1.2	Opposed to Implausible Distractors	45
4.2.1.3	Discussion.....	47
4.2.2	Students should analyze data.....	47
4.2.2.1	Discussion.....	49
4.2.3	Number of concepts an item should test.....	50
4.2.3.1	Test many concepts per item	50
4.2.3.2	Test few concepts per item	51
4.2.3.3	Discussion.....	53
4.2.4	Algorithmic questions should be free response; Conceptual questions should be multiple-choice	53
4.2.4.1	Discussion.....	54
4.2.5	Levels of Understanding Tested	55
4.2.5.1	Discussion.....	56
4.2.6	Assess widely applicable skills.....	57
4.2.6.1	Discussion.....	58
4.2.7	Equitable items for all learners	59
4.2.7.1	Discussion.....	60
4.3	Knowledge of Assessment Purposes	61
4.3.1	Use MC assessment data to shape instructional decisions	61
4.3.2	Discussion.....	62
4.4	Knowledge of what to assess	62
4.4.1	What to assess.....	63
4.4.1.1	Discussion.....	70

4.4.2	Determining what to assess	72
4.4.2.1	Discussion.....	75
4.4.3	Finding the items	77
4.4.3.1	Discussion.....	80
4.5	Knowledge of Assessment Strategies	82
4.5.1	General Strategies	83
4.5.1.1	Collaboration	83
4.5.1.1.1	Discussion	85
4.5.1.2	Revise, Edit, and Trial	85
4.5.1.2.1	Discussion	86
4.5.1.3	Using multiple types of assessment items	87
4.5.1.3.1	Discussion	88
4.5.1.4	Benefits and Disadvantages of MC Assessment	88
4.5.1.4.1	Discussion	90
4.5.1.5	Lack of Knowledge of Multiple-Choice Assessment Design	90
4.5.1.5.1	Discussion	92
4.5.2	Item and Exam Properties.....	94
4.5.2.1	The stem	94
4.5.2.1.1	Negative phrasing.....	94
4.5.2.1.2	Complete problem statements	96
4.5.2.2	The response set.....	98
4.5.2.2.1	Distractors	99
4.5.2.2.2	All of the above	104
4.5.2.2.3	Answer choice length and order.....	106
4.5.2.3	The item overall.....	109
4.5.2.3.1	Item clarity	109
4.5.2.3.2	Item Conciseness (amount of information).....	111
4.5.2.3.3	Significant figures and units.....	112
4.5.2.3.4	Formatting consistency	114
4.5.2.4	The exam overall	116
4.5.2.4.1	Discussion	117

4.6	Knowledge of assessment interpretation and action taking	117
4.6.1	Discussion.....	121
CHAPTER 5. DEVELOPMENT AND USE OF A MULTIPLE-CHOICE ITEM WRITING FLAWS EVALUATION INSTRUMENT IN THE CONTEXT OF GENERAL CHEMISTRY		
	123
5.1	Abstract.....	123
5.2	Introduction.....	124
5.2.1	Multiple-choice item writing format and writing guidelines.....	126
5.2.2	Preparing the exam	127
5.2.3	The overall item.....	128
5.2.4	Stem creation guidelines.....	132
5.2.5	Answer choice creation guidelines	133
5.2.6	Guidelines for the exam overall.....	134
5.2.7	Item writing guideline violations in higher education.....	136
5.2.8	Purpose of the study.....	136
5.3	Methodology.....	137
5.3.1	Instrument Development	137
5.3.2	Item analysis of past exams	139
5.4	Results.....	140
5.4.1	Item writing guideline evaluation instrument.....	140
5.4.2	Item Analysis	143
5.5	Discussion.....	146
5.5.1	Development.....	146
5.5.2	Use of the Instrument.....	147
5.5.3	Limitations.....	151
5.6	Conclusions and Implications	152
CHAPTER 6. CONCLUSIONS AND IMPLICATIONS.....		154
6.1	General Conclusions	154
6.2	Recommendations for professional development.....	156
6.3	Recommendations for future research	159
REFERENCES		161

APPENDIX A.....	174
APPENDIX B.....	182
APPENDIX C.....	197

LIST OF TABLES

Table 3.1 Demographic Information.....	35
Table 5.1. Recent work on improving assessment practices in chemistry courses.....	124
Table 5.2. Item writing guideline violations in higher education	136
Table 5.3. Instrument development phase: Demographic information of inter-rater reliability raters.....	139
Table 5.4. Inter-rater Reliability of the IWFEI Criteria.....	143
Table 5.5. Item writing guideline adherence and violation (per item).....	145
Table 5.6. Item writing guideline adherence and violation (per exam)	145

LIST OF FIGURES

Figure 1.1. Example of a multiple-choice item with a stem, response set, and distractors.	17
Figure 1.2. Project Overview	20
Figure 2.1. Test development guidelines	23
Figure 2.2 Assessment triangle model	26
Figure 2.3. Science Teacher Assessment Literacy Model	27
Figure 3.1 Geographical area of participants institutions	35
Figure 3.2. Three-phase interview process	36
Figure 3.3. Phase two MC items.....	37
Figure 4.1. Fourth item evaluated during phase two of the interview protocol. Answer choice B is implausible.....	44
Figure 4.2 Item six evaluated during phase two of the interview protocol. This item requires the interpretation of graphical data	48
Figure 4.3. Dr. Sorenson’s stoichiometry item created during phase three of the interview.....	50
Figure 4.4. Dr. Sorenson's VSEPR item created during phase three of the interview.....	51
Figure 4.5. Item three evaluated during phase two of the interview.....	52
Figure 4.6. Item five evaluated during phase two of the interview	59
Figure 4.7. Item three evaluated during phase two of the interview.....	60
Figure 4.8. MC item created by Dr. Johnson about hydrogen bonding.....	66
Figure 4.9. Knowledge of Assessment Strategies.....	83
Figure 4.10. Created item during phase three of Dr. Johnson's interview	99
Figure 5.1. Example from (Towns, 2014) wordy item left; succinct item right. https://pubs.acs.org/doi/abs/10.1021/ed500076x Further permissions related to this figure should be directed to the ACS.	129
Figure 5.2. Grammatically Inconsistent Item left; Grammatically consistent item right (Examples from appendix 1, see supplemental information)	129
Figure 5.3. Example of k-type item format.....	131
Figure 5.4. Unfocused (left) focused (right) Reproduced from (Dudycha & Carpenter, 1973) with permission.....	132
Figure 5.5. Flowchart of instrument development and item analysis	138

Figure 5.6. Item Writing Flaws Evaluation Instrument	141
Figure 5.7. Frequency of item writing guideline violations per item	144
Figure 5.8. Overly wordy item identified using the IWFEI (left); Overly wordy item example from (Towns, 2014) (right) https://pubs.acs.org/doi/abs/10.1021/ed500076x Further permissions related to the right side of this figure should be directed to the ACS.....	148
Figure 5.9 K-type items identified using the IWFEI (left) and found in the literature (right)....	149
Figure 5.10. Items with implausible distractors identified using the IWFEI; Percentages of students choosing each answer choice is indicated.....	150
Figure 5.11. Flawed item identified by using the IWFEI and a revised version	151

ABSTRACT

Assessment of student learning is ubiquitous in higher education chemistry courses because it is the mechanism by which instructors can assign grades, alter teaching practice, and help their students to succeed. One type of assessment that is popular in general chemistry courses, yet difficult to create effectively, is the multiple-choice assessment. Despite its popularity, little is known about the extent that multiple-choice general chemistry exams adhere to accepted design practices or the processes that general chemistry instructors engage in while creating these assessments. Further understanding of multiple-choice assessment quality and the design practices of general chemistry instructors could inform efforts to improve the quality of multiple-choice assessment practice in the future. This work attempted to characterize multiple-choice assessment practices in undergraduate general chemistry classrooms by, 1) conducting a phenomenographic study of general chemistry instructor's assessment practices and 2) designing an instrument that can detect violations of item writing guidelines in multiple-choice chemistry exams.

The phenomenographic study of general chemistry instructors' assessment practices included 13 instructors from the United States who participated in a three-phase interview. They were asked to describe how they create multiple-choice assessments, to evaluate six multiple-choice exam items, and to create two multiple-choice exam items using a think-aloud protocol. It was found that the participating instructors considered many appropriate assessment design practices yet did not utilize, or were not familiar with, all the appropriate assessment design practices available to them.

Additionally, an instrument was developed that can be used to detect violations of item writing guidelines in multiple-choice exams. The instrument, known as the Item Writing Flaws Evaluation Instrument (IWFEI) was shown to be reliable between users of the instrument. Once

developed, the IWFEI was used to analyze 1,019 general chemistry exam items. This instrument provides a tool for researchers to use to study item writing guideline adherence, as well as, a tool for instructors to use to evaluate their own multiple-choice exams. The use of the IWFEI is hoped to improve multiple-choice item writing practice and quality.

The results of this work provide insight into the multiple-choice assessment design practices of general chemistry instructors and an instrument that can be used to evaluate multiple-choice exams for item writing guideline adherence. Conclusions, recommendations for professional development, and recommendations for future research are discussed.

CHAPTER 1. OVERARCHING INTRODUCTION

Educational assessment has been described as ‘reasoning from evidence’ and provides educators with a way to know what students know (Pellegrino, 2001). Simply put, assessment provides data to instructors that can then be used to make claims about student understanding. Those claims about student understanding can then be used to assign grades, alter teaching, and provide feedback. The importance of assessment in education cannot be understated and plays a critical role in educational decision making (Holme et al., 2010).

With this said, and with higher education chemistry instructors being largely responsible for the assessment of their students (Bretz, 2012), it is important for chemistry instructors to be able to create effective and meaningful assessments. The process of effective assessment design is important for all chemistry instructors to know and be able to use.

Knowing how to design, use, and interpret assessments properly is known as assessment literacy (DeLuca, LaPointe-McEwan, & Luhanga, 2016). Being assessment literate is important for anyone tasked with the design, use, or interpretation of educational assessments.

One popular form of assessment in chemistry courses (specifically general chemistry) is the multiple-choice (MC) assessment (Goubeaud, 2010). A MC assessment item typically consists of a stem, or problem statement, and a list of possible answer choices known as a response set. The response set usually contains one correct answer and incorrect options known as distractors. An example of a multiple-choice item is shown in Figure 1.1.

What is the molarity of a 0.50 L solution that contains 2.0 moles of NaCl?

- a) 1.0 M
- b) 2.0 M
- c) 3.0 M
- d) 4.0 M

Figure 1.1. Example of a multiple-choice item with a stem, response set, and distractors.

MC assessments are known for being difficult to construct because as with any assessment, creating a MC exam requires planning, objective-test alignment, revision, and a large time commitment (Fuhrman, 1996). Additionally, because MC items have a stem, response set, and distractors, there are numerous ways to lessen a MC exams validity based on how it is designed (Dell & Wantuch, 2017). When designing MC exams, it is important to follow appropriate design recommendations to avoid reducing the assessment's validity. Many of these recommendations are outlined in the literature as item writing guidelines (Haladyna, Downing, & Rodriguez, 2010).

Although MC assessments are popular in the chemistry classroom and are known for being difficult to construct, there is relatively little known about the current quality of MC chemistry assessments or what undergraduate chemistry instructors consider when they are designing these assessments for their students. No previous studies have explored what chemistry instructors consider during the MC assessment design process or to what extent MC chemistry exams currently adhere to item writing guidelines.

While there is not much known about MC assessment design practices in undergraduate chemistry classrooms, we do know that many chemistry instructors are not completely familiar with assessment terminology (Raker, Emenike, & Holme, 2013). It has also been noted that assessment of student learning can often feel overwhelming and foreign to chemistry instructors

(Bretz, 2012). Moreover, chemistry instructors tend to receive little to no formal training in appropriate assessment or pedagogical practices (Lawrie, Schultz, Bailey, & Dargaville, 2018). Furthermore, it has been suggested that chemistry instructors may benefit from assessment focused educational tools that could help them in their efforts to assess student learning effectively (M. E. Emenike, Schroeder, Murphy, & Holme, 2013).

1.1 Purposes of this Study

There were several purposes of this study which all focus on the broader idea of characterizing MC assessment practices in general chemistry classrooms. The purposes of this study were:

1. to investigate the assessment design practices of undergraduate general chemistry instructors by determining what they consider when they are creating or evaluating MC general chemistry exams
2. to develop an instrument (rubric) that can assist instructors and researchers to assess whether MC general chemistry exams are adhering to accepted guidelines
3. to determine to what extent a sample of general chemistry exams adhere to accepted item writing guidelines

1.2 Significance of this Study

With the increased demand of quality control in higher education (Holme, 2003), it is important for chemistry instructors to understand and create valid assessments (Bretz, 2012; Towns, 2014). A valid assessment is one in which evidence and theory support the interpretations of assessment's scores (AERA, 2014). Valid assessments provide a foundation upon which instructors can use the assessment data to make instructional decisions that can improve their teaching practice and student learning (Holme, 2011; Holme et al., 2010). Poorly constructed assessments can lead to invalid test results which can be unhelpful, or even detrimental, to educational decision making (Thorndike & Thorndike-Christ, 2010). If we are concerned with

student learning and improving chemical education practice, assessment needs to be a major focus (Holme, 2011). David Boud (1995) once said in the *Journal of Assessment for Learning in Higher Education*:

There is probably more bad practice and ignorance of significant issues in the area of assessment than in any other aspect of higher education. This would not be so bad if it were not for the fact that the effects of bad practice are far more potent than they are for any aspect of teaching. Students can, with difficulty, escape from the effects of poor teaching, they cannot (by definition if they want to graduate) escape the effects of poor assessment. Assessment acts as a mechanism to control students that is far more pervasive and insidious than most staff would be prepared to acknowledge. (p. 35)

Characterizing MC assessment practices in general chemistry classrooms will provide information about the current state of MC assessment and how it may be improved with targeted professional development opportunities. More specifically, understanding what general chemistry instructors consider when creating and evaluating MC exams may provide a building block for further professional development and research. Additionally, the development of a rubric that can be used to analyze general chemistry MC exams for their adherence to item writing guidelines, can fill a need for assessment focused educational tools that are accessible to undergraduate chemistry instructors.

1.3 Research Questions

The research questions guiding this study are:

1. To what extent do first semester general chemistry multiple-choice exams adhere to item writing guidelines?
2. What do general chemistry instructors consider when constructing/evaluating multiple-choice assessments?

1.4 Project Overview and Organization

This research has two main threads which are organized by research question. These threads are outlined below in Figure 1.2. The first thread of this research was a qualitative study of chemistry instructors MC assessment design practices. This thread will be presented throughout the document in Chapters 2, 3, 4, and 6. The second thread of the project was focused on the development of an instrument that can detect item writing guideline-violations in general chemistry exams. This thread of the research will be presented fully in Chapter 5.

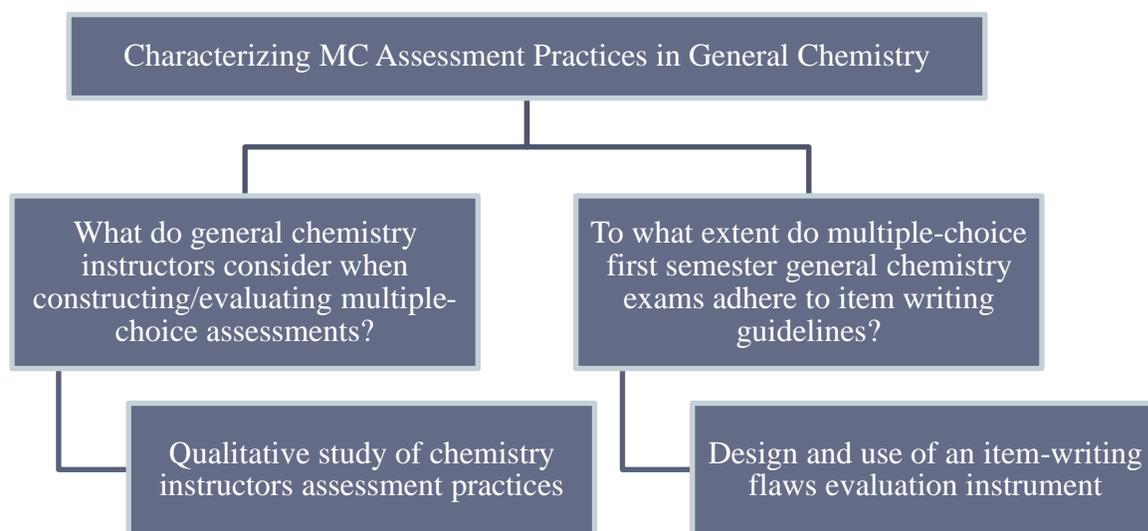


Figure 1.2. Project Overview

CHAPTER 2. LITERATURE REVIEW

2.1 Basics of Educational Assessment

Effective educational assessment is a complex task because it is impossible to measure a person's knowledge directly (Pellegrino, 2001). Instructors can only design tasks that elicit a response that would represent a person's knowledge of the subject at hand. For example, it is impossible to measure what a person understands about chemical kinetics, however tasks can be designed that would elicit what a person understands about chemical kinetics and then performance on those tasks can then be measured.

This is denoted mathematically in classical test theory by Equation 1 where, X represents a person's observed score on the test (the designed task), T represents the persons true score or understanding of the subject, and e represents error associated with the measurement process (Thorndike & Thorndike-Christ, 2010).

$$X = T + e$$

Equation 1.

The goal of effective educational assessment is to have an observed score as close to a person's true score as possible, or in other words, to minimize e , the error term. Therefore, it is advantageous that instructors ensure that tests and test items truly represent the content they want students to have mastered. This is known as test validity. Additionally, instructors want to make sure that items and exams will perform similarly with repeated administrations. This is known as test reliability.

In addition to validity and reliability, how difficult and discriminating test items are for students, are also important measures in educational assessment (Thorndike & Thorndike-Christ, 2010). Item difficulty, typically reported as a percentage, is the proportion of students who got the item correct over the total number of students who attempted the item. Item difficulty can give an instructor an idea of how hard the test item was for the students who attempted it. A higher percentage for item difficulty represents an easier item, whereas a lower item difficulty percentage represents a harder item. Item discrimination is a measure of how well an item can separate students based on ability. While test and item statistics can clarify how a test is functioning, they are not a replacement for qualitatively analyzing an exam for design flaws, trialing the exam, and thinking about the exam with students in mind (Thorndike & Thorndike-Christ, 2010).

2.2 Test Development

According to the Standards for Educational and Psychological Testing, test development is:

“the process of producing a measure of some aspect of an individual’s knowledge, skills, abilities, etc. by developing questions or tasks and combining them to form a test, according to a specified plan (AERA, 2014) (Pg. 75)”

Furthermore, the standards, which can be used as a guide for test development, describe the process as having four phases including: 1) the development of test specifications; 2) the creation, trialing, and evaluation of items; 3) assembly of the test; and 4) the development of procedures for administration and scoring (AERA, 2014). These general phases are outlined in Figure 2.1.

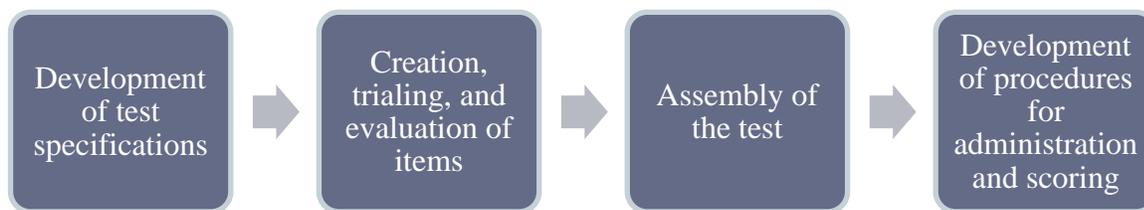


Figure 2.1. Test development guidelines

The first phase of this process is the development of test specifications. Test specifications are plans that can guide the test creation process (Fuhrman, 1996). These plans can include goal or purpose statements, content and coverage specifications, decisions about item format, level of cognitive skill specifications, and test length specifications (AERA, 2014). The use of test specifications is encouraged in the literature (Coderre, Woloschuk, & McLaughlin, 2009; Dills, 1998; Fives & DiDonato-Barnes, 2013) and has even been shown to improve student's attitudes toward chemistry courses (Young, Lashley, & Murray, 2019).

Once test specifications have been determined, items need to be developed, trialed, and evaluated before being included into an exam. The standards note that test development will be different in every situation, however they describe the ideal steps that should occur during this phase of test development (AERA, 2014). These steps include:

1. Assemble an item pool that contains more items than are needed
2. Evaluate items for content quality, clarity, and construct irrelevant variance (CIV)
3. Administer items to a representative group of test takers
4. Conduct a statistical analysis of the items, including differential item functioning (DIF)
 - If DIF is present, conduct interviews to identify problematic item features
5. Revise items

Literature on appropriate item writing practices, item clarity, and CIV have been reviewed and can be found in section 5.2. Generally, this process of assembling an item pool, evaluating

items for quality, and revising items before administration has been described and promoted in the higher education literature (Dell & Wantuch, 2017; Regan, 2015; Towns, 2014).

After the item pool has been evaluated, revised, and tested, an instructor can then populate the exam with items that meet the requirements set by the test specification. Aligning exam items with the requirements set out by the test specification helps to increase exam validity by ensuring appropriate content, difficulty, and thinking skill distribution throughout the exam (Fives & DiDonato-Barnes, 2013). Once the test form has been assembled, instructors may want to field test the exam prior to actual use. This would be done mainly to evaluate how the items function as a whole and to check for exam reliability prior to official use (AERA, 2014).

The last stage of test development is to create procedures for administration and scoring. Regarding administration, it is important to develop clear, and standardized instructions for all examinees. However, test accommodations should be made for those students who need them. Regarding scoring, assigning points and weights to each item and developing grading rubrics for open-ended items is important for the reliable interpretation of exam scores. Lastly, test security measures should be implemented, including, storage of test materials and scores, non-disclosure agreements for examinees, and exam proctoring procedures (AERA, 2014).

These four stages of test development as outlined by the AERA's Standards for Educational and Psychological Testing can serve as a guide for what practices should be included in test development and design.

2.3 Assessment Literacy

While designing curriculum, creating a motivating learning environment, and implementing effective pedagogical practices are all part of being a successful instructor, it is an instructor's ability to assess student learning that is the building block for improving education.

This is because without determining what students have learned, it is difficult to know what to change or improve upon. An instructor's knowledge of and ability to implement effective assessment is typically referred to as his or her assessment literacy (AL) (Stiggins, 1991). Stiggins describes assessment literate instructors as those who have a basic understanding of high- and low-quality assessment. Furthermore, instructors who are assessment literate know how to create assessments that provide useful data about student learning and can interpret the results in meaningful ways (Stiggins, 1991). Additionally, three essential areas of assessment literacy that instructors need to have an understanding of were outlined including, basic concepts and terminology of assessment, the uses of assessment, and assessment planning and development (Schafer, 1991).

As time progressed and more understanding of assessment literacy emerged, others added to the working model of what it means to be assessment literate.

Incorporated into Magnusson's model of pedagogical content knowledge (PCK), is the idea that instructors need to understand how to assess scientific literacy (S Magnusson, Krajcik, & Borko, 1999). In this model, Magnusson splits assessment of scientific literacy into two subdomains of, "dimensions of science learning to assess" and "methods of assessing science learning" (S Magnusson et al., 1999). Although not overly specific, these subdomains do indicate the importance of knowing what to assess and how to assess it.

Additionally, in a report from the National Academy of the Sciences, assessment is described as "reasoning from evidence" and a model, known as the assessment triangle, outlines what instructors need to consider when creating assessments (Pellegrino, 2001). The three parts of this model, as shown in Figure 2.2, are cognition, observation, and interpretation. The cognition corner of the triangle refers to the fact that every assessment is built upon a model of learning that

can help identify appropriate assessment tasks. The observation corner of the model refers to the set of beliefs about what kind of tasks will elicit responses from students that represent desired

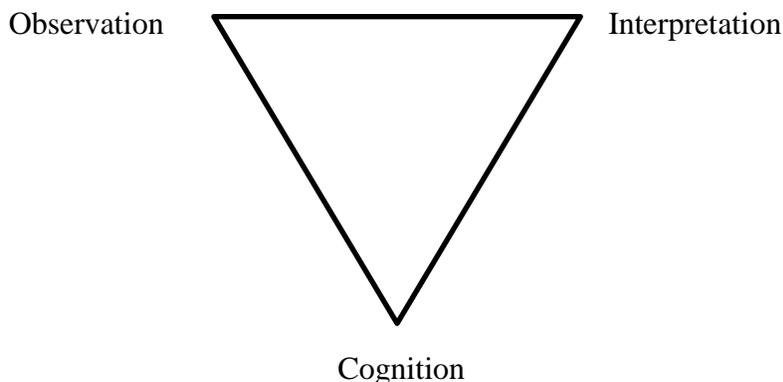


Figure 2.2 Assessment triangle model

knowledge or skills. The interpretation corner of the model represents the ideas, tools, and methods used to analyze the data received from an assessment. The three corners of this model all must be considered by one who is designing effective assessment tasks (Pellegrino, 2001). This model has added more depth and structure to what it means to be assessment literate.

Moreover, the idea of AL was further advanced through the creation of the science teacher assessment literacy model (STALM) (Abell & Siegel, 2011). This model used the foundations of PCK and the assessment triangle to develop a more detailed method to look at the assessment literacy of science teachers. As shown in Figure 2.3, the STALM is centered around “views of learning” and “assessment values and principles.” Radiating outward, there are four domains in the model: Knowledge of Assessment Purposes, Knowledge of What to Assess, Knowledge of Assessment Strategies, and Knowledge of Assessment Interpretation and Action-Taking. This model has been used to assess the assessment literacy of pre-service teachers and college instructors (Presley & Hanuscin, 2015; Siegel & Wissehr, 2011).

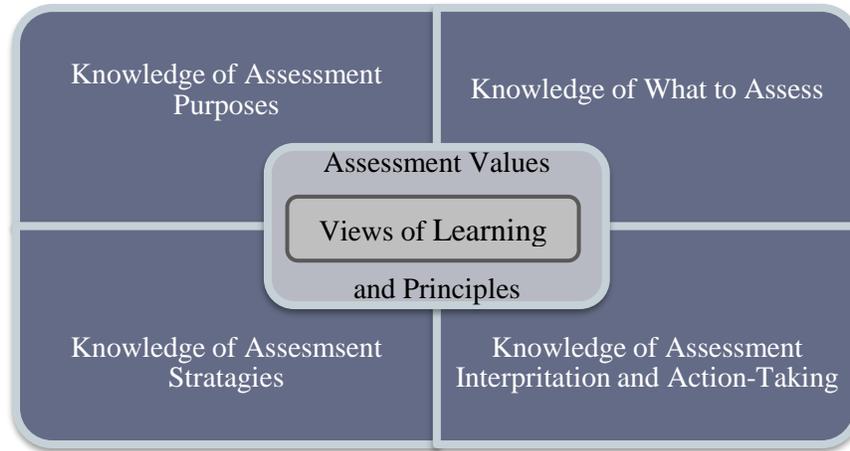


Figure 2.3. Science Teacher Assessment Literacy Model

Most recently, in a paper that reviewed over 100 articles on AL and connected the fields of educational assessment and teacher education, a more detailed model of AL was developed called the Teacher Assessment Literacy in Practice (TALiP) model (Xu & Brown, 2016). This model situates AL in the context of social, cognitive, affective, and identity dimensions. This model provides an intricate model of what AL may look like in practice.

2.3.1 Studies of assessment literacy

Many of the studies on assessment literacy have been conducted with pre-service or in-service K-12 teachers. From these studies, a few general conclusions emerge. First, pre-service and in-service teachers lack proficiency in using a variety of assessment tasks (Mertler, 1999; Siegel & Wissehr, 2011). Secondly, assessment literacy tends to develop over time with experience (Duffee & Aikenhead, 1992; Sato, Wei, & Darling-Hammond, 2008; Zhicheng & Burry-Stock, 2003). Lastly, in a review of over 100 studies on AL there were three main conclusions for instructors and assessment education efforts (Xu & Brown, 2016). First, instructors need to have a firm understanding of assessment if they are to be successful in assessing their students. Secondly, assessment training needs to be long term to allow instructors to engage in deep learning about

assessment. Third, it is needed to better understand instructors as individuals because improving assessment literacy is a complex task that involves prior experiences, relationships, and emotions. These findings, although from studies with K-12 instructors, do provide insight into the field of AL.

2.3.2 Assessment literacy of higher education chemistry instructors

Although few studies have examined the assessment literacy of chemistry instructors at the undergraduate level directly, there are some studies, when looked at together, that provide understanding of the state of assessment practices in the undergraduate chemistry classroom. First, it is important to note that chemistry educators have focused on their assessment efforts since the 1920's and continue to make assessment of student learning a priority (Bretz, 2013). Secondly, although improvements in assessment practice have been made, chemistry instructors at the undergraduate level often feels unfamiliar with, and overwhelmed by, the responsibility to assess their student's learning effectively (Bretz, 2012). This may, in part, be due to the fact that they tend to receive little to no formal training in effective teaching or assessment practices (Lawrie et al., 2018). Thirdly, chemistry instructors are actively trying to improve their assessment practices. In a survey of over 1,500 chemistry instructors, 72% were aware of plans for enhanced assessment efforts in their departments (M. E. Emenike et al., 2013). Additionally, it was noted that 38% of departments used in-house chemistry exams (M. E. Emenike et al., 2013). In another study it was found that 93.2% of a sample 1,282 chemistry instructors from institutions that confer bachelor's degrees in chemistry, reported creating their own assessments (Gibbons et al., 2018). Furthermore, in a report by Towns (2010), the assessment improvement plans of four institutions were described as evidence of efforts to improve assessment practice. These findings indicate that chemistry departments are engaged in improving assessment efforts and that many create their own exams

in the process. Lastly, even though many are engaged in assessment improvement efforts, chemistry instructors have been shown to be unfamiliar with assessment terms such as validity, reliability, item difficulty, item discrimination, Cronbach alpha, and item response theory (Raker, Emenike, et al., 2013; Raker & Holme, 2014). These self-reported results were validated using an internal standard as well (M. Emenike, Raker, & Holme, 2013). In summary, chemistry instructors at the undergraduate level, although not completely familiar with, or trained in assessment methods, are actively trying to improve their assessment practices.

CHAPTER 3. METHODOLOGY

To gain a more detailed foundational knowledge of general chemistry instructors multiple-choice assessment literacy, a qualitative, phenomenographic study was conducted. The guiding research question and related sub-research questions (3.1), frameworks (3.2 and 3.3), participants (3.4), recruitment and sampling plan (3.5), data collection (3.6), data analysis (3.7), and reliability studies (3.8) are outlined within the following sections.

3.1 Guiding Research Question

The research question which guided the development and execution of this project was:

- What do general chemistry instructors consider when constructing/evaluating multiple-choice assessments?

Several related research sub-questions were also identified based on the methodological framework used in this study. These research questions refer to different aspects of the methodological framework used to analyze the data from this project. These related research questions are:

- What values and principles are guiding chemistry instructors in their assessment decisions?
- What assessment strategies do general chemistry instructors use or consider when assessing their students?
- What do general chemistry instructors find important to assess?
- What influences how general chemistry instructors determine what to assess?
- How do general chemistry instructors generate MC items for their exams?
- How do general chemistry instructors interpret and use exam data?

3.2 Theoretical Framework

Phenomenography is a research framework that is used to explore the similarities and differences between peoples' experiences with a phenomenon (Marton, 1981). The purpose of

phenomenography is to describe the ways in which people experience, perceive, or conceptualize a phenomenon (G. Bodner & Orgill, 2007). Furthermore, this framework characterizes the variation of experiences within a group (Trigwell, 2000).

Phenomenographic studies have useful benefits and implications in education. They can provide data on how both teachers and students experience the learning and teaching process (Svensson, 1997). Additionally, the data gained from phenomenographic studies can be used to inform change from one way of thinking/experiencing a phenomenon to a “better” way (Marton, 1986). Using phenomenography to collect data on how people experience an event (i.e. constructing multiple-choice exams), could lead to an improvement in how people experience that phenomenon in the future.

Phenomenography is a framework that has been shown to be useful in educational research and has been used in numerous science/chemistry education research studies (Bhattacharyya & Bodner, 2005; Bretz, 2010; Dekorver & Towns, 2015; Orgill & Sutherland, 2008; Tomanek, Talanquer, & Novodvorsky, 2008). Some of the relevant research questions associated with these studies are: What are the different ways graduate students propose organic mechanisms (Bhattacharyya & Bodner, 2005)? What do undergraduate general chemistry, analytical chemistry, and biochemistry students’ (‘chemistry students’) understand about chemical buffers (Orgill & Sutherland, 2008)? What do students enrolled in a general chemistry course hope to accomplish in the laboratory across the cognitive, affective, and psychomotor domains of learning (Dekorver & Towns, 2015)? What factors influence science teachers’ reasoning when evaluating and selecting formative assessment tasks (Tomanek et al., 2008)?

3.3 Methodological Framework

The methodological framework that was used for this study was the Science Teacher Assessment Literacy Model (STALM). The purpose of this model is to capture the types of assessment knowledge and skills that are needed to create an assessment-centered learning environment (Abell & Siegel, 2011). This model, (Figure 2.3) incorporates six domains including: the instructor's views of learning, assessment values and principles, knowledge of assessment purposes, knowledge of what to assess, knowledge of assessment strategies, and knowledge of assessment interpretation and action-taking (Abell & Siegel, 2011).

The instructor's views of learning and assessment values and purposes are in the center of the model because they influence all other parts of the model (Figure 2.3). What a teacher believes about student learning and assessment influences not only how he or she teaches, but how he or she assesses his or her students.

Knowledge of Assessment Purposes refers to why a teacher chooses to assess students. Some of the reasons could include: understanding what students know before instruction, providing evidence of student learning, and providing data that can be used to change instructional practice.

Knowledge of What to Assess refers to a science teacher knowing what to test in his or her classroom (Abell & Siegel, 2011). This is closely related to the goals and learning objectives a teacher has for his or her classroom.

Knowledge of Assessment Strategies refers to what a science teacher knows about different types assessment (e.g. multiple-choice, short answer, clickers, etc.) (Abell & Siegel, 2011). Additionally, this includes what a teacher knows about how to create assessments that produce usable, reliable, and valid data (Abell & Siegel, 2011).

Knowledge of Assessment Interpretation and Action-Taking refers to what science teachers know about what to do with assessment data (Abell & Siegel, 2011). Knowing how to interpret and use assessment data is an important part of assessment literacy.

The STALM is founded in the frameworks of the assessment triangle (Pellegrino, 2001) and Pedagogical Content Knowledge (PCK) (Shirley Magnusson, Krajcik, & Borko, 1999) as described previously. The STALM was created to describe science teachers assessment literacy in more detail than can be done with Magnusson's model of PCK (Abell & Siegel, 2011).

The STALM has been used to measure assessment literacy in middle school science teachers (Gottheiner & Siegel, 2012), science instructors at the community college level (Presley & Hanuscin, 2015), and in experienced biology professors and supplemental instruction leaders (Vanmali & Siegel, 2012).

This was an appropriate framework for this study because this study explored what general chemistry instructors consider when they are creating/evaluating multiple choice assessment. The STALM allowed for a detailed analysis of the interview data and provided appropriate categories to be used in the data analysis.

3.4 Participants and setting

Instructors of general chemistry in the United States who have used MC items or exams in their courses were solicited for participation in this study. An instructor of general chemistry was defined for this project as a person who has taught undergraduate general chemistry, as a lead instructor, within three years of being interviewed.

3.5 Recruitment and sampling

Participants were recruited through a combination of purposeful sampling techniques (Patton, 2015). To find an adequate number of participants that would provide useful data about the creation of MC exams in undergraduate chemistry courses, snowball and saturation sampling were used.

In snowball sampling, participants are asked to refer others who would fit the requirements of the study and may be willing to participate (Patton, 2015). Those who were referred were then contacted. Initial participants were identified through their university websites and then contacted about the study.

To determine how many participants would need to be interviewed, a saturation sampling technique was employed. In saturation sampling, recruitment is ended when no new information is obtained from newly interviewed participants (Lincon & Guba, 1985). This type of sampling requires that data collection and analysis occur at the same time, and that those interviewed come from a wide enough background within the context of the study so as to not reach premature saturation (Patton, 2015).

Recruitment occurred mainly through email with some recruitment occurring in person. No incentives were given to participants in this study.

3.6 Data collection

The data collected to answer the research question was collected from June 2017 – August 2019. The data collected included demographic information and interviews with instructors about their MC assessment design practices.

3.6.1 Demographic Data

Demographic information on each participant was collected through a Qualtrics survey. Information collected included: chemistry subdiscipline, job title, teaching experience, and MC exam creation experience. Demographic information of the participants is found in Table 3.1

Table 3.1 Demographic Information

<i>Pseudonym</i>	<i>Chemistry discipline most associated with:</i>	<i>Job title:</i>	<i>Experience teaching (yrs)</i>	<i>Approx. number of MC gen. chem. exams administered:</i>
<i>Dr. Madison</i>	Physical	Associate Professor	10+	20+
<i>Dr. Sanders</i>	Organic	Lecturer	10+	20+
<i>Dr. Deagen</i>	Biochemistry	Full Professor	10+	20+
<i>Dr. Brown</i>	Chem Ed.	Full Professor	10+	20+
<i>Dr. Irvine</i>	Physical	Associate Professor	10+	20+
<i>Dr. Tracy</i>	Inorganic	Associate Professor	10+	20+
<i>Dr. Lopez</i>	Physical	Associate Professor	5-7	20+
<i>Dr. Sorenson</i>	Inorganic	Assistant Professor	5-7	10-19
<i>Dr. Crawford</i>	Chem Ed.	Assistant Professor	5-7	5-9
<i>Dr. Smith</i>	Physical	Lecturer	2-4	5-9
<i>Dr. Bennett</i>	Biochemistry	Assistant Professor	2-4	5-9
<i>Dr. Patrick</i>	Chem. Ed.	Other	0-1	1-4
<i>Dr. Johnson</i>	Inorganic	Other	0-1	5-9

3.6.2 Instructor interviews

Due to the wide geographical area where instructors were located (see Figure 3.1), interviews occurred both in person and online through video chat software. Interviews followed a semi-structured interview protocol that consisted of three phases as shown in Figure 3.2.

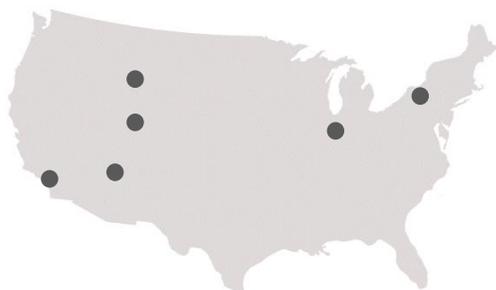


Figure 3.1 Geographical area of participants institutions

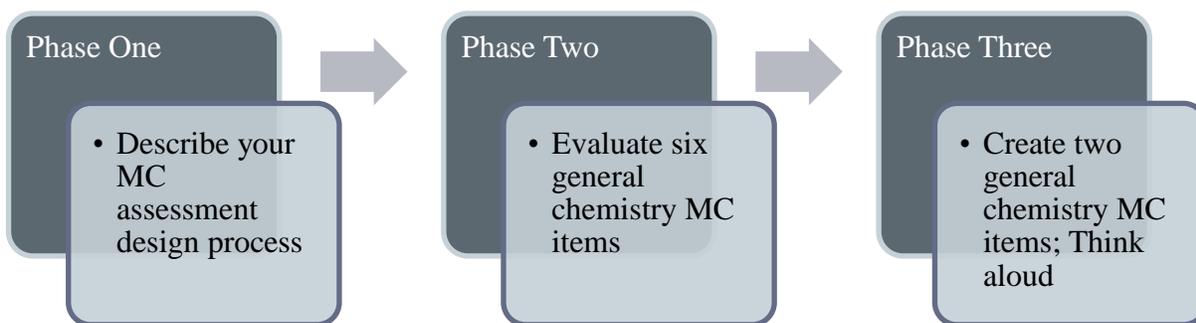


Figure 3.2. Three-phase interview process

In the first phase, instructors were asked questions about their teaching background and the process they go through when designing MC general chemistry exams. Background questions were asked, in part, to allow participants to get used to the interview procedure by answering non-threatening questions. Although the interviews were semi-structured in nature, there were seven typical questions that were asked during this phase of the interviews as shown below:

1. How long have you been teaching at the college level?
2. How many semesters have you taught general chemistry?
3. How many multiple-choice exams have you been involved in creating for general chemistry?
4. How long before the exam date do you usually start to create it?
5. Do you typically make your own MC exam questions, or do you get them from other sources?
6. Do you normally make general chemistry exams by yourself or with colleagues?
7. In a broad overview, would you be able to walk me through how you typically go about creating a multiple-choice exam for your general chemistry students?

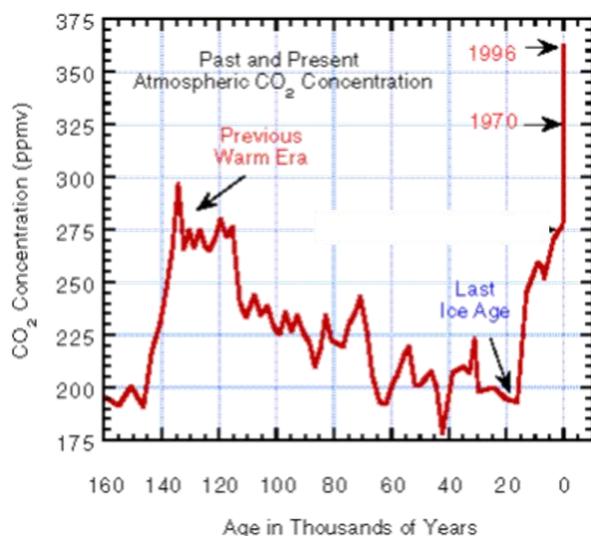
In the second phase, instructors were asked to evaluate six MC general chemistry exam items as if they were considering putting them onto an exam for their students. They were asked what they liked/disliked about each of the items and if there was anything about the items they would change. The six items were chosen because they exemplified common item writing guideline violations. These items are shown below in Figure 3.3.

<p>1. What is the molarity of a 35.0 mL solution of 9.00 M H₂SO₄ diluted to 0.500 L?</p> <p>(a) 6.30 M (b) 0.624 M (c) 61.1 M (d) 630. M (e) 0.630 M</p>
<p>2. Which quantum number effectively describes the shape of the probability region that an electron in an atom can occupy?</p> <p>(a) m_s (b) n (c) m_ℓ (d) ℓ (e) All of the above</p>
<p>3. Which of the following are key differences between chemical and nuclear reactions?</p> <p>I. Atoms do not change identity in chemical reactions, whereas in nuclear reactions they do II. Nuclear reactions release a greater amount of energy than chemical reactions III. Nuclear reactions have rates that depend on temperature, concentration, and catalysts, whereas chemical reactions do not</p> <p>(a) I (b) I, II (c) II, III (d) I, III (e) I,II,III</p>
<p>4. Electronegativity:</p> <p>(a) has no periodic trends. (b) is the term for a common attitude among pessimistic electrons. (c) is generally greatest for the transition metals. (d) can be used to determine chemical properties and generally increases left to right across a period and decreases down a group. (e) generally decreases left to right across a period and increases down a group.</p>
<p>5. What volume of 1 molar hydrochloric acid would be neutralized by 10 grams of chalk?</p> <p>(a) 50 mL (b) 100 mL (c) 150 mL (d) 200 mL (e) 250 mL</p>

Figure 3.3. Phase two MC items

Figure 3.3. Continued

6. Answer the following question based on the graph provided.



Which of the following cannot be determined based on the provided graph?

- (a) 42,000 years ago the CO₂ concentration was at an all-time low
- (b) CO₂ levels increase during warm periods
- (c) There was a sharp increase in the CO₂ concentration after the last ice age
- (d) There were no warm eras prior to 160,000 years ago
- (e) 135,000 years ago the CO₂ level was about 33% higher than it was 35,000 years ago

In phase three of the interview, instructors were asked to create two MC exam items for a first semester general chemistry course using a think-aloud protocol. They were asked to create one item that they would consider to be conceptual and one that they would consider to be algorithmic. This was done to elicit different considerations when the instructors were creating different types of MC exam items. The instructors were given four possible general chemistry topics to choose from when deciding to create their test item. These four topics included, VSEPR theory and molecular and electron geometries, stoichiometry, acid-base titrations, and intermolecular forces. To capture the written data from this phase of the interview, a Livescribe pen was used during in-person interviews and scanned work was used for online interview participants.

3.7 Data Analysis

3.7.1 Demographic data

The interview data collected in this project were not analyzed by demographic data such as teaching experience. The demographic data were only used to provide context of the participant's background and experiences.

3.7.2 Transcription

Once interview data were collected, they were transcribed removing filler words such as, "like" or "um." Multiple-choice items from phase two and created MC items from phase three were then added to the appropriate spot in each transcript.

3.7.3 Coding Process

Interview transcripts were coded in both an inductive and deductive manner. Codes were created in an inductive manner from what emerged from the analysis of interview transcripts (Saldana, 2016). The codes were then put into appropriate STALM categories in a deductive manner for organizational purposes.

To create an initial coding scheme, the first five interviews collected were analyzed. Coding occurred in an iterative manner where the transcripts were analyzed for one portion of the STALM at a time. For example, all five initial transcripts were analyzed and coded for data that would be deemed fit into "Assessment Values and Principles." Once coded, those five transcripts were then analyzed again for data that would be deemed to fit into another portion of the STALM, for instance, "Knowledge of Assessment Strategies." This process was repeated until all transcripts were analyzed for all STALM categories. This coding method was used to increase the detail and consistency of the coding process. In addition, a definition for each code was created to help guide the interpretation and use of the coding scheme (Appendix A).

After the initial coding scheme was created, subsequent interviews were transcribed and then analyzed immediately looking for new codes to emerge from the data. Any new codes were defined and added to the coding scheme. Once new codes stopped appearing in the data, further recruitment ended as data saturation was reached. The coding scheme can be found in Appendix A.

3.8 Establishing Trustworthiness

Because the researcher is the instrument in qualitative research, it is important to establish the reliability or trustworthiness of the data analysis methods (Patton, 2015). To establish trustworthiness, the role of the researcher, potential biases, inter-rater reliability studies, and limitations of this study will now be outlined.

3.8.1 Role of the Researcher

As the researcher, my role has affected all aspects of this work. I have designed the study, collected the data, and completed the data analysis. In the design phase, I identified a gap in the literature, determined research questions, conducted pilot interviews, and finalized the research methodology with the help my advisors. In the data collection phase, I conducted the semi-structured interviews and gathered the demographic information from the participants. During the data analysis phase, I inductively coded the transcripts using the STALM as a guide.

Additionally, I have experience that qualifies me for this work. As part of my graduate education I have taken courses in qualitative research design and educational measurement. These courses have given me a foundation to make me successful as the researcher in this qualitative project focused on faculty assessment practices. In addition to the courses I have taken, I have an undergraduate degree in chemistry education, and have taught general chemistry as a graduate

teaching assistant for 4.5 years. This has helped me to become familiar me with the content and student abilities in general chemistry courses and how these courses are typically assessed.

3.8.2 Potential Biases

To further establish trustworthiness, it is important to identify the potential biases I as the researcher hold that may affect this work. I have hypothesized throughout this project that chemistry instructors may not be aware of successful MC assessment design practices. I may be prone to confirmation bias of this hypothesis although I have taken efforts to remediate this bias through memoing and coding reliability studies with another researcher.

3.8.3 Inter-Rater Reliability

To assess the inter-rater reliability of the coding scheme, an additional rater used the codebook definitions to code a randomly selected transcript using the entire codebook. The additional rater was given a 30-minute orientation on using the codebook. The additional rater analyzed the transcript independently using N’vivo software. Kappa was then calculated between the raters as a measure of inter-rater reliability. A 0.74 Kappa was calculated which suggests a substantial level agreement between the raters when chance agreements are taken into account (Landis & Koch, 1977; Viera & Garrett, 2005).

3.8.4 Limitations

Much like any educational assessment, the research methods in this study can not fully capture the knowledge that an instructor has about assessment design, and thus, the interview tasks serve as indirect measures to approximate multiple-choice AL.

Furthermore, every research study has limitations that need to be considered when interpreting the results. Several limitations exist in this study that will now be outlined. First, because

participants had to choose to participate in the study, the participants may be more comfortable with and knowledgeable about assessment design than chemistry instructors at large. This was seen during recruitment when several potential participants declined to participate because they did not feel knowledgeable enough about assessment design to be interviewed. Thus, results may be skewed toward higher assessment literacy. Secondly, because the interviews were conducted in a research setting with an interviewer present, observer effects may have influenced participant considerations in comparison to if the participants were unobserved. Although this limitation is an inherent part of interview-based studies, it is important to consider this when interpreting results.

CHAPTER 4. RESULTS AND DISCUSSION

The results of this study are organized by the categories of the Science Teacher Assessment Literacy Model (STALM; Figure 2.3). Each category within this framework describes a slightly different aspect of assessment literacy (AL) and thus provides a different view on what general chemistry instructors are considering when they create or evaluate multiple-choice (MC) assessments. The results will now be presented and discussed by category of the STALM, which include: views of learning (Section 4.1), assessment values and principles (Section 4.2), knowledge of assessment purposes (Section 4.3), knowledge of what to assess, (Section 4.4), knowledge of assessment strategies (Section 4.5), and knowledge of assessment interpretation and action taking (Section 4.6).

4.1 Views of Learning

The instructors interviewed did not discuss their views of learning during this study. This may be due to the nature of the interview protocol which did not directly probe into the chemistry instructors' views of learning. However, it is interesting to note that the instructors did not discuss their views of how students learn in regard to how they design assessments. Further research into how chemistry instructors' views of student learning influence assessment design is recommended.

4.2 Assessment Values and Principles

What values and principles are guiding chemistry instructors in their assessment decisions?

Abell and Siegel proposed that assessment values and principles are overarching ideas and beliefs that guide assessment decisions (Abell & Siegel, 2011). As the transcripts were inductively coded, the assessment-focused overarching ideas and beliefs of the participants emerged from the

data. A total of 17 values and principles were identified. Substantial findings will now be reported and discussed.

4.2.1 Obviously Implausible Distractors

Many of the participants interviewed made statements that reflected a belief about including obviously incorrect answer choices in MC items. Instructors responses fell into two categories, namely, those who were in support and those who were opposed to the practice.

4.2.1.1 In Support of Implausible Distractors

Some of the participants held a belief that including obviously implausible distractors is a desirable practice in MC assessments. Those who were in support of including obviously implausible distractors tended to cite reducing test anxiety as a main reason for the practice. Many of these statements occurred during the second phase of the interview while participants were evaluating the fourth item (Shown below in Figure 4.1).

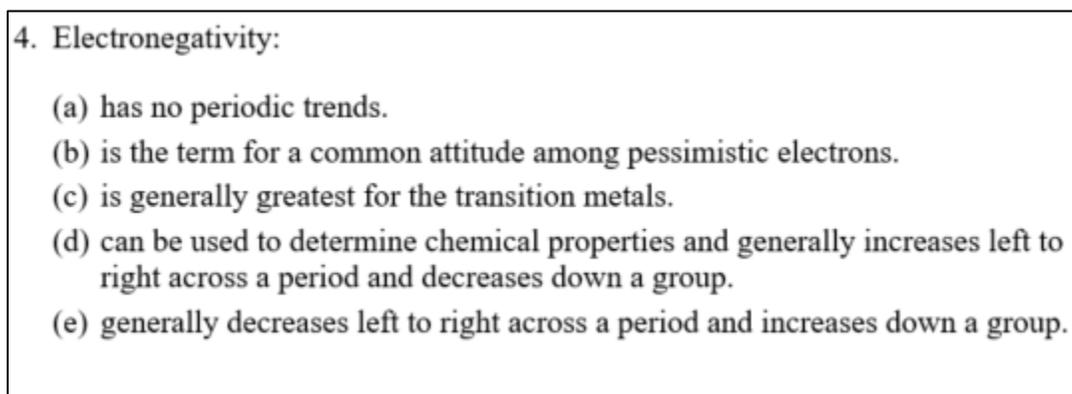


Figure 4.1. Fourth item evaluated during phase two of the interview protocol. Answer choice B is implausible.

Dr. Madison who had over 10 years of teaching experience in general chemistry and had administered over 20 MC exams said about item four:

And so, I actually like answer B in question number four only because that's kind of my sense of humor anyway, and I do try to inject it sometimes into an exam just to kind of try to lighten the mood sometimes, because I mean, students ... some of them get very, very intense while they're taking an exam and you have to remind them to breathe, because some of them, you look at them, you think they're holding their breath for some inordinate amount of time, and you're like, 'Come on. Just breathe, please.'

In another example, Dr. Lopez who had 5-7 years of teaching experience and had administered over 20 MC exams said about item four:

Good features, it's kind of fun. I like option B, it just brings a smile to the students and maybe relax them while they're taking their test.

Furthermore, While Dr Bennett, who had 2-4 years of teaching experience in general chemistry and had administered 5-9 MC exams, was creating an item during phase three of the interview said:

I usually try to have an obvious wrong... If it can be right so that they can eliminate, so that essentially they are picking between four and it makes them feel a little better.

These data show that some of the participants hold a belief that including obviously implausible distractors can act as a stress reliver for students as they take a potentially stressful exam. Dr. Madison, Dr. Bennett, and Dr. Lopez described answer choice B as something that is desirable and could help their students to relax during an exam.

4.2.1.2 Opposed to Implausible Distractors

Contrarily, some of the participants held a belief that including obviously implausible distractors is not a desirable practice in MC assessments. Many of these participants cited the lack of students choosing the implausible answer choice as a reason to avoid the practice. These comments also tended to emerge as participants evaluated the fourth item in phase two of the interview protocol (Figure 2).

Dr. Crawford who had 5-7 years of general chemistry teaching experience and had administered 5-9 MC exams said about item four:

And then obviously there's stupid answers. Like a common attitude among pessimistic electrons or something, which is just a throw away answer. So again why does somebody put that there? Maybe it's funny and people will think that's a funny little answer, but actually it's just a throw away. You're asking about how do I make an assessment that measures student learning, this is not a good way to do that because you're just throwing away a distractor that's not useful.

Dr. Sanders who had more than 10 years of general chemistry teaching experience and had administered over 20 MC exams said about item four:

Response B, common attitude among pessimistic electrons. I think that's nonsense and that's just a filler and it needs to be taken out. No student in their right mind is going to select that response and so why put it?

Dr. Smith who had 2-4 years of general chemistry teaching experience and had administered 5-9 MC exams said about item four:

Dr. Smith: Question number four I'm not crazy about. While answer B is super cute, I don't like to have total throwaways in multiple choice questions.

Interviewer: Okay. Can you tell me why?

Dr. Smith: Yeah. I want all of the answers, someone should be able to select any of the answers and there's a reason why they selected it.

Interviewer: Okay.

Dr. Smith: No student is gonna select B and if they do, they're trying to be funny. I'm not learning anything about their understanding of electronegativity. So, B's a waste of space for me. While I teach, I'm quite silly and very quirky. But when I am actually giving them exams, I'm pretty serious.

Interviewer: Yeah.

Dr. Smith: Usually like, I know some people will put funny extra credit things. I never do. And so for me, while B is very funny, I would never bother with it on an exam.

These examples show that some of the participants hold a belief that including obviously implausible distractors are not useful in MC exams. Although some participants, like Dr. Crawford and Dr. Smith understand that implausible distractors may be viewed as funny, they believe that

including them does not help assess what students know and therefore, should be left out of a MC item.

4.2.1.3 Discussion

When it comes to including obviously implausible distractors in MC items, there are two ways to view the issue, from a psychometric perspective or from a humor perspective. From a psychometric perspective, it is recommended to include only as many distractors as are plausible without including additional choices (Rodriguez, 2005). This is because including extra answer choices will increase test taking time, decrease the number of items that can be included in an exam, and will not provide you with more information about student ability (Rodriguez, 2005). However, although empirical results vary considerably in regards to the effect of using humorous response options, using humor is supported only if the humor is consistent with the instructor, the testing is low stakes, and the testing time-limit is not an issue (Mcmorris, Boothroyd, & Pietrangelo, 1997).

As is shown in this research, chemistry instructors tend to view this issue either from a psychometric perspective or from a humor perspective. Although there is merit behind both perspectives, in summative assessment circumstances used to assign grades, such as are often used in general chemistry courses, it may be best to avoid the use of humor and obviously implausible distractors.

4.2.2 Students should analyze data

Many of the participants held beliefs that students should analyze data on exams. To demonstrate this, three example excerpts will be presented. Many of the comments referring to this belief emerged from the second phase of the interview while participants evaluated item number six (Figure 4.2).

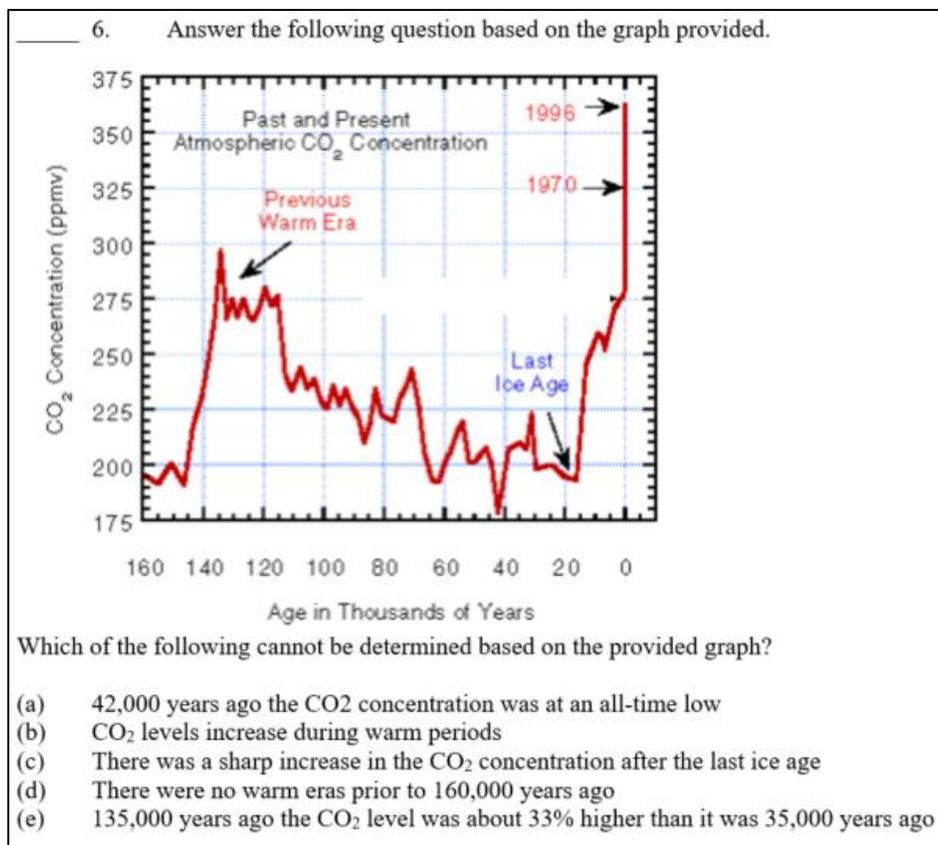


Figure 4.2 Item six evaluated during phase two of the interview protocol. This item requires the interpretation of graphical data

First, Dr Sorenson who had 5-7 years of general chemistry teaching experience and had administered 10-19 MC exams, said of item six:

I like the kind of question where you have to interpret actual data because I think that's really good for the students not only because they have to do that in lab but that's like what scientists do generally. So I like that.

Furthermore, Dr. Irvine, having over 10 year of teaching experience and having administered over 20 MC exams, said:

I really like this question. I think it's very important that students are able to analyze a graph. That's what scientists do. And people that aren't scientists can benefit a lot from learning to make good graphs and reading graphs correctly. So, I think this is a very important question.

Lastly, Dr. Deagen who had over 10 year of teaching experience and having administered over 20 MC exams, said:

Dr. Deagen: Number six I like because it's a graph. I love showing graphs, I love making them analyze data on an exam.

Interviewer: And why is that?

Dr. Deagen: Because it's more real life, it's more what they would see if they actually became a scientist, and it's really important to be able to interpret data if you're going to think like a scientist, more than just memorizing quantum numbers, right? I like these kinds of questions. Often, I'll show a titration curve or something, "What can you learn from this titration curve and why is it important?" No, I like that question.

From these data, it shows that many of the chemistry instructors interviewed believe that assessments should include the analysis of data. Many cite the fact that analyzing data is what scientists actually do as a reason for the practice.

4.2.2.1 Discussion

The fact that many of the instructors held beliefs that assessments should elicit data analysis skills is a promising finding in-line with literature recommendations such as those promoting the Next Generation Science Standards (NGSS) (M. M. Cooper, 2013; Lavery et al., 2016; Reed, Brandriet, & Holme, 2016). An integral part of the NGSS standards, which aim to provide a deeper level of learning, is the role of science practices in instruction and assessment (M. M. Cooper, 2013). Science practices are skills that scientists use to understand the world, such as analyzing and interpreting data (Next Generation Science Standards: For States, By States, 2013). The fact that instructors believe there is importance in assessing scientific practices such as the interpretation of data, provides a foundation to build upon in hopes to improve to quality of general chemistry assessments.

4.2.3 Number of concepts an item should test

Many of the participants held beliefs about how many concepts a MC item should test. These beliefs fell into two main categories, namely, those who like to test many concepts within one item and those who prefer to test very few, preferably one, concept per item.

4.2.3.1 Test many concepts per item

Those who believed that MC assessment items should test many concepts per item often cited content coverage as a reasoning for the practice. For example, Dr. Sorenson who had 5-7 years of experience teaching general chemistry and had administered 10-19 MC exams said about a stoichiometry item she created (Figure 4.3) during phase three of the interview:

But yeah, I picked this question and I wrote this question because it combines balancing equations, it combines stoichiometry, and it combines limiting reagents. And when we write tests for Course X we're only allowed twenty questions so (oh and it combines sig figs) so it's like four different things I can cover with one question. And that's why this is a great question for me because I can cover a lot of concepts with only one question. So, I like that.

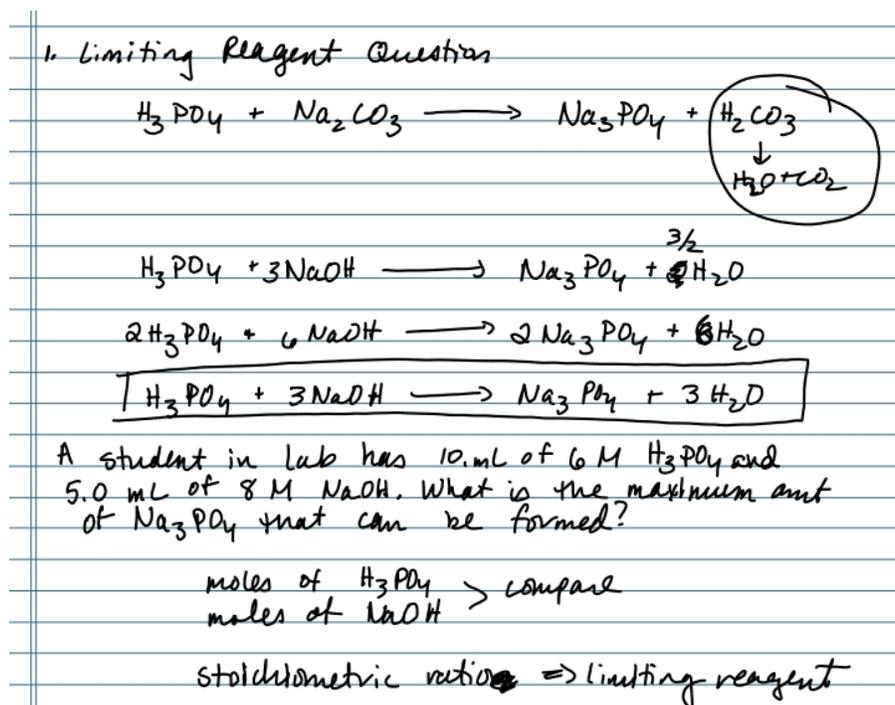


Figure 4.3. Dr. Sorenson's stoichiometry item created during phase three of the interview

Furthermore, when discussing an item she created about VSEPR (Figure 4.4) she said:

So, this would be a concept question that I would write because again it combines multiple concepts.

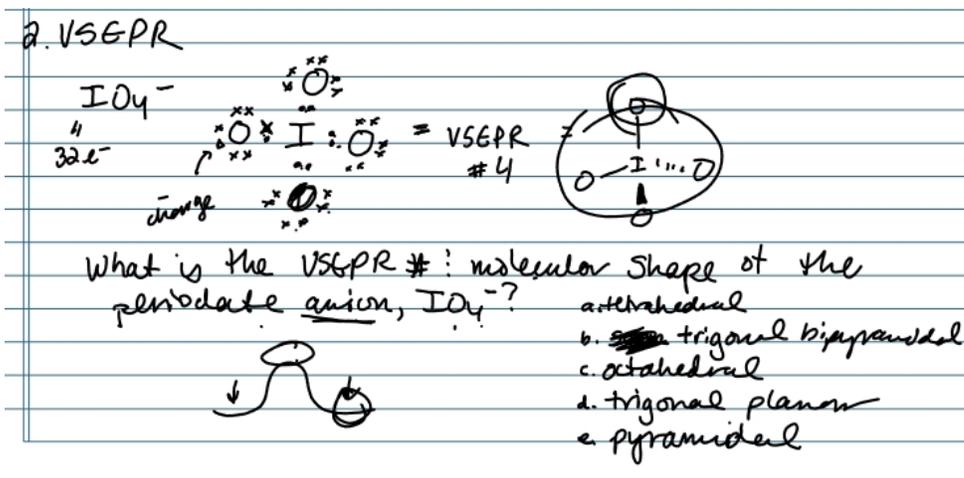


Figure 4.4. Dr. Sorenson's VSEPR item created during phase three of the interview

From these examples, we see that some instructors hold beliefs that testing multiple concepts in a MC item for content coverage purposes is desirable.

4.2.3.2 Test few concepts per item

Alternately, some instructors held beliefs that testing single concepts or skills in a MC item was the preferred method. Many of them cited being able to understand what a student knows about a specific topic from the assessment as the reason for testing only one concept per item. For example, Dr. Crawford who had extensive training in educational measurement, 5-7 years of teaching experience, and had administered 5-9 MC exams said:

One is this idea of in terms of measurement, it's always better to have an item that's testing only one skill. But in terms of exam writing, a lot of times people like to have items that are the "hard" items, and the reason they're hard is because they require multiple steps. So I try to be careful about doing that. Although both of the questions that I've written probably require multiple steps to get to. But I try to be careful that certainly not all of the questions should be that way, and it that it can be possible to write challenging questions that still are testing one skill. The problem is if you test multiple skills, if they got the question wrong, you don't know

if it's because they don't understand A or B, you know what I mean? It could be multiple things that are affecting it. Sometimes if you write good distractors, then you can get at that. So if I'm using a question that's going to have multiple skills within it, then I want to make sure that my distractors are going to be like this is the answer you would get if you knew this skill, but not this skill, does that make sense?

Furthermore, Dr. Tracy who had over 10 years of teaching experience and had administered over 20 MC exams said while evaluating item three (Figure 4.5) during phase two of the interview:

_____ 3. Which of the following are key differences between chemical and nuclear reactions?
I. Atoms do not change identity in chemical reactions, whereas in nuclear reactions they do
II. Nuclear reactions release a greater amount of energy than chemical reactions
III. Nuclear reactions have rates that depend on temperature, concentration, and catalysts, whereas chemical reactions do not
(a) I
(b) I, II
(c) II, III
(d) I, III
(e) I,II,III

Figure 4.5. Item three evaluated during phase two of the interview

Dr. Tracy: I try to keep multiple choice questions a little bit simpler and cleaner than that one.

Interviewer: Okay. Why do you like to keep them simpler and cleaner?

Dr. Tracy: Well, I don't like situations where if a student understands something mostly and just doesn't get one aspect of it, I don't want them to get an entire problem wrong. I think that one is carefully written, so I don't know that it would happen as much. I think if they've got the idea of what a chemical reaction is and what a nuclear reaction is, they should be able to get that right if you kind of circle one and two and put an X through three, and then look for just one and two response. Most of them should be able to get that, but if a student's a little bit ambiguous on three, there would be a probability they'd second guess themselves and get the entire thing wrong when they still mostly understand the idea.

From these examples we see that some of the instructors believed that including fewer concepts in an item was beneficial in order to enable appropriate interpretation of assessment results.

4.2.3.3 Discussion

The chemistry instructors tended to believe in either testing many concepts per item or in testing few concepts per item. These beliefs were centered around the ideas of content coverage and assessment interpretation, respectively. The number of concepts to include in any single MC item has been discussed in the literature (Haladyna & Downing, 1989) and measurement textbooks (Thorndike & Thorndike-Christ, 2010). The “correct” belief and practice comes down to the test creators’ purpose for the assessment or assessment item. If the test creator desires to assess if a student can use multiple pieces of knowledge to solve a complex problem, then including multiple concepts in an item can be desired. However, if the purpose of the assessment is to provide evidence that a student understands *specific* concepts and skills, then including many skills into a single MC item may be undesirable. Additionally, including too many concepts into one item has been shown to increase cognitive load and decrease performance due to working memory capacity overload (Johnstone & El-Banna, 1986; Niaz, 1987).

4.2.4 Algorithmic questions should be free response; Conceptual questions should be multiple-choice

Most of the participants made comments in regard to a belief that algorithmic (mathematical) items should be free response and that conceptual items should be multiple-choice. Many cited being able to interpret test results accurately and being fair to students for partial credit earned as reasoning for the belief and practice.

Dr. Deagen said after creating an algorithmic acid-base chemistry question during phase three of the interview:

Dr. Deagen: Normally, when I have other choices, often I will make this as a short answer, so there's lots of places for partial credit. That's why I do the short answer, not only can I see where they're going wrong

when their thinking is wrong, but I can give partial credit. Here it's all or nothing.

Interviewer: Okay, and so this is the type of question you would more likely do as short answer?

Dr. Deagen: Right. This is partially why I do fewer algorithmic questions in multiple choice, because there's so many places to go wrong that if you ... Having 10 points based on a math error is too much for me.

Furthermore, Dr. Johnson, who had 0-1 years of general chemistry teaching experience and had administered 5-9 MC exams, said about item one during phase two of the interview:

So, I'm not a big fan of question one, just because like I mentioned before, I don't like to put calculations in the multiple choice. And the reason for that is because students, they may get certain steps but somehow, maybe because of a calculation error, maybe just one small, the last step they made a mistake or even in the first step, just something that's a common misunderstanding, that they did. It really is not reflective of the students understanding and we cannot really gauge where they're not understanding something. And so that's why I don't like question one because it's purely mathematical and there's no partial credit and there's no way for us to see the student's thought process.

Dr. Smith said;

I won't give my students just multiple choice because sometimes I want them to do the math and things like that. But because of that, I also pretty much avoid doing multiple choice questions that are math. Just because I feel like the point of the multiple choice question for me is to get at their reasoning and the way to get to their reasoning for math, I think is easier to look at a page of their math and see where they went wrong.

4.2.4.1 Discussion

These thoughts and considerations are in-line with how The National Academy of Sciences described the purpose of educational assessment in their report *Knowing What Students Know*. In that report, assessment is described as a process of 'reasoning from evidence' where an instructor uses assessment data to make claims about what a student knows and can do (Pellegrino, 2001). As such, if a chemistry instructor wants to understand what his or her students know about an algorithmic or mathematical process, using a short answer item may provide more detail to 'reason

from the evidence' than using a MC item. However, MC items can be designed in a way where the selection of a distractor would suggest a certain mistake or lack of knowledge (Gierl, Bulut, Guo, & Zhang, 2017; McClary & Bretz, 2012). An instructor could then use student response patterns to provide evidence of mathematical or algorithmic process understanding. In conclusion, with the purpose of 'reasoning from evidence' in mind, a chemistry instructor can choose to use MC or free-response item format if the claims they are making about student understanding are supported by the evidence of the assessment type.

4.2.5 Levels of Understanding Tested

Many of the participants discussed a guiding belief that decisions about assessments and assessment items should be governed by the type of understanding they elicit. A type or level of understanding would include processes such as memorization, calculation, analysis, synthesis, etc (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956).

Dr. Crawford who had, extensive educational measurement training said about the items evaluated during phase two of the interview:

Dr. Crawford: Yeah, so all of the questions except for question one which required a calculation, which might be considered apply or understand at least, all of the questions two, three, and four have just been remember questions.

Interviewer: Okay. And why is that level of understanding something that's important to you when you're designing your assessments?

Dr. Crawford: Well chemists always talk about oh, chemistry's a place where we work on our critical thinking and problem solving skills and whatever, but if you look at our assessments and they're just like uh, can somebody memorize their way through it? Then we're not actually assessing those things that we claim to be teaching and we claim students are learning. So it's important that an assessment work at the level that we want students to be performing.

Dr. Crawford: It depends what class you're teaching. If you're in general chemistry, maybe you're fine getting to understand, apply, analyze, maybe evaluate, you know what I mean? You probably don't need students in general chemistry to create new chemical compounds or something. So maybe you don't need the highest levels of Blooms Taxonomy. At least you want to be working at understand and apply as opposed to just remember.

Interviewer: Okay. Yeah thank you-

Dr. Crawford: It's just bad teaching. The problem with having questions like this that are remember type questions is that it is contrary to the nature of science. It teaches people that science is a collection of facts to be memorized instead of a process of discovery and thinking and analyzing and learning things about the world. And that's wrong. It's wrong to give students the impression that science is dead and you just have to memorize it. That's not what science is.

Furthermore, Dr. Deagen said about the questions she creates:

I try and make them not a waste of time. You know? It's more like if they're too straightforward, it's like, "Why bother?" I try and make it useful for understanding why they should care about the material that they're looking at, not just for memorization sake or something. That's kind of where I stand.

Lastly, Dr Lopez who said:

I try the questions to not be boring in the sense that they should not be questions that students can answer without really having a understanding of the material, and should ideally not be too similar to what they've done before. So that it's not ... if they get it right, it's not because they memorized something, but because they really understood.

4.2.5.1 Discussion

Dr. Crawford, Dr. Deagan, and Dr. Lopez made comments that exemplified the belief that the level of understanding an item probes at helps to guide assessment decisions. Particularly, Dr. Crawford said it well as she spoke about how her assessments should be functioning at the same level she desires her students to be at as well.

Considerations of level of understanding are not foreign in the world of education or educational assessment. Perhaps the most well-known way to characterize this is Blooms Taxonomy, a hierarchy of cognitive processes ranging from memorization to creation (Bloom et al., 1956). These mental processes have guided educators' design of assessments in higher education science courses. For example, many studies have analyzed exam items for level of Blooms taxonomy (Stark, 2008; Thompson & O'Loughlin, 2015). Additionally, it is recommended that college instructors center their assessments around higher levels of blooms taxonomy than memorization and recall levels in order to promote deeper understanding (Lord & Baviskar, 2007). Thus, it is good that chemistry instructors hold beliefs that assessing deeper levels of learning is desired.

4.2.6 Assess widely applicable skills

Some of the participants expressed a belief that assessing widely-applicable skills is valuable in chemistry exams. For example, Dr. Deagen said:

Dr. Deagen: And third, who cares? That's a lot of what it is. It's like, "Why would I care about doing that?" A lot of the times when I'm asking these questions it's like, "Why do I want them to know it? Why do I care that they know it? And why is it important for what they're going to do next?" That's a lot of what informs me.

Interviewer: And those are big factors in what you used to determine what you put on an exam?

Dr. Deagen: Right, so for my course especially, they're going right into organic afterwards, and a lot of what drives me for that is what do they need to know to be successful in organic and biochemistry? This is certainly not something that I would care about.

Dr. Johnson said about writing out units in exam items:

Interviewer: Okay. You mentioned that you don't like writing out the units, you like to just write it out as M or g in this case, could you tell me why?

Dr. Johnson: Yes, because I think that students need to be exposed to units as they are, based on previous knowledge and based on what they learn, they should be familiar with the scientific notations and scientific units. And so, they should know that M is molar for example, and the

options, we have 50 mL and everybody knows mL is milliliters. And so, this is something that students, that is important for them as they go on in other courses, as they go on through the rest of the course, and so it almost seems like a crutch to write all these things out fully so that students don't remember it.

Lastly, Dr. Sorenson said after creating an item during phase three of the interview:

Dr. Sorenson: After I write the question I step back and and I'm like what is this really teaching them? Is this actually teaching them something important? Or is this a stupid detail they maybe got or maybe didn't get. So, I guess the difficulty level is something that I often consider as well.

Interviewer: And when you say important what do you mean by important?

Dr. Sorenson: Like important concepts you mean? Is that what I said? I guess it's concepts that are broadly applicable to other areas of science versus concepts that maybe you would only need to know if you were only studying chemistry.

4.2.6.1 Discussion

These examples show that some of the instructors hold a belief that assessing skills and knowledge that are useful in a wide variety of situations, is important to them in their assessment practices. Dr. Johnson spoke specifically about assessing students understanding of units because that knowledge would be useful to them even in other disciplines. Furthermore, Dr. Deagen spoke about using the thought, 'Is this going to be important for what they will do next?' as a guide for making decisions about assessments.

This belief is in-line with popular educational motivation theories, such as Expectancy Value Theory (Eccles, 2013). Additionally, the NGSS supports the idea that we should connect our teaching to Crosscutting Concepts which are applicable to many disciplines (M. M. Cooper, 2013; *Next Generation Science Standards: For States, By States*, 2013). It is appropriate that chemistry instructors believe that assessing widely-applicable skills is an important aspect of their chemistry courses.

4.2.7 Equitable items for all learners

A few of the participants held a belief that assessment items should be fair to all students. Although this belief was not described by a majority of the participants, it was a noteworthy finding, and should be reported and discussed. Dr. Sorenson who had 5-7 years of general chemistry teaching experience and had administered 10-19 MC exams, said about the words ‘hydrochloric acid’ in item five (Figure 4.6) of the second phase of the interview:

And the other thing I will say about this question is, a lot of times when I write multiple choice questions for Course X I have to be very conscious about the fact that not everyone in the class is a native English speaker, so if there's words they're not going to recognize, I don't use them. So they might know, they might not know what the word hydrochloric acid is in English but if you wrote HCL they would definitely would know that was, so I also think that's a consideration but I don't know. I did my post doc in Germany and I didn't really speak German and all of the chemical labels in the building were in German, so I had to learn a lot of German names of chemicals so I guess from that experience I'm kind of like, not everyone knows.

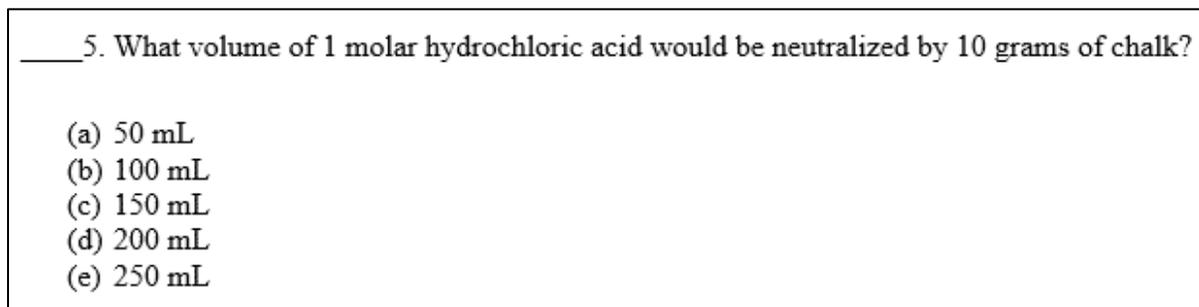


Figure 4.6. Item five evaluated during phase two of the interview

Additionally, Dr. Brown who had over 10 years of general chemistry teaching experience and had administered over 20 MC exams made comments about equitable items while evaluating the third item (Figure 4.7) during phase two of the interview. She said:

Dr. Brown: One thing I would maybe do, just because with enough students with disabilities is I would maybe do like 1, 2, 3, not Roman numerals one two three.

Interviewer: Okay. And that's based on your experience with students with disabilities, you said?

Dr. Brown: Yes, students with disabilities, it would be easy for them to misread, where they all look so similar to each other, and I just wouldn't want to, again, I've learned after my many years of the trauma that students can bring to me that I have to then deal with, and so I try and eliminate that.

_____ 3. Which of the following are key differences between chemical and nuclear reactions?

I. Atoms do not change identity in chemical reactions, whereas in nuclear reactions they do

II. Nuclear reactions release a greater amount of energy than chemical reactions

III. Nuclear reactions have rates that depend on temperature, concentration, and catalysts, whereas chemical reactions do not

(a) I

(b) I, II

(c) II, III

(d) I, III

(e) I,II,III

Figure 4.7. Item three evaluated during phase two of the interview

Lastly, Dr. Madison said about the graph used in item 6 during phase two of the interview:

What I don't necessarily like about the graph is the use of blue and red, only because, and I'm not, but I'm always conscious of those people who are colorblind, and that red-blue, I don't know. Does it make any difference? They might see something.

4.2.7.1 Discussion

From these data, we see that some of the instructors held a belief that making items equitable for their students is important. Dr. Sorenson discussed non-native English speakers understanding of chemical terminology, Dr. Brown discussed students with disabilities, and Dr. Madison discussed colorblind student's ability to interpret a graph.

This belief of equitable assessment is in-line with beliefs outlined by Abell and Siegel in their work creating the Science Teacher Assessment Literacy Model (Abell & Siegel, 2011). It is hopeful to see beliefs emerge about being equitable to students in this study. Creating equitable

assessments is an important aspect of assessment design (Thorndike & Thorndike-Christ, 2010) and seeing it discussed among chemistry instructors is promising. Many studies have focused on equality in testing, particularly studies looking at differential item functioning among various demographics of students (Kendhammer, Holme, & Murphy, 2013). It is hoped that this belief of ensuring equitable chemistry assessments may be propagated and studied further in the future.

4.3 Knowledge of Assessment Purposes

Abell and Siegel identified four main categories that assessment purposes can be grouped into: diagnostic, formative, summative, and metacognitive (Abell & Siegel, 2011). The data from this research has shown that the chemistry instructors interviewed considered formative assessment purposes when designing MC assessments.

4.3.1 Use MC assessment data to shape instructional decisions

Some of the instructors made comments which showed their understanding that one purpose of assessment is to inform teaching practice. To illustrate, Dr. Deagen said:

Interviewer: Testing more than one concept in a question, you like doing that in an exam? Can you tell me why you like that?

Dr. Deagen: Because it probes more deeply into their understanding of something. It's like, you have to know this part to know this part, so it lets me know. Then when I get the data back, then I know, "Where's this falling apart? Is it falling apart in the geometry or is it falling apart in the polarity?" Then you can go back and remediate that.

Furthermore, Dr. Smith said about MC exams:

When I first started teaching, I was of the opinion that multiple choice was a cop out. I used to think that an instructor did it because they didn't want to have to deal with the grading. But then later I realized that oh no, wait. If you have a really good multiple choice question and you have really good distractors, it can actually be

really insightful. And it's also very easy then to tabulate the data and start to figure out where the misconceptions are. So, now I'm a pretty big fan.

4.3.2 Discussion

In the examples above, we see two instances of instructors using assessment for the purpose of informing decisions in their classrooms. Dr. Deagen discussed using the data she gets back from an assessment to see where students may be struggling and then using that information to know what to re-teach in class. Additionally, Dr. Smith discussed how her views of MC assessments have changed as she began to use them for the purpose of uncovering her students' misconceptions.

It is interesting to note that in these cases the instructors were referring to using summative assessment tasks for the formative purpose of changing teaching practice. The instructors were referring to assessments used to assign grades, typically given at the end of a unit or semester. While it is always good practice to learn from and adapt practice based on assessment data, seeing instructors discuss using their summative assessments in a formative manner is worth noting. This is because formative and summative assessment types tend to have different purposes (Pellegrino, 2001). Formative assessments are to assist in learning and teaching while summative assessments are to give a snapshot of individual achievement (Pellegrino, 2001). Using a summative assessment task for a formative purpose is therefore unexpected but would not be considered poor practice. Additionally, although not specifically mentioned, it would be desired that the instructors were using homework assessments in a similar way to inform teaching practice.

4.4 Knowledge of what to assess

Part of assessment literacy according to Abell and Siegel is an instructors knowledge of what to assess in their classroom (Abell & Siegel, 2011). During the interviews, many instructors did

consider aspects of what to assess while creating and evaluating assessments. While coding the transcripts, three main codes emerged within the knowledge of what to assess category. These codes include what to assess, determining what to assess, and finding the items. ‘What to assess’ refers to what an instructor chooses to assess in an exam. ‘Determining what to assess’ refers to how an instructor decides what to put onto an assessment. ‘Finding the items’ refers to how an instructor generates items for his or her exams. Substantial results from each of these codes will now be presented and discussed.

4.4.1 What to assess

What do general chemistry instructors find important to assess?

What to put onto a chemistry exam is one of the paramount decisions that an instructor makes during the exam creation process. Considerations of what to assess were common during the interviews. Many instructors mentioned assessing data analysis/lab skills, conceptual understanding, misconceptions, representative course coverage, and widely applicable knowledge. Substantial results will now be presented and discussed.

Assessing data analysis and lab skills was a common thread among those interviewed. For example, Dr. Deagen discussed making students analyze data on an exam after evaluating item six during phase two of the interview (see Figure 3.3 in section 3.6.2):

Dr. Deagen: Number six I like because it's a graph. I love showing graphs, I love making them analyze data on an exam.

Interviewer: And why is that?

Dr. Deagen: Because it's more real life, it's more what they would see if they actually became a scientist, and it's really important to be able to interpret data if you're going to think like a scientist, more than just memorizing quantum numbers, right? I like these kinds of questions. Often, I'll show a titration curve or something, "What can you learn from this titration curve and why is it important?"

Furthermore, Dr. Smith discussed the importance of assessing data analysis skills in exams.

Dr. Smith said:

Dr. Smith: Okay. I like this question from the standpoint of it's interpreting graphs is really important and I think it's a nice way to do it in a multiple choice. In fact, it makes me think that maybe I could turn graphs into multiple choice questions. So I like that about it.

Interviewer: And why is interpreting graphs important?

Dr. Smith: Well, because I teach a lab course, right? The students collect data, they then plot their data, and it's important that they can figure out what the graph is telling them. But I feel like that's something that they're actually quite weak at. And also, so I think they have a hard time interpreting how to read it and how to interpret which ... I think it's important for a scientist in general that they can interpret graphs because that's how usually we share our data with each other. It's the most convenient way to share your data.

In addition to data analysis skills, several instructors mentioned choosing to assess conceptual understanding in their exams. Dr. Johnson mentioned this when discussing her students:

We've noticed that in our students, I'm sure it's the same other places, that they just cannot make connections between previous courses and future courses, and that's mainly because of some conceptual deficit. And so that's why I want to address that in my multiple-choice questions

Furthermore, Dr. Smith considered assessing conceptual understanding when designing an item in the third phase of the interview. Dr Smith decided to create an item based on acid-base titration. In the following excerpt she was deciding whether to make the item more focused on the concept or the math by changing the concentrations of the reagents. She said:

Dr. Smith: So, I would know the concentration of NaOH. But even then, I would start making decisions about, do I want to have the same concentration of H_2SO_4 , or different concentrations. 'Cause if they're the same concentrations, right, then I can make this very mathematically easy. But getting more at the concept. Or I can give

them really wonky concentrations and then it turns into some math as well as the concept.

Dr. Smith: I have to start making decisions about what exactly I'm trying to get at in the problem.

Interviewer: Okay. And so, it's based on what you would want to test them on that you would make that decision?

Dr. Smith: Yeah. What I think is more important.

Dr. Smith: That you punch these numbers into their calculator and calculate the moles and then go ahead and calculate new volumes. Or if I care more about at face value they can look at it, and maybe even do the calculation in their head. But understand the ratio is two to one, right. So I think in my case I would care more about, forget them trying to calculate weird numbers of moles. Because that's a trivial calculation. And instead focus on the main part of the interest here, the misconception about how that ratio factors into the volume needed for the titration.

Several of the instructors mentioned that they find value in assessing student misconceptions in their exams. For example, Dr. Johnson said:

I use my clicker questions, one misconception students had or what they answered wrong, and then I think of a deeper level multiple choice question based on that. I like testing my students, I know it sounds mean but I like testing my students on things that they got wrong in class just to see if they actually did go back and study it or not. And usually my clicker questions are based on topics that are important or very conceptual. And so, I like to go through my slides first to see what topics are kind of needed in the exam for the multiple choice

While Dr. Johnson created a MC item on hydrogen bonding, she demonstrated designing an assessment item based on misconceptions. Figure 4.8 shows an item Dr. Johnson created. The following two quotes were from the think-aloud portion of her interview referring to the item in Figure 4.8.

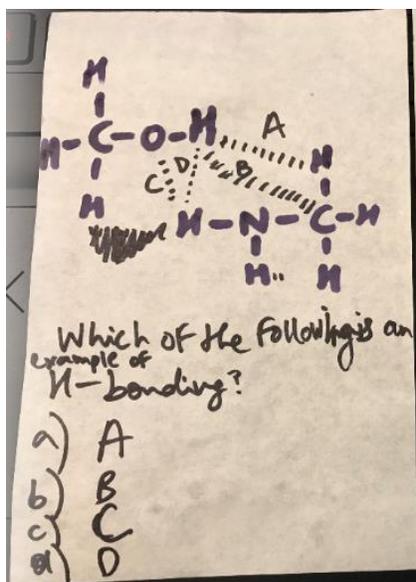


Figure 4.8. MC item created by Dr. Johnson about hydrogen bonding

And so, I now want to address the topic of hydrogen bonding, which is something that students don't get, that's actually what they have the most trouble with. And so I want to create a question with hydrogen bonding involving molecules.

And so, a common misconception that students have is that they think that a hydrogen bond is between two hydrogens, and instead of a hydrogen bonded to an oxygen nitrogen fluorine, that gets bonded to another oxygen nitrogen fluorine, they often think that it's between two hydrogens because of the name. And so, I'm going to make that as one option as well, where I'm gonna show a bond between two hydrogens, does that make sense?

Furthermore, Dr. Smith focused on misconceptions while creating an item in the third phase of the interview, she said:

So that would mean I would first generate an acid base titration graph, right. There's so many places I could go. There's so many misconceptions here. The question is where do I want to take this. So now I need to figure out, I want to use a graph and I want to ... I have to figure out what misconception I want to try to get to.

One of the most widely talked about aspects of what to assess was a desire to have representative content coverage on an exam. For example, Dr. Sorenson discussed the exam

creation process in a team-taught course and how representative coverage was a key factor. Dr.

Sorenson said:

We usually meet every week during the semester and so our individual lists of questions are due two weeks before the exam and so that way what we will do we will get together at a meeting and we'll go through everybody's questions and then we'll pick the ones that we like that represent the topics that we want to test on. And so we'll come up with kind of a short list and then we'll make a test based on that and we'll meet the following week and we'll finalize the test and make ensure that all the questions are representative of all the topics we want to cover.

Furthermore, Dr. Bennett mentioned assessing representative coverage. He said:

I go through the learning objectives that we've taught and make sure that we identify what we think are the critical things they should have learned. And there... And obviously we can't have a question for every single one, but we pick ones both that are representative of what we have taught for that module.

Dr. Sanders also valued assessing representative coverage in general chemistry exams. He said:

Coverage, amount of time spent on each topic. That ought to be reflected in the number of questions on the exam. Four lectures versus six lectures, two different topics, the six lecture should have more questions associated with it than the four lectures.

Dr. Johnson also wanted to assess representative coverage in her exams. She said:

Dr. Johnson: Yes. And so we have 4 units and so, first, I know that there's a certain number of points that I allocate for multiple choice questions, and so each of the multiple choice questions is worth 3 points, sometimes 4, and so depending on the number of points that are allocated, I divide it up. And so, let's say I have 8 multiple choice questions per exam, I then divide those 8 up between the chapters that I taught and so 2 for thermodynamics, 2 for electro chemistry, 2 for kinetics, and then 2 for organic, whatever else is there or sometimes it's like 3 and 2. But, I make sure that I cover all the topics equally in the multiple choice question.

Interviewer: Can you tell me why you like to do that?

Dr. Johnson: Yes, because I, each of the, all the topics, all the units that I cover, I take, when I prepare my course plan I make sure that they're all covered within the same amount of time, I give same importance to all the topics and because all these topics are equally important, and so I want to test my students equally too, I don't want them to feel that one topic is more important and then they spend more time studying that and then they neglect the other topic. And so that's why I do it.

While many of the instructors discussed desiring representative coverage for their exams, only a few mentioned actual strategies for ensuring this to happen. Dr. Brown mentioned making a list of learning objectives and making sure he has items for each objective. He said:

Dr. Brown: So, I usually start by listing the main topics that we have covered during that unit. And then sometimes I break those up into sub-topics. But start by listing and make sure I have at least one question on each of those topics and then some that I think I need more on.

Interviewer: Okay. That's great, thank you so much. And this is related to what I've asked previously, but just in a broad overview, would you be able to walk me through how you would typically create a multiple-choice general chemistry exam for your students?

Dr. Brown: Yes. So, just broad, I would start by making a list of all the topics that we had covered that needed to be on the exam. And then I usually write the individual questions first, and make sure I have a question for each topic or if I need more than one.

Furthermore, Dr. Crawford who had extensive training in educational measurement, discussed making a list of learning objectives and a table of specifications to guide her in the assessment design process. She said:

So I take my list that I have of learning objectives of things that we've done in the class and then I usually, like I said I'm starting from an existing test, so I'll sort of map it out onto the learning objectives and see how it lines up and am I doing a good job representing the range of all the learning objectives that we've done?

...What I'm saying is I look at the exam level to make sure that we're covering the objectives and having the range of Blooms Taxonomy that we want to have. So, something like a table of specifications. I make tables of specifications. I usually

make them from the existing assessment and then modify to make the table of specifications look the way I want it to, you know what I mean? So, if I have too much in one area, I'll move things around a little bit.

The instructors interviewed also discussed the importance of assessing knowledge and skills that would be applicable to their students in future courses or careers. Often, these considerations emerged around discussions of data analysis. For example, Dr. Deagan discussed how data analysis skills would serve her students when they become scientists. She said:

Dr. Deagan: Number six I like because it's a graph. I love showing graphs, I love making them analyze data on an exam.

Interviewer: And why is that?

Dr. Deagan: Because it's more real life, it's more what they would see if they actually became a scientist, and it's really important to be able to interpret data if you're going to think like a scientist, more than just memorizing quantum numbers, right? I like these kinds of questions. Often I'll show a titration curve or something, "What can you learn from this titration curve and why is it important?"

Furthermore, Dr. Smith described how interpreting graphs and data would be useful for students when they become scientists. She said:

Interviewer: And why is interpreting graphs important?

Dr. Smith: Well, because I teach a lab course, right? The students collect data, they then plot their data, and it's important that they can figure out what the graph is telling them. But I feel like that's something that they're actually quite weak at. And also, so I think they have a hard time interpreting how to read it and how to interpret which ... I think it's important for a scientist in general that they can interpret graphs because that's how usually we share our data with each other. It's the most convenient way to share your data.

Moreover, Dr. Deagen explained why she did not like the fifth item evaluated during phase two of the interview (see Figure 3.3 in section 3.6.2). In this explanation she discussed how the applicableness of the knowledge tested in an exam was an important consideration to her. She said:

Interviewer: Okay. What would be the reasons why you would throw it out again?

Dr. Deagen: ...And third, who cares? That's a lot of what it is. It's like, "Why would I care about doing that?" A lot of the times when I'm asking these questions it's like, "Why do I want them to know it? Why do I care that they know it? And why is it important for what they're going to do next?" That's a lot of what informs me.

Interviewer: And those are big factors in what you used to determine what you put on an exam?

Dr. Deagen: Right, so for my course especially, they're going right into organic afterwards, and a lot of what drives me for that is what do they need to know to be successful in organic and biochemistry? This is certainly not something that I would care about.

4.4.1.1 Discussion

In this research, the chemistry instructors interviewed thought there were several important things to assess in general chemistry exams, including data analysis/lab skills, conceptual understanding, misconceptions, representative course coverage, and widely applicable knowledge.

It is promising, and not surprising, to see chemistry instructors value the assessment of data analysis and lab skills. Laboratory skills and analysis techniques have been a staple of general chemistry courses for years, although there is little evidence of their necessity for chemistry learning (Bretz, 2010, 2019). With that said, being able to analyze data has been identified as a 'big idea' in the anchoring chemistry content map for general chemistry (Murphy, Holme, Zenisky, Caruthers, & Knaus, 2012) and a science and engineering practice in the Next Generation Science Standards (*Next Generation Science Standards : For States, By States*, 2013). A focus on assessing data analysis and lab skills is in-line with the literature on what to assess in general chemistry.

Assessing conceptual understanding of general chemistry has been a focus of the chemical education community (Bretz, 2014; Nakhleh, 1993; Smith, Nakhleh, & Bretz, 2010). Thus, seeing chemistry instructors choose to assess conceptual understanding is no surprise and is in-line with literature and prior practice.

Much like assessing conceptual understanding, characterizing student misconceptions has been the focus of many chemical education research studies (Banerjee, 1991; Duis, 2011; Herrmann-Abell & DeBoer, 2011; Orgill & Sutherland, 2008). Focusing assessment on student misconceptions is an interesting finding. Seeing Dr. Smith and Dr. Johnson design MC items based on misconceptions and incorporate misconceptions into their distractors shows that they place value on exposing (and possibly mitigating) their own students' misconceptions.

The most common thing participants wanted to assess in their exams was not a thing at all, but the idea of representative coverage. This was seen among many participants with only a few participants citing strategies for achieving this. Representative coverage in an exam is important for establishing validity and drawing appropriate conclusions from exam results (Thorndike & Thorndike-Christ, 2010). Outlining the content of a course through listing learning objectives is the beginning of achieving representative exam coverage and has this been seen with the ACS exams in the development of the anchoring chemistry concept maps (Murphy et al., 2012). Once learning objectives have been established, representative exam coverage can be achieved by making a written description of the skills, knowledge, and or the types of abilities desired to be tested on the exam (Fuhrman, 1996). These written descriptions are often referred to as a table of specifications or a test blueprint. It has been shown that creating, using, and then distributing test blueprints to students facilitates positive student attitudes about exam fairness without inflating exam scores (Young et al., 2019). Few instructors interviewed in this research mentioned using test blueprints to guide exam development. However, it is probable that instruction on the use of test blueprints may be received favorably by chemistry instructors due to the strong desire for representative exam coverage.

Lastly, chemistry instructors expressed a desire to assess widely applicable knowledge and skills. This finding shows that the chemistry instructors interviewed were thinking about the importance of general chemistry in the lives of their students and how their exams could best represent widely applicable knowledge and skills. These findings are in-line with current literature suggesting that the content taught in general chemistry should be thoughtfully considered and chosen, as shown in the CLUE curriculum (M. Cooper & Klymkowsky, 2013). Additionally, curriculum reform efforts should be driven by our assessments (Holme et al., 2010), thus, seeing chemistry instructors value the assessment of widely applicable knowledge and skills is promising for the future reform efforts of general chemistry.

4.4.2 Determining what to assess

What influences how general chemistry instructors determine what to assess?

Determining what to assess refers to how an instructor decides what to put onto an assessment. Throughout the interviews, three main determinations emerged as being important to chemistry instructors. The chemistry instructors mainly determined what they would assess based on, lecture, learning objectives, and the level of understanding tested by an item. Example quotes will now be presented and discussed.

Many instructors used lecture, lecture notes, and or lecture slides to help them determine what to assess. For example, when asked how he would make a general chemistry exam for his students, Dr. Lopez said,

And then I would go through my notes, I think that's how I do it. I go through my notes that I used for preparing class

Furthermore, Dr. Sanders mentioned using lecture as a guide for determining the content and the coverage of what to assess. He said:

Whether I'm teaching as part of a team or by myself, my interest in exam questions is really focused on the lecture material. I tend to use my lecture notes primarily as a guide to try and figure out what it is I want to see if they know. I do pull old exam questions as part of that process.

If I cover that material in lecture and they did an experiment on that in the lab, and you don't always have that, there would be a greater weight of questions on that particular topic because they had it in two places. That's kind of my reasoning in terms of number of questions and what the questions should address. It's subjective, very subjective. There are things in lecture, for example, that I really try to emphasize, things that I want them to know four years later. Those will get hammered on exams more than cursory topics.

Moreover, Dr. Johnson used lecture slides to help her to determine what to put into her exams. She said:

And then I go through my lecture slides, I go through important sub topics within each of those topics, I look at my clicker questions and I try to think about, then I use my clicker questions, one misconception students had or what they answered wrong, and then I think of a deeper level multiple choice question based on that. I like testing my students, I know it sounds mean but I like testing my students on things that they got wrong in class just to see if they actually did go back and study it or not. And usually my clicker questions are based on topics that are important or very conceptual. And so, I like to go through my slides first to see what topics are kind of needed in the exam for the multiple choice

Another prominent way that the chemistry instructors determined what to assess, was through consulting their learning objectives for the course. Dr. Irvine discussed using his learning objectives in a backward design approach to determine what to assess. He said:

The idea is they have to be tied to the course outcomes. I need to make sure that I'm giving them enough practice and enough exposure, so that they can, on their own, do the skills that we need them to do. And so based on those outcomes, I look for questions that will, if answered correctly, provide some suggestion of evidence, right? And so we typically, all of us, do that backwards design approach.

Furthermore, Dr. Irvine continued by demonstrating this practice during the third phase of the interview. While deciding which topic to create an item on, he cited course outcomes in the process.

He said:

I'm going to choose the first one, and one of my outcomes is that students need to be able to produce and name the geometry of a molecule, both the electronic and the molecular geometry. So, a question I could write for that, would be, "Which of the following molecules has a square pyramidal molecular geometry?"

Additionally, Dr. Lopez mentioned using learning objectives to guide his exam development process. He said:

Okay, yes. First of all, I would go back to the learning objectives that were set at the beginning of the semester, and the learning objectives for each topic, and I would think of those goals before writing the question.

Also, Dr. Crawford mentioned using learning objectives to guide the content coverage of an exam. Dr. Crawford said:

And then in the event that we didn't have a question that would address a particular learning objective that we have, then I might write a question and I might use the textbook or I might use whatever lecture slides I had, you know what I mean? A variety of different sources.

The third prominent way that the chemistry instructors determined what to assess was through considerations of the cognitive skills an item may test. Cognitive skills refer to mental processes such as memorization, evaluation, analysis, synthesis, etc (Bloom et al., 1956). For example, while evaluating the second item during the second phase of the interview (see Figure 3.3 in section 3.6.2), Dr. Smith used considerations of cognitive skills to determine if she would use a MC assessment item. She said:

Dr. Smith: I would probably not use question number two either.

Dr. Smith: Just because question number two is less of an application and more of a general recall. Have you memorized what each quantum number describes. Rather than using that quantum number in a way that I think is a little more interesting.

Interviewer: And why is that something that's important to you?

Dr. Smith: Well, I guess when I write my exams, I usually try to write them with higher order questions in mind. It's very rare that I write a question that is more about just general recall. I think my students don't love that about me. But I feel like for me, I just feel like ... I give quizzes, and the quizzes have more general recall type of things. But when it comes to actual exams, then I want my students to be operating at a higher level. So maybe question two I could use a quiz, right. Just kind of a very general.

Furthermore, while creating an item during phase three, Dr. Sanders considered the cognitive skills required by the item when determining what to assess. He said:

I think a solid question for VSPER would be, 'what is the molecular shape of sulfur tetrafluoride?' That's a molecule that they're not necessarily familiar with, so it's not going to be simple recall.

Additionally, Dr. Crawford thought about cognitive skills while deciding to make an item on intermolecular forces during phase three of her interview. She said:

Okay, so intermolecular forces... well I think I'm going to go with a question that's higher on Blooms Taxonomy, so I'm going to go with a ranking question because that's like an evaluate question. I've used questions that ask students to rank things as evaluation because they're deciding on the value and putting them in order.

4.4.2.1 Discussion

Three substantial ways instructors used to determine what to assess were though using lecture, learning objectives, and considerations of cognitive skills.

It is interesting, yet not surprising that many instructors rely on their lecture material as a guide when determining what to assess. Being that chemistry instructors tend to assess what they teach, seeing the use of lecture material as a guide for what to assess is not a surprising practice.

However, in the literature it is recommended to use learning objectives (not simply lecture material) to guide what to put onto chemistry exams (Towns, 2014). This helps to ensure appropriate content coverage and assists in establishing exam validity (Thorndike & Thorndike-Christ, 2010). The practice of using lecture material alone as a guide, although convenient, is not considered best practice for assessment design and should be discouraged from use.

It is promising to see that many of the instructors interviewed used learning objectives to guide them in what to include in their assessments. This is promising because assessment design guided by a set of learning objectives is considered best practice (Butler, 2018). An additional (and practical) aspect of aligning assessments with learning objectives is the use of test blueprints (Fuhrman, 1996). Although using learning objectives as a guide for their assessments was a common practice, very few instructors mentioned using a test blueprint to facilitate exam-objective alignment. Recent literature has even shown that the use of test blueprints in chemistry courses can help to facilitate positive student views about course transparency (Young et al., 2019).

The use of learning objectives to guide exam development is a literature-supported practice that should be continued to be used by chemistry instructors. The use of test blueprints, although not discussed heavily by the instructors in this study, is also a recommended practice and should be adopted by instructors of general chemistry.

Considerations of cognitive skills tested was also a substantial way that the chemistry instructors determined what to include in their assessments. For example, when Dr. Sanders and Dr. Crawford were creating MC items, they avoided recall items and desired to test higher-order thinking skills, respectively. These findings are in-line with relevant literature which shows the chemistry education community being interested in testing higher-order cognitive skills in their assessments (Fensham & Bellocchi, 2013; Lord & Baviskar, 2007). Additionally, the NGSS

science standards which aim to promote higher-order thinking (*Next Generation Science Standards: For States, By States*, 2013) have been used and promoted in undergraduate level chemistry courses (M. M. Cooper, 2013; Lavery et al., 2016). Thus, seeing instructors consider higher-order thinking skills when deciding what to assess is encouraging, and not surprising, based on current interest in higher-order thinking skills in chemistry education.

4.4.3 Finding the items

How do general chemistry instructors generate MC items for their exams?

A large part of MC assessment design and an instructor's knowledge of what to assess is generating items for an exam. Strategies for how the participating general chemistry instructors accomplished this emerged from the data. The most prominent strategies were finding items in textbooks or item banks, in previously administered exams, or personally creating items. In connection to using items from textbooks and previously administered exams, a theme of modifying those items before inclusion was also present. Example quotes will now be presented and discussed.

Many of the instructors used textbooks or item banks to generate items for an exam. For example, Dr. Crawford discussed using (and modifying) items from a textbook to use in her exam. She said:

So, I went through and kind of deleted some [items] and added some different ones. Usually when I was doing that, I would take them from the test bank that came with the textbook, although they often need some modification because the test bank is not amazing, right? I work with what was existing material.

Additionally, Dr. Tracy mentioned that he uses test banks to find items because it is easier than coming up with answer choices himself, and he believes test bank items have been vetted for item quality. Each of the following three paragraphs is a different quote from Dr. Tracy. He said:

I get a moderate fraction from a couple different test banks from different textbooks. I think the most difficult thing about writing a good multiple-choice question is coming up with the distractors. And so, for a lot of the problem types, I'd rather find one that's kind of been used and vetted that has kind of a good number of distractors that are associated with it.

I try to use test banks because there's a lot of problem types where I don't feel I can come up with the distractors well. Or maybe where I'm too lazy to sit there and try to think of all the mistakes that they're likely to make and come up with them. That might be a better way of putting it.

I will also say, though, with multiple choice, it is easy to write questions that seem good and students don't quite get the idea behind, there's a mistake in there. I figure with test banks, they've been tested on students previously, and you can kind of see that students at this level are likely to be able to get that question right.

Furthermore, Dr. Johnson also uses a textbook to acquire items for her exams although she does modify them before inclusion. She said:

So, I use the textbook, we have a textbook which is by Atkins which is not great for multiple choice questions, but I use a textbook by Tro. And so, he has some conceptual questions in the end of every chapter and so even though they're not multiple choice, I try to take the questions from them and make my own answers.

Another source that the participating instructors used to generate items was the use of previously administered exams. As was seen with the use of test banks, many of the instructors would modify items before including them in an exam. For example, Dr. Tracy discussed using old exams as a resource. He said:

Usually, what I will do is I will go back to an exam from two to four years ago, and kind of find a set of multiple-choice questions that isn't too far from what I'm going to do, and I'll keep about half of them, and then I'll go through test banks. I'll put in maybe a new half, and sometimes I'll kind of change a couple of the multiple choice around a little bit. That's kind of the main thing that I'll do when I'm preparing an exam

Additionally, Dr. Patrick who had 0-1 year of general chemistry teaching experience and had administered 1-4 MC exams, used previously administered exam items. She used exam

statistics and item topic to help her determine which old exam items to use. She said:

So first I'll go through the pool of multiple choice questions that have already shown to be effective, like in that sense I'm looking at the [statistics] report that we get, see how effective those questions were and how well the students answered those and how well they targeted the learning outcomes that I really wanted to get at.

I mean, as I told you I basically use the database that the Gen Chem Office has offered, with all the different exams that they have had and I went like back 10 years and I looked at all the different questions, looked at the topics I was teaching at that time, and picked out all the questions that related to the topic.

Furthermore, Dr. Sanders mentioned using previously administered exams to generate items. He said:

Yeah. I recycle questions. I do write some new questions. I use old exams as a major resource. Sometimes I use questions as written. Sometimes I modify them.

Another significant way that the participating instructors generated items for their exams was by personally creating them. For example, Dr. Madison mentioned that he and his colleagues creates their own items. He said:

But in that time, we've written hundreds of our own multiple-choice questions. Or, lots of times, fill in the blank calculated questions, and so, I don't know, probably ... This coming up break, I might add a few questions, but I do reuse questions, but I'd say, during this coming up break, I'll probably write a few more.

Additionally, Dr. Smith mentioned personally writing exam items and then getting input from colleagues. She then went on to describe an item she created by herself that she was proud of. She said:

I write my exams myself. And then oftentimes I'll share my exams with a colleague afterwards to get their input.

That's probably the proudest multiple-choice question, the thing I'm most proud of because I made it pretty much entirely from scratch. So, I was like, oh but it was really hard.

Furthermore, Dr. Crawford described a culture in her department of using previously administered exams and how she would use those as a starting point in her exam creation process. She then mentioned creating her own items in addition to using other item acquisition strategies such as using a textbook or lecture slides. She said:

And then in the event that we didn't have a question that would address a particular learning objective that we have, then I might write a question and I might use the textbook, or I might use whatever lecture slides I had, you know what I mean? A variety of different sources.

Lastly, Dr. Patrick created some of her own items for her exams. After using previously administered exam items, she described creating additional items to address unassessed learning objectives. Although she did mention that creating her own items was less common than other strategies. She said:

Dr. Patrick: And then I would just create new questions based on if there's any holes to fill.

Interviewer: Okay. Yeah. Along those lines, do you usually make your own questions, or do you usually find them in other sources or modify them? How has that worked for you in the past?

Dr. Patrick: I did create a few on my own, but I would say that the majority I found in different resources.

4.4.3.1 Discussion

Seeing the participating instructors use textbooks, old exams, and their own minds as sources for general chemistry exam items is not surprising and may be partially explained by current literature. This literature describes some of the main barriers to MC item writing as being time and a lack of motivation (Karthikeyan, O'Connor, & Hu, 2019). More specifically, a lack of

time has been described as a difficulty for chemistry instructors in regards to assessment design (M. E. Emenike et al., 2013). A lack of time and motivation may be a determining factor for how chemistry instructors are generating MC items; more research on the subject is recommended.

Using available resources such as textbooks or old exams is not poor practice if the instructor is able to effectively evaluate the quality of an item before including it into a current exam. Thus, it is promising to see that many of the instructors modified existing items before including them into their exams.

Contrarily, there may be a lack of understanding of, or ability to evaluate the quality of test bank items. This was demonstrated by Dr. Tracy when he described using test bank items because they may have been vetted for quality. However, test banks are known for containing flawed items (Hansen, Dexter, & Hansen, 1997).

Furthermore, having instructors personally create items for their exams is unsurprising, as over 90% of chemistry instructors use personally written assessments (Gibbons et al., 2018). It is interesting to note, and is supported by other finding in this study, that instructors described creating MC items as being difficult. Particularly, several instructors, including Dr. Smith and Dr. Tracy, mentioned creating the distractors as being the biggest contributor to the difficulty level. The creation of effective MC assessment has been described as being a difficult skill to learn (Thorndike & Thorndike-Christ, 2010). Furthermore, chemistry instructors have been described as often feeling unfamiliar with assessment design (Bretz, 2012). Because of the difficult nature of MC assessment design and chemistry instructor's potential unfamiliarity with the topic, these findings support what is described in the literature.

4.5 Knowledge of Assessment Strategies

What assessment strategies do general chemistry instructors use or consider when assessing their students?

A science teachers knowledge of assessment strategies (KOAS) refers to the ways they assess their students in a particular course (Abell & Siegel, 2011; S Magnusson et al., 1999). These strategies can include their knowledge of what types of assessment items to use and their knowledge of how to design effective tests or test items.

In this study, which was focused on what chemistry instructors consider when creating MC exams, it was found that the coded instructors discussions fell into this category more than any other category of the STALM. This is not surprising though, as the interview protocol was focused on instructors' considerations during the exam creation process, a time when considerations of assessment strategies would be particularly important. Thus, it follows that KOAS codes are prominent.

Two main groups of codes were detected. General strategies and knowledge of exam and item properties. This is shown below in Figure 4.9. Substantial KOAS findings will now be presented and discussed.

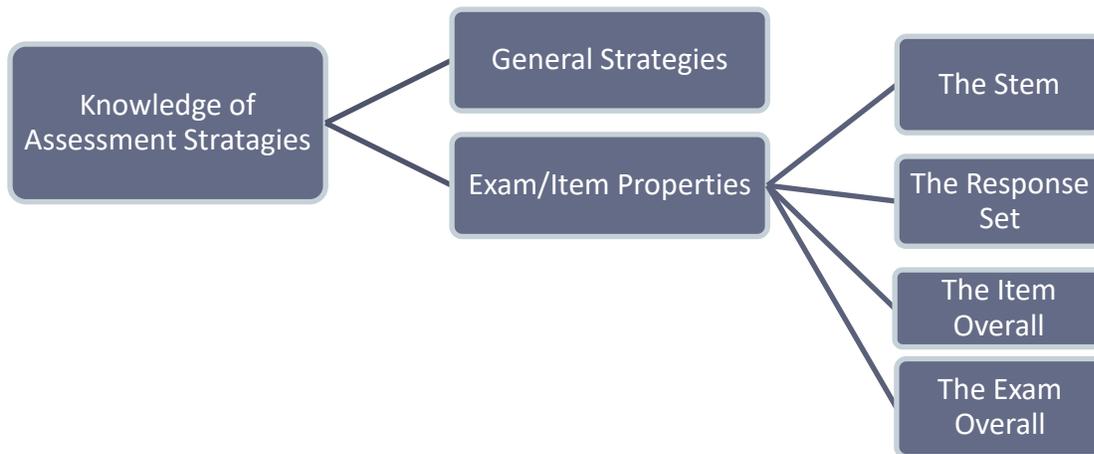


Figure 4.9. Knowledge of Assessment Strategies

4.5.1 General Strategies

General strategies encompass the techniques and considerations an instructor may use while designing assessments. Some prominent general strategies found in this study include: collaboration, using multiple types of assessment items, revising, trialing, and editing exams, perceived benefits and disadvantages of using MC assessment, and a lack of knowledge of MC assessment design.

4.5.1.1 Collaboration

Nearly all the participants mentioned working with others while developing MC exams for their students. Most participants did collaborate during assessment design however some worked independently. It is interesting to note that many worked together due to external factors such as teaching a team course and not necessarily due to personal desire. A few example quotes will now be given.

Dr. Irvine said:

Interviewer: As you're creating your exams, do you ever work with your colleagues on creating exams for your courses, or do you normally do that independently?

Dr. Irvine: The department has divided us into teaching groups, and then general chemistry teaching group meets together and we have created some standard questions that everybody has to give, so those ones are part of my question bank, and we report how the students do in each question, including what option they each selected. And so the person over the teaching group gathers the data and produces a report, and lets us know how our students are doing compared to the average, and still we can see if we need to change any teaching.

Furthermore, Dr. Deagen said when asked about making exams with colleagues:

So, in course X, which has just been around since 2010, I've been the only one teaching it so I haven't had to collaborate. Before that I taught course Y for a little while, and there we had to work together.

Additionally, Dr. Lopez said:

Sometimes here it depends on the semester I'm teaching and the colleagues I'm teaching with. Because if I teach general chemistry in the fall semesters, then I don't discuss with my colleagues because it's only one section and I'm in charge of the section. But if it's in the spring semester, we have to coordinate with the other sections, so we all need to agree on the questions. Sometimes the other lecturer may have disagreements with some of my questions, and I may have disagreements with his or her questions, so we need to fine tune and work together.

The above quotes show examples of instructors working together mainly due to course and departmental structure requirements. On the other hand, some of the instructors interviewed did collaborate out of personal desire. A couple example quotes will now be given below. Dr. Smith who had 2- years of teaching experience and had administered 5-9 MC exams mentioned sharing her exams with a colleague for feedback:

I write my exams myself. And then oftentimes I'll share my exams with a colleague afterwards to get their input.

Dr. Patrick, mentioned using examples from other colleagues:

As an instructor, I based the design of my multiple-choice questions on examples I got from Person X and Person Y, 'cause at that time I didn't have a whole lot of experience doing multiple choice questions.

4.5.1.1.1 Discussion

Working with others on exam creation is recommended in measurement textbooks because it can help the test creator to identify errors that may affect student performance (Thorndike & Thorndike-Christ, 2010). Many of the instructors interviewed in this research did collaborate, although many cited course structure as the reason for doing so. Although course structure requirements may promote collaboration within chemistry departments, collaborating based on a desire for higher test quality is the ideal. Additionally, collaboration has many perceived benefits including an increase of productivity, motivation, the effective use of limited resources, promotion of creativity, and the improvement of teaching quality (Austin & Baldwin, 1991).

4.5.1.2 Revise, Edit, and Trial

Many of the instructors discussed or used the strategy of revising, editing, or trialing exam items before giving them to students. Dr. Sorenson illustrates this by describing the test development process typically used in his or her department. Dr. Sorenson said:

Dr. Sorenson: So, we definitely make sure we are pretty rigorous about going through the questions. So after that we make a short list of questions and then Person X's office will go through and compile the questions and format them and stuff so when that is done then the following week, we'll get a rough draft and we'll go through and check it again. And then usually after that we usually have somebody take it. So sometimes Person X takes it, sometimes a TA takes it or something like that. Just double check all of our answers are alright. Then after that then it gets published.

Interviewer: So you usually have an expert take it? Like a TA or a professor?

Dr. Sorenson: Yeah, we definitely do not give a test without having a dry run because sometimes we find mistakes. Actually, usually we give the test and then we find mistakes and we have to fix it. So the more mistakes we can prevent the better because if you make a mistake with 2,000 people it's a huge pain to correct.

To further illustrate this, Dr. Crawford discussed during phase one of her interview refining and modifying test questions before giving them to students:

So, I went through and kind of deleted some [test questions] and added some different ones. Usually when I was doing that, I would take them from the test bank that came with the textbook, although they often need some modification because the test bank is not amazing, right?

Additionally, Dr. Tracy edited an item during phase three of the interview. This practice was seen with many of the participants.

Dr. Tracy: I'm not sure that I'm happy about the wording on what I wrote out. "Using the supplied table of electronegativities, which of the following compounds will show the highest character of ionic bonding?" Probably I would reword that a little bit.

Interviewer: Okay. Can you tell me why you would reword it? What about the question would you like to change?

Dr. Tracy: I don't like the word character in there. Let's see. I'll cross that out and then I'm going to change that to "would exhibit ionic bonding."

4.5.1.2.1 Discussion

Revising, editing, and trialing exams is a recommended practice (Pellegrino, 2001) and is seen in the design process of ACS exams (Eubanks & Eubanks, 1995; Holme, 2003). Additionally, many chemistry departments have assessment plans that utilize a feedback loop of assessing students, analyzing assessment data, and then modifying assessments based on those results (Towns, 2010). The data in this research shows that revising exams and exam items is an important

consideration of many of the participants in this study. This fact that chemistry instructors are revising, editing, and trialing their exams is a promising finding that could be explored in further research and promoted among general chemistry instructors.

4.5.1.3 Using multiple types of assessment items

Many of the instructors interviewed used multiple types of assessment items in an exam. For example, using multiple-choice and short answer items in the same exam. This is illustrated by a quote from Dr. Tracy:

My exams always have about 40-50% of the points available as multiple choice questions. And so, every general chemistry exam that I've done has had between 15 and 25 multiple choice questions on it.

Furthermore, Dr. Johnson said:

Yes. And so I make, every quarter I make 3 quizzes, 2 midterms, and 1 final and all of them have multiple choice in them. And so my quizzes are all multiple-choice and my midterm is 50% multiple-choice, 50% short answer, and my final is about 40% multiple-choice, 60% short answer. So, all of mine have multiple-choice in them.

Additionally, Dr. Crawford said:

In terms of the exams, in Course X and Course Y, we usually have three or four midterms and then a final exam. Each of the midterms is usually split about half and half with multiple choice and free response. So, a typical exam will have 20 multiple-choice questions and then three or four or five free response, on the order of four free response questions as well. And then the final exam is usually all multiple-choice and it might be something like 50 questions or something like that.

Lastly, Dr. Tracy said about item five in phase two of the interview:

Dr. Tracy: And that's another thing that's a little bit tricky. There's an additional step here. You've got to figure out which one would be completely neutralized, so you're going to calculate something like 183 milliliters, and then, okay, 150 milliliters would be completely neutralize, 200 wouldn't.

And that's one additional step, so this is testing a lot more concepts than I would like to see for a multiple choice question. I like to do this sort of thing in short answer, but I like to kind of break it up into a few different steps.

Interviewer: Okay. The reason you would prefer to do this short answer versus multiple choice, what would be your reasonings there?

Dr. Tracy: If you calculate the formula mass for calcium carbonate wrong, just a calculator mistake, and then spend five more minutes figuring this problem out, you're sunk. I think at the level that I teach, students would have a very difficult time navigating something like this, and breaking it into a couple steps is helpful.

4.5.1.3.1 Discussion

Different types of assessment items (multiple-choice, short answer, essay) can be used to assess different levels of cognitive skills and each item type have distinct advantages and disadvantages (Jacobs & Chase, 1992). Because of this, it is important for an instructor to consider the types of assessment items he or she will use while designing an exam (Jacobs & Chase, 1992). The data from this study shows that many of the chemistry instructors interviewed did use multiple types of assessment items. Dr. Tracy for example preferred item five to be written as a free-response in order to capture more of the students thinking. Using multiple types of assessment items in general chemistry exams is a practice that should be encouraged and promoted throughout chemistry departments.

4.5.1.4 Benefits and Disadvantages of MC Assessment

When it comes to the assessment strategy of using MC assessment items, many of the instructors mentioned various advantages and disadvantages of using MC assessment. These advantages and disadvantages fell into three main categories, namely, MC assessments can be used to assess many concepts quickly (advantage), MC assessments are easy to grade (advantage), and

MC items is difficult and time consuming to create (disadvantage). Example quotes for each category will now be given.

To demonstrate the view that MC assessments can be used to assess many concepts quickly, a quote from Dr. Johnson will be presented:

So, our exams are usually 50 minutes or so. And because of that, it's hard to hit as many topics as I want to. And I find that multiple choice is a really nice way to get directly to a particular concept and not have it a bunch of time for the students. So, if I have certain topics I want to make sure I address, I find that multiple choice is a really nice way for me to make sure I hit all of the points that I wanted.

The ease of grading MC assessments was mentioned by several participants. A quote from Dr. Crawford illustrates this point. Dr. Crawford said:

Interviewer: Yeah, certainly. That's great. Thank you for sharing that. You mentioned that your exams tend to be half multiple choice, half write on during the semester, and then the final exam is typically all multiple choice. Could you tell me a little bit about your reasonings behind that?

Dr. Crawford: That's sort of what people do here, so I think that's the main reason. But also it's really hard, our final exams. A lot of times the TAs will leave and so it would be, with our classes, with 250 people, it would be impossible to grade a free response exam right at the end of the semester like that. So we mostly do it to make it easier because of grading.

One disadvantage of MC assessment that was discussed by many of the participants was the difficult and time-consuming nature of creating them. Dr. Tracy exemplified this when he said:

As I said, I find the multiple choice to be the most difficult to write myself, and I try not to do it. I try to use test banks because there's a lot of problem types where I don't feel I can come up with the distractors well. Or maybe where I'm too lazy to sit there and try to think of all the mistakes that they're likely to make and come up with them. That might be a better way of putting it. For stoichiometry problems, I almost never write them myself. I use them out of banks. They're straightforward problems to write. It's coming up with the five choices that's the challenge.

4.5.1.4.1 Discussion

Multiple-choice exams are known for having advantages and disadvantages such as being able to be used to test a wide variety of content, being easy to grade, being difficult to construct and being susceptible to guessing or the use of test taking strategies (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005; Jacobs & Chase, 1992). Many of the instructors interviewed were cognizant of the advantages and disadvantages of using MC assessment.

4.5.1.5 Lack of Knowledge of Multiple-Choice Assessment Design

While the instructors interviewed considered many correct aspects of assessment design, there was a demonstrated lack of knowledge present among many of the participants. To demonstrate this several excerpts will be presented. First, Dr. Sorenson had several occasions where a lack of understanding was present during her interviews. The following two excerpts are from Dr. Sorenson.

Dr. Sorenson: ‘Which of the following cannot be determined based on the provided graph?’ I’ve been told by people I teach with that you’re not supposed to ask negative questions so I guess it would be good to try to frame it more positively or whatever but obviously that changes the distractors you can use.

Interviewer: Do you have any reasoning behind why you have been told that about using negative questions?

Dr. Sorenson: No, because I teach with a lot of chem ed people and that’s what they tell me. No, I don’t, because my teachers used to write questions like that all the time. So, I guess I don’t really know why.

Furthermore, Dr. Sorenson said while trying to come up with a distractor for an item created during phase three of the interview:

Dr. Sorenson: So, I need one more distractor. And usually when I get to this point, I start running out of distractors and I don’t really know what to do.

Interviewer: You said that you needed one more distractor after you came up with the four. Can you tell me a little bit about why? Why you would need one more?

Dr. Sorenson: We always write questions with five answers. Is there some scientific study or something that talks about the difference between four answers versus five answers and it's effectiveness? I don't know why we do that. It's just that we do. Maybe because they have more of a chance of, it's just harder right if you have one extra choice. Statistically speaking, it's more difficult. So, we always write them with five.

At the end of phase one of his interview, Dr. Bennett, expressed that he was still learning how to create MC exams so he wasn't completely comfortable with it yet.

Interviewer: Okay. Is there anything else you would like to share about, in general what you think about or consider when you're creating an exam?

Dr. Bennett: Um, I'm not sure. I mean I'm still learning how to do it at this level. It's not easy for me yet to write a multiple-choice question. So, a lot of teaching in Course Y was where I learned a lot of it because of Person X and Person Y really taught me a lot about... When I started out my questions were 10 times too hard. I wrote questions that were too complex and too hard, and I really had to learn how to step back so I don't... So, I'm still in that learning process so I don't know that I can give you more about it.

Dr. Bennett goes on to say while choosing distractors for an item he created during phase three of the interview:

Dr. Bennett: Then, I'm also not sure how I feel about 'none of the above' and 'all of the above'. Sometimes I include that sometimes I don't.

Interviewer: What do you mean by you're not sure how you feel about them?

Dr. Bennett: Well I don't really know... I just don't know. Actually, I don't know in terms of... Maybe you can tell me. In terms of its sort of value as a choice on a question. Is it a cop out? Is it good? I would never have it on all of them but if you have it on occasionally is that okay? Or is it really just not a good thing to put on there. I don't really know. I don't think I've ever asked that before.

Dr. Lopez demonstrated a lack of understanding while determining the order of the answer choices in an item created during phase three of the interview.

Dr. Lopez: Slightly less than one mole of water. Then A, B, C, D. What is slightly less than one mole of water? I'd move it to A. Slightly less than one mole of water.

Interviewer: Why did you decide to move it to A?

Dr. Lopez: I like to put them in order.

Interviewer: Okay.

Dr. Lopez: Don't ask me why.

Dr. Brown who had over 10 years of general chemistry teaching experience and had administered over 20 MC exams said about the number of answer choices typically included in a MC item:

Dr. Brown: And I usually do four, so like A, B, C, D. I occasionally do more than that. Usually four answers, sometimes five.

Interviewer: Okay. Do you have any reasoning for why you typically choose four?

Dr. Brown: It's what I've always done.

4.5.1.5.1 Discussion

Having the data from this research reveal a lack of understanding of assessment design is supported by the literature that has investigated college level chemistry instructors. Chemistry instructors at the college level tend to receive little to no training in appropriate pedagogical or assessment practices (Lawrie et al., 2018). Furthermore, in a survey of chemistry faculty about their familiarity with assessment terminology, it was found that although faculty were familiar with many assessment related terms, holes in their understanding were present (Raker, Emenike,

et al., 2013; Raker & Holme, 2014) Additionally, Bretz has commented that that assessment of student learning is often a difficult and unknown process for many chemistry instructors (Bretz, 2012). This deficiency in formal training and knowledge of assessment terminology among chemistry instructors may help explain why a lack of understanding of assessment design principles is present in these data. Further research is recommended.

Dr. Sorenson, Dr. Bennett, and Dr. Lopez all expressed a lack of understanding of the reasonings behind typical item writing guidelines, such as avoiding negatively phrased items, avoiding ‘all of the above’, and putting answer choices in a logical order (Haladyna et al., 2010). Although they were aware of the practices, they were still unsure about the underlying purposes guiding those practices.

Another common area where the instructors demonstrated a lack of understanding of assessment design was in the number of answer choices to include in MC items. Dr. Sorenson and Dr. Brown both expressed that they typically include five answer choices in their MC items. When asked why, they did not have appropriate reasonings for their actions. Dr. Sorenson and Brown both cited tradition as why they choose to have five answer choices even when they were struggling to come up with a viable fifth option. This practice was seen among many of the instructors interviewed where tradition or a desire to have uniform items with the same number of answer choices guided their decision of how many answer choices to include in an item.

The literature describes the reasonings that should guide these MC assessment design practices. Negatively worded items are advised to be used with caution (Haladyna, Downing, & Rodriguez, 2002a) because they can introduce error that can increase item difficult (Cassels & Johnstone, 1984; Dudycha & Carpenter, 1973). ‘All of the above’ is to be avoided as an answer choice because it can enhance student performance due to a cuing effect (Harasym, Price, Brant,

Violato, & Lorscheider, 1992). When determining the number of answer choices to put in an item, it is recommended to include only as many distractors as are plausible because extra distractors do not improve the psychometric properties of the item (Papenberg & Musch, 2017).

The expression of a lack of understanding of MC assessment design is not surprising among those interviewed. However, it does highlight an area of focus for future research and professional development opportunities.

4.5.2 Item and Exam Properties

Under the category of Knowledge of Assessment Strategies, many codes fell into a group characterized by the term Item and Exam Properties. These codes relate to an instructor's knowledge of the various aspects of an effective MC item or exam. There are four sub-categories within this group, namely, the stem, the response set, the item overall, and the exam overall. This is shown in Figure 4.9.

4.5.2.1 The stem

Two considerations of assessment design were discussed in relation to the stem of MC items. Those interviewed mentioned the use of negative phrasing and the completeness of the problem statement during their interviews.

4.5.2.1.1 Negative phrasing

Within the discussions of negative phrasing, three distinct views were present, those in support of, those opposed to, and those not sure of the practice. An example will now be given of each view.

While evaluating item 6 (see Figure 3.3 in section 3.6.2) during phase two of the interview Dr. Smith expressed support of using negative phrasing in the context of the item. Dr Smith said:

But I think also what's nice about this question in particular is that it's specifically asking you what you can't determine from the graph. So, it's like hey, we can use graphs for all sorts of things, but what can't you use this graph for. So, I think that's a really, really nice spin on a graphical question.

Conversely, Dr. Sanders, while evaluating the same item (see Figure 3.3 in section 3.6.2), expressed being opposed to the practice of including negative phrasing in an item. Dr. Sanders said:

I think this question ought to be which of the following can be determined based on the provided graph? Then it's more positive. It says what can we figure out based on what I can see here, based on the limits of this axis? That I think is a better approach to this question than saying which things can't we figure out. Turning it around I think makes it a much better question.

The third view of negative phrasing was one of uncertainty about the reasoning behind the practice. Dr. Sorenson, who advocated for removing the negative phrase, was uncertain about why removing the negative phrase would be desirable. Dr. Sorenson said:

Dr. Sorenson: 'Which of the following cannot be determined based on the provided graph?' I've been told by people I teach with that you're not supposed to ask negative questions so I guess it would be good to try to frame it more positively or whatever but obviously that changes the distractors you can use.

Interviewer: Do you have any reasoning behind why you have been told that about using negative questions?

Dr. Sorenson: No, because I teach with a lot of chem ed people and that's what they tell me. No, I don't, because my teachers used to write questions like that all the time. So, I guess I don't really know why.

4.5.2.1.1.1 Discussion

The inclusion of negative phrasing in an item is only recommended to be used cautiously and if used it is important to ensure the negative phrase is bolded or capitalized (Downing, Haladyna,

& Rodriguez, 2010). In study of negative phrasing in chemistry exams, it was found that including a negative phrase increased the item difficulty (Cassels & Johnstone, 1984). Additionally, Tamir found that for items that require high cognitive reasoning (like the graphical interpretation question is this study) negative phrasing makes items more difficult (Tamir, 1993).

It is interesting to note the variety of views on negative phrasing found in this study. Dr. Smith was in support of negative phrasing in the context of this item if the learning objective was to assess student's ability to recognize what can not be learned from a graph. As it says in the literature, use of negative phrasing is to be used cautiously and in some situations its use may be justified. On the other hand, Dr Sanders recognized that phrasing the question positively may improve its quality. In the context of this item, (see Figure 3.3 in section 3.6.2) the negative phrase may add additional cognitive load to an already complex item. Furthermore, although Dr. Sorenson mentioned the strategy of avoiding negative phrasing, she did not understand the reasoning for the practice. The variety of views shown in the data of this research may indicate that chemistry instructors' views on negative phrasing in item stems fluctuate significantly.

4.5.2.1.2 Complete problem statements

Some of the instructors interviewed considered the completeness of the problem statement in MC items. Those who mentioned this were all in support of complete problem statements. Many of the comments about this strategy came while evaluating the fourth item in phase two of the interview (See Figure 3.3 in section 3.6.2). Dr. Bennett discussed how making the problem statement complete would improve the item:

So, while I can go through it, they have to read and comprehend every single solution so that's what sort of Person X and Person Y pointed out that that can be very challenging especially under a stressful situation for a student because they have to process each one and discern it as if you had something where the question is more elaborated and the solutions are more simplified. That's better.

Furthermore, Dr. Johnson discussed how items with incomplete statements can be demotivating to students:

You don't even know what the question is asking because it's just one word, it says electro negativity and then a bunch of statements. It's a bit de-motivating to students to see questions like this and they would just pick the first thing that seems right without actually reading carefully.

Additionally, Dr. Patrick spoke about how complete problem statements is a desired practice and then expressed uncertainty in the practice for this specific item:

Dr. Patrick: "Electronegativity ... " So, what I learned is that's not good practice. If you just do one statement and then you let them select a lot of different statements, that's not good. You should rather be very descriptive in your question, in your stem question that you provide, and then look for right answers.

Interviewer: Can you-

Dr. Patrick: This is almost like a fill-in-the-blank question, right?

Interviewer: Could you tell me a little more about why that's something that you think?

Dr. Patrick: Well, in this case it's not that tricky because it's only one word, right? So you can easily just read that through. Maybe I'm wrong in this case. Maybe I'm wrong. See, I wouldn't know in this particular case exactly. Because it is actually, the more I read it, it's actually pretty clear.

Interviewer: Okay.

Dr. Patrick: "Generally increases left to right ... " Yeah, it's a ... but I think what I would have liked for me as a student if I would imagine myself in the role of a student, I would like to have a phrase here like, "Electronegativity does show the follow trends. With which do you agree?" It's just more of a coherent sentence.

Lastly, Dr. Crawford, who had extensive experience in educational measurement, discussed the strategy of using complete problem statements and its reasonings. Dr. Crawford said:

Okay. Uh huh, okay. So, the first rule of writing a multiple-choice question is that the question's stem should be an actual question that has an answer. And so we haven't met that rule here. A second rule is that as much of the words that need to be in the, like as many of the words as possible should be in the stem, and not the response options because it makes it so much easier for students to read them in the stem and then have smaller things to read as the choices because it's easier to distinguish between those choices when they're smaller and easier to understand. So, this question has not met those requirements.

4.5.2.1.2.1 Discussion

The fact that chemistry instructors were aware and in support of using complete problem statements in the stems of MC items is in-line with the literature on best practice (Towns, 2014). Having complete statements reduces the cognitive load placed on students, tends to decrease answering time, and increases the number of items that can be included in an exam (Downing et al., 2010; Dudycha & Carpenter, 1973).

Drs. Bennett, Johnson, Patrick, and Crawford all made comments about making the items easier to understand for their students as reasonings for this practice. Considerations of the student testing experience is promising to see among the chemistry instructors interviewed.

4.5.2.2 The response set

All the chemistry instructors, to some extent, discussed the response set in their interviews. The response set is the array of answer choices that a test-taker can select from. The presence of a response set is what makes a MC item unique from a short answer item. While discussing the response set, the chemistry instructors mainly talked about distractors, using all/none of the above, symmetry, answer choice length, and answer choice order. Substantial results will now be presented and discussed.

4.5.2.2.1 Distractors

While discussing distractors, the chemistry instructors focused on the number of distractors to use and making distractors from common student errors or misconceptions. Several instructors also noted that creating the distractors is the most difficult part of MC assessment design. Quotes and examples will now be given.

Regarding the number of distractors to include in a MC item, the instructors interviewed fell into two groups. Those who tended to include a set number of answer choices (typically 4 or 5) and those who included only plausible answer choices. Dr. Johnson demonstrates the practice of including a set number of distractors while creating an item during phase three of the interview. In the quote, Dr. Johnson determines the number of answer choices (four) before considering how many distractors could actually be plausible. The quote is shown below, and the item is shown in Figure 4.10.

And so, okay, let me write the four options. I think I got it.

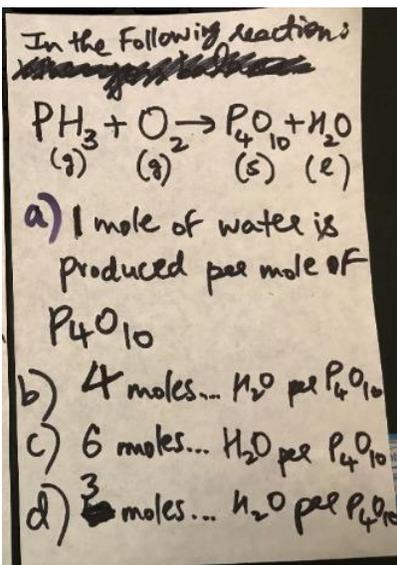


Figure 4.10. Created item during phase three of Dr. Johnson's interview

Furthermore, Dr. Tracy decided to include five answer choices in the item he created during phase three of his interview. When asked about why, he said:

Interviewer: Okay. You said that you would like to or need to come up with two more. Could you tell me about why you feel you would need to come up with two more after-

Dr. Tracy: Our Scantrons have five circles, and so I always just do A through E.

Interviewer: Okay.

Dr. Tracy: You're right. There's no reason I couldn't just have A through D, but I've never done that before.

Lastly, Dr. Irvine discussed why he choose to always include five answer choices in MC items. He said:

Interviewer: And you mentioned that you would choose to have five different answer choices. Could you tell me why you would choose to have five?

Dr. Irvine: I've always liked five because I felt that four is just too easy, and more than five can be difficult for some people to kind of retain the differences. I believe on the processing space theory, where people say, "Well, you can only hold between five and seven things in your head at a time." So I like to stay safe, right, and I think that the chances of getting it right by guessing are just a little lower.

Dr. Irvine: We also have a restriction from the testing center. Their software only allows five. So, they can deal with more, but they really just prefer five. They've been that for a number of years.

The other group of comments about the number of answer choices to include in a MC item focused more on the plausibility of the answer choices and less on the overall number of answer choices. Dr. Crawford typified this strategy when discussing the first item in the second phase of the interview (see Figure 3.3 in section 3.6.2) by saying:

Most questions, the vast majority of questions I've been limiting to four answers, not five. Obviously that increases the probability that somebody's going to guess,

but usually you can't come up with a fifth answer that's a good distractor which is like what's happened here. They've put in three 630s, right? There's no point to adding those, they're not helping you discriminate between people who know this and don't know it.

Additionally, Dr. Sanders explained that he does not always include the same number of answer choices in his MC items. Dr. Sanders said:

Dr. Sanders: I don't feel it's necessary to have five distractors for every question. Sometimes four is better actually.

Interviewer: Can you tell me why you feel that way?

Dr. Sanders: Question one or two was it? For example, all of the above on question number two. The question is really asking which one of these quantum numbers and I think all of the above is wasteful. It doesn't need to be there. I think the students won't choose it because the question really is kind of asking which one.

Another aspect of MC distractors that was discussed by the instructors was that they typically use common student errors or misconceptions as distractors. Dr. Sorenson demonstrated this during phase three of the interview when he or she said:

So, the first thing I'll do is I'll work out the answer and then what I'll do is I'll go back through my math and I'll pretend that I started making mistakes. So first thing I'll do is go back to my stoichiometric ratio step and where I'm figuring out where the limiting reagent is. So, I'll assume I forgot to use the ratio where I balance the question or I balance the equation wrong. So, one of my distractors will definitely come from that. Then what I'll do is I'll make a mistake like either in one of these calculations where I am determining the moles of something and another distractor will come from that.

Furthermore, Dr. Crawford discussed using common student errors as distractors in MC items. Dr. Crawford said:

Sometimes when I've written new questions like that especially, I'll have my TAs do them, like as a free response question and say these are all the things that you could do wrong, so these are some good distractors and stuff like that.

Lastly, Dr. Deagen said:

Dr. Deagen: I pretty much know a lot of what their misconceptions are and I can build those into the multiple choice distractors.

Interviewer: Okay, great. So, you consider their misconceptions when you're making the questions?

Dr. Deagen: Oh, absolutely.

In addition to the number of distractors in a MC item and using common errors as distractors, several instructors mentioned how creating the distractors was the most difficult part of MC assessment design. Dr. Tracy said:

I think the most difficult thing about writing a good multiple-choice question is coming up with the distractors.

Furthermore, Dr. Smith discussed the creation of a MC item that she was proud of creating.

She said:

Dr. Smith: That's probably the proudest multiple choice question, the thing I'm most proud of because I made it pretty much entirely from scratch. So I was like, oh but it was really hard.

Interviewer: Yeah. Can you tell me why it's difficult? I know you mentioned it before. But what was it that made creating them difficult?

Dr. Smith: Just because I- coming up with four good distractors was really, really tricky.

Additionally, Dr. Sanders mentioned that creating distractors can be time consuming and difficult to do. He said:

Getting distractors written correctly oftentimes takes some time because you want them to be good distractors but if they're not written well, and I'm thinking about responses that are actually text not numbers, if they're not written well, they [students] can pick up on the fact that that one doesn't make any sense.

4.5.2.2.1.1 Discussion

The design and use of distractors has been extensively studied in the field of educational research (Gierl et al., 2017).

When it comes to the number of distractors to use, it is recommended that test developers only include as many distractors as are plausible (Haladyna & Downing, 1989). Studies have shown that two distractors and the correct response is typically sufficient in most testing situations. (Rodriguez, 2005; Tarrant & Ware, 2010; Tarrant, Ware, & Mohammed, 2009). It has also been shown that students answer 3 option MC items on average 5 seconds faster than 4 or 5 option items (Schneid, Armour, Park, Yudkowsky, & Bordage, 2014). This can allow for more items to be included in an exam which can increase content coverage and exam reliability (Thorndike & Thorndike-Christ, 2010).

Although results were varied from the chemistry instructors interviewed, it is concerning to note that many of the instructors opt for including a set number of answer choices in a MC item over only as many as are plausible. For example, Dr. Tracy exemplified this when he said, “Our Scantrons have five circles, and so I always just do A through E.” The practice of always including four or five answer choices in a MC item is not in-line with literature recommendations and should be addressed by further research and training.

The design of effective distractors for MC items typically follows two strategies, namely, creating distractors based on similarity or based on student misconceptions (Gierl et al., 2017). Therefore, observing many of the chemistry instructors discuss and practice using student misconceptions to design MC distractors is promising. This practice should be encouraged and promoted among those designing MC chemistry examinations.

Lastly, with several instructors mentioning the creation of distractors as being one of the most difficult parts of MC assessment design, it may be useful to focus future professional development and research efforts around easing this difficulty for chemistry instructors.

4.5.2.2.2 All of the above

Most of the instructors interviewed discussed the inclusion of ‘all of the above’ as an answer choice in a MC item. Most were opposed to its use with some unsure about its value as an answer choice.

Dr. Deagen discussed the reasoning behind avoiding ‘all of the above’ in a MC item:

Dr. Deagen: Yeah, I mean I go to it, but I don't really like to, is using the all the above, none of the above, because they [students] can usually eliminate those pretty fast.

Interviewer: Okay.

Dr. Deagen: Or if they know two of the three are right, they'll know that it's all of the above. I tend to stay away from those too just to circumvent their test taking strategies.

Additionally, Dr. Patrick further emphasized that students can exploit ‘all of the above’ as an answer choice. Dr. Patrick said:

Dr. Patrick: I also do not like to include questions such as like ... that say, "All the answers are right. Above all." Right? So like, I have A, B, C as the correct answers and then D is "Above all is also correct."

Interviewer: Okay.

Dr. Patrick: Like, "All of the above are correct." I try to not do that, because test-takers-

Interviewer: Yeah, can you tell me why.

Dr. Patrick: Test-takers who are really good in taking multiple choice questions, they can select the right answer based off that last choice. So, they can say, "Oh, okay. I'm not so sure about C, but then if she says that, then that might be an option, so maybe A, B, C is right because of

that." So, it's hard and I don't think I'm really successful at this yet, but I'm aware of the fact that people who are trained to take tests like these and that I think is a big disadvantage of multiple-choice questions. You can get trained, even though you do not have the necessary knowledge pool. You can get trained to get a higher score in these tests if the questions have a certain bias in them.

Furthermore, some of the instructors were unsure about using 'all of the above' as an answer choice. Dr Bennett expressed uncertainty regarding 'all of the above' when he said:

Dr. Bennett: Then, I'm also not sure how I feel about 'none of the above' and 'all of the above'. Sometimes I include that sometimes I don't.

Interviewer: What do you mean by you're not sure how you feel about them?

Dr. Bennett: Well I don't really know... I just don't know. Actually, I don't know in terms of... Maybe you can tell me. In terms of its sort of value as a choice on a question. Is it a cop out? Is it good? I would never have it on all of them but if you have it on occasionally is that okay? Or is it really just not a good thing to put on there. I don't really know. I don't think I've ever asked that before.

Moreover, Dr. Brown, while discussing item two during phase two of the interview (see Figure 3.3 in section 3.6.2), expressed uncertainty in using 'all of the above' as an answer choice. He ultimately decided its inclusion would be appropriate if one was trying to have five answer choices for each item in the exam.

I don't think I have anything else to say. Well, question two, the "all of the above", I guess if you're trying to have an A, B, C, D, E for each one... Yeah, I think that would be okay.

4.5.2.2.2.1 Discussion

The inclusion of 'all of the above' as an answer choice in the response set is not recommended as it can be easily eliminated or selected by students (Harasym, Leong, Violato, Brant, & Lorscheider, 1998). Although some instructors were unsure about its practice, most who discussed 'all of the above' were opposed to its practice and could cite the appropriate reasoning why it

would not be desired as an answer choice. Seeing some of the chemistry instructors follow this item writing guideline and discuss the reasoning behind it offers promise that training on other item writing guidelines may be well received. With this said, in future assessment focused professional development opportunities, it would be recommended to still emphasize the guidelines around ‘all of the above’ as some of the instructors interviewed expressed uncertainty about its practice.

4.5.2.2.3 Answer choice length and order

Many of the instructors discussed the respective length of the answer choices and the order they prefer to put the answer choices into when designing MC items. In regard to the length of answer choices, many instructors desired them to be approximately the same length to prevent a cuing effect. Dr. Tracy exemplified this view when he said about the length of the answer choices in item 4 during the second phase of the interview (see Figure 3.3 in section 3.6.2):

Dr. Tracy: Yeah. I like the question overall. I don't like how the correct answer is a little bit longer and more specific than the others one. It kind of stands out a little bit. I probably would have tried to make it a little bit shorter and a little bit more similar to the others in what it says.

Interviewer: Why is it a concern for you if one of the answers is longer than the others?

Dr. Tracy: I don't want them to be able to kind of look at the question and figure it out using multiple choice skills. I want them to know the concept.

Furthermore, Dr. Smith prefers to have answer choices be approximately the same length to prevent students from using test-taking strategies to determine the correct answer.

Dr. Smith: And then question four has the very common problem where the right answer is the one that's the longest. And I'm like well, students are gonna see that and be like well this has a lot of good information in it, it's probably the right answer, right? Electronegativity's not

gonna be summed up in three words, so I'm probably gonna pick the thing that's the longest.

Interviewer: Have you seen that with your students?

Dr. Smith: That's a really good point. I don't know where that feeling comes from. Maybe it comes from my own personal [experiences] when I took multiple choice tests. That the answer that seems to be the most robust is usually the answer that is correct because correct answers usually have a lot of subtleties involved. And if you go back to my very long worded problem that I just sent you that I wrote, the correct answer is it's not the longest, it's not the shortest. I wanted to make sure that it wasn't the longest answer.

Regarding the order to arrange answer choices, many of the chemistry instructors interviewed preferred to put the choices in a logical order as opposed to a random order. Dr. Smith discussed this when commenting on item one during the second phase of the interview (see Figure 3.3 in section 3.6.2):

I don't like the answers. They're all written in the wrong order. This increases cognitive load without actually improving the... it just makes people get the question wrong because they're confused about the way you've written the answer as opposed to because they don't know the answer. So they should be written, if they're going to be written, in order.

Furthermore, Dr. Madison discussed arranging answer choices in a logical order while evaluating item four during the second phase of the interview (see Figure 3.3 in section 3.6.2):

And again, maybe I would order them as n, l, m_l, m_s, only because that's the order that I would teach them in, just to kind of keep them ... You know, you gotta keep their minds thinking the way they saw things. I mean, it's a little thing, but it's just probably what I would do

Moreover, Dr. Brown explained the reasoning behind why he prefers to arrange answer choices in order. He said:

Dr. Brown: And then another thing I would probably do, and this is just me, is I would probably put A, B, C, D, E in order, like A would be N, B would be L, C would be M₁ then D would be M_s.

Interviewer: Okay. Do you know why you, why do you prefer to do it that way?

Dr. Brown: Two reasons. One reason is, I like things in the right order, and the second reason is, and I don't know the literature to back this up, but I need to find the literature to back this up, maybe you know the literature and you can tell me. But I have heard that if the answers are in a more readable format, then the discrimination of the question is better.

Dr. Brown: So randomizing answers doesn't necessarily make it a better question. And so even if I went back up to question one, if I was writing it, I would write, like A, B, C, D, E, A would be the lowest value, so .64, B would be the next lowest, and it'd go from lowest to highest value.

4.5.2.2.3.1 Discussion

Several of the chemistry instructors interviewed did discuss the importance of the length and order of answer choices. Observing the use of these two strategies further shows that chemistry instructors are considering appropriate assessment design strategies when designing MC exams.

This is particularly important with these two strategies because it is recommended that answer choices be kept to approximately the same length and be placed in a logical order (Haladyna et al., 2010). Keeping answer choices to approximately the same length may help to prevent student from using the “choose the longest answer heuristic” when answering a MC item (Towns, 2014). Arranging answer choices in a logical order is also recommended for aesthetic reasons and because illogical ordering may affect lower ability student’s exam performance (Huntley & Welch, 1993).

4.5.2.3 The item overall

Assessment strategies that focus on the entire item (the stem and response set) include considerations such as item clarity, item conciseness, significant figures and units, and formatting consistency. Substantial results will now be presented and discussed.

4.5.2.3.1 Item clarity

All the instructors interviewed discussed how easily an item would be for students to interpret. Item clarity was certainly a large consideration of those interviewed with all instructors taking measures to ensure easy interpretation of exam items for their students. Example quotes will now be presented and discussed.

Dr. Brown expressed his desire to have clearly written items while he evaluated item one in the second phase of the interview (see Figure 3.3 in section 3.6.2). He thought that the item was written unclearly and might confuse students. He said:

Dr. Brown: Okay. So, for question one, the wording of the question seems a little weird to me. I think it could be worded... So, "What is the molarity of 35 milliliter solution of nine molar diluted to .5 liters?"

Dr. Brown: I think it could be asked a little differently. Like, "What would be the new molarity?" or "What would be the molarity..." Start with the nine molar and the .5 liters... I don't know.

Interviewer: Okay.

Dr. Brown: Wait, which number goes where? (laughter)

Interviewer: Yeah. Why would you be considering changing the wording of the question?

Dr. Brown: I think it might confuse the students. So, when I write questions, I try, I've learned if I don't make them as clear as I possibly can, then students often have a way to come back and complain about them. So I've learned what students complain about and I don't like that.

Dr. Brown: I just am really careful with the way I word things.

Additionally, Dr. Crawford thought that the second item evaluated during phase two of the interview (see Figure 3.3 in section 3.6.2) was unclear due to the wording of the item's stem. Dr. Crawford said:

Hm... Okay. Uh huh. So, this is like somebody who's trying to avoid using the word orbital for some reason that I don't know. I don't know why they're trying to avoid using the word orbital. This question would be better if it said, "Which quantum number effectively describes the shape of an orbital?" So, they've used this sort of convoluted language to say the same thing. And maybe it's because there's another question on the exam where they're going to ask somebody what is an orbital, so they don't want to give that away here. But I just think it's unclear so it would be better. It took me a couple times reading it to understand what the question was saying. So, I think the question's stem could improve there.

Moreover, Dr. Patrick noticed confusing aspects of the first item evaluated during phase two of the interview (see Figure 3.3 in section 3.6.2). Dr. Patrick noticed that students would need to decode numerous pieces of information as well as interpret unclear answer choices. Dr. Patrick said:

The students that read that question need to decode a lot of information. First of all, they need to know what H_2SO_4 is. They need to know that it's like an ... it can split into ions when it's dissolved in water and they need to know, okay, difference between milliliters and liters. How do I convert that? I cannot just plug that into my equation. I do not like when questions ... here, D, "630 dot." Great. Is there a zero behind that? Right? What does that mean? Is it the end of the sentence? I mean, yes, we can assume what it means but the students are confused by that for sure.

4.5.2.3.1.1 Discussion

The clarity of assessment items was a front-runner regarding what the chemistry instructors consider while creating assessments. All instructors interviewed desired their MC items to be clear and easy to understand. This common sense practice is a well-supported item writing guideline (Haladyna, Downing, & Rodriguez, 2002b) and adhering to its practice improves item validity. It

has been shown that even small changes in item wording (clarity) can have large impacts on students performance in general chemistry assessments (Cassels & Johnstone, 1984; Schurmeier, Atwood, Shepler, & Lautenschlager, 2010). With item clarity being a large consideration of the chemistry instructors interviewed, it shows that the instructors are thinking about their students and viewing assessment items from their student's perspective. This practice should be continued and encouraged among general chemistry instructors.

4.5.2.3.2 Item Conciseness (amount of information)

Many of the instructors considered the amount of information given in an item during the evaluation and design process. Instructors preferred to include enough information for students to be successful without providing so much as to confuse or distract students from the task at hand. Example quotes will now be presented and discussed. Dr. Deagen expressed that the fifth item in phase two of the interview (see Figure 3.3 in section 3.6.2) didn't have enough information for students to be successful. This was an example of where including more information in an item was desired.

Dr. Deagen: Number five makes me a little nuts.

Interviewer: Can you tell me why?

Dr. Deagen: Because first of all, what's chalk? "What's the molecular weight of chalk?" There's not enough information to answer the question in my estimation. That'd be one of those I would throw out.

Furthermore, Dr. Patrick expressed the view that keeping items simple and concise was important for assessment design. Dr. Patrick discussed some criteria for how she evaluates her MC items. She said:

When I've done that, I would then go back and see, okay, is the ... simple criteria. Is the question itself that I'm posing, the stem question, is it coherent? Is it clear?

Does it have unnecessary information in there? So I personally think, and it's based on research, right? And you probably know this really well. That, if you include unnecessary information it just confuses students and it does really not help the understanding or the targeted approach for each of the questions.

4.5.2.3.2.1 Discussion

Seeing the instructors consider the conciseness and amount of information given in an item is promising because item conciseness is an important part of MC assessment design (Towns, 2014). Instructors interviewed during this study considered including enough information for students to be successful yet keeping the items simple as to not promote confusion.

4.5.2.3.3 Significant figures and units

Many of the instructors discussed the use of significant figures and units in assessment items. Including appropriate and consistent significant figures and units was an important consideration and assessment strategy of many interviewed. Conversely, some interviewed disregarded appropriate use of significant figures in MC items. Example quotes will now be given and discussed.

Dr. Bennett mentioned how he considers the use of significant figures to be an important consideration while designing MC items. While evaluating the fifth item in the second phase of the interview (see Figure 3.3 in section 3.6.2). Dr. Bennett said:

Also, not a fan of... This is my thing... If we are going to teach sig figs lets teach sig figs and there is ambivalence in terms of sig figs here. So yeah.

Furthermore, Dr. Madison also mentioned the value of consistent use of significant figures while evaluating the same item. Dr. Madison said:

I tried to emphasize to my students the use of significant figures, and so the answers don't match the significant figures in the problem, because you've got A, B ... Answer A, B and D all have one significant figure, answer C and E have two significant figures, whereas one molar has one significant figure, 10 grams has one significant figure, and so the inconsistency there, I don't like that. Not that a student might necessarily ... I mean, given those choices, if they get close, they might say, "Okay." But I do try to be consistent when I write a question.

Conversely, some instructors were not concerned with the appropriate use of significant figures in MC items if the significant figures don't effect the validity of the item. For example, Dr. Lopez said while designing an item during phase three of the interview:

And here my colleagues would start telling me that I am disregarding significant figures, and I'm okay with that because for this question I am not caring about significant figures.

Moreover, Dr. Crawford mentioned not being concerned with significant figures. She said while evaluating the fifth item in phase two of the interview (see Figure 3.3 in section 3.6.2):

Some people might be upset that some of the answers have to two sig figs and others don't. That doesn't super bother me because I don't care that much about sig figs. People will learn stuff like that in analytical chemistry. I teach them, but I'm not a stickler about them. But some people might be upset by that.

Regarding units, Dr. Johnson discussed the importance of using correct units in MC items.

Dr. Johnson said:

Interviewer: Okay. You mentioned that you don't like writing out the units, you like to just write it out as M or g in this case, could you tell me why?

Dr. Johnson: Yes, because I think that students need to be exposed to units as they are, based on previous knowledge and based on what they learn, they should be familiar with the scientific notations and scientific units. And so, they should know that M is molar for example, and the options, we have 50 mL and everybody knows mL is milliliters. And so, this is something that students, that is important for them as they go on in other courses, as they go on through the rest of the course, and so it almost seems like a crutch to write all these things out fully so that students don't remember it

4.5.2.3.3.1 Discussion

Knowledge of using significant figures and units appropriately may be a discipline, or at least science specific, assessment strategy. The variation in viewpoints on the use of significant figures was worth noting. Several instructors such as Dr. Bennett and Dr. Madison try to be consistent and correct with significant figure use. On the other hand, some instructors, like Dr. Lopez and Dr. Crawford, felt that significant figures were not important to consider in general chemistry assessments if their ambiguity does not affect student ability. Furthermore, the appropriate use of units was looked upon favorably among those interviewed.

Literature on MC item writing does not address the use of significant figures or units directly, however, in an article geared toward chemistry faculty, Towns advocates for appropriate science (nomenclature, chemical symbolism) as a minimum standard in MC items (Towns, 2014). Additionally, avoiding cues to students that could help them to select or eliminate answer choices is certainly recommended (Haladyna et al., 2002a; Haladyna & Rodriguez, 2013). In specific circumstances incorrect significant figures or units could provide that cuing. These data show that there are chemistry specific assessment strategy considerations.

4.5.2.3.4 Formatting consistency

Another overall item assessment strategy considered by the chemistry instructors was ensuring that item formatting was consistent. Example quotes will now be given and discussed.

Dr. Madison discussed the importance of consistent formatting in MC items. Dr. Madison said:

I mean, of course there are all of the mechanics of it. Because making an exam, you got all that other stuff that's entailed in that as well, making sure it looks consistent, make sure that you've got ... if you're gonna use ABC, whatever, don't use A-parenthesis, B-parenthesis, C-parenthesis on some of them and not on others, because then it just looks sloppy and students are like, "Oh wow, he doesn't really care about this test. Why should I care about this?" So, I mean it's just little things like that.

Furthermore, Dr. Brown noticed inconsistency in the formatting of the chemical nomenclature of item six evaluated during phase two of the interview (see Figure 3.3 in section 3.6.2):

Dr. Brown: Also, I just noticed that the 2 on CO₂ is not subscript.

Interviewer: Okay.

Interviewer: And not having that subscripted, that's something that you would change?

Dr. Brown: Well, that is something I would change, absolutely.

Interviewer: And why is that?

Dr. Brown: Because you need to be consistent. Every other time CO₂ is mentioned, both in the graph and in the other possible answers, the 2 is subscripted.

Interviewer: Okay. Thank you.

Dr. Brown: I think consistency is important.

Moreover, Dr. Bennett mentioned formatting consistency after evaluating the six items in phase two of the interview (see Figure 3.3 in section 3.6.2):

Um... Let's see... So (laughing) here's the little things that I notice that I try to do now. Person Y was really particular about the formatting, so I notice that all the time now. so, the indentations are different here. So, I'd make sure they all lined up. And like with the questions and everything like that so it was very clear. Yeah so, I think mostly making sure that the format is very consistent throughout it.

4.5.2.3.4.1 Discussion

Many of the instructors considered formatting while creating or evaluating MC items. For example, Dr. Brown felt that consistency was important regarding chemical nomenclature and Dr. Madison believed that the attention to formatting detail represents how much the instructor cares about the course. These data about formatting consistency are aligned with recommendations in

the literature to have MC items written in a way that is consistent and easily understood (Haladyna & Downing, 1989).

4.5.2.4 The exam overall

The chemistry instructors interviewed did consider assessment strategies that apply to the creating the exam as a whole. The strategies that fell into this category were item order and exam key balancing. These codes were discussed by very few instructors. Examples will now be given and discussed.

Regarding item order, those who discussed this preferred to start their exams with easy items and then increase the difficulty toward the end in hopes to encourage students. Dr. Bennett said:

Just being a test taker, it was nice if your first couple questions were easy questions and that's how I try to have it so that people who not good test takers are don't get thrown off right away and get really freaked out. So, I definitely like to have very easy questions but it's a small number of points so I'm not really worried that it's going to get 100% correct hopefully in the class.

Furthermore, Dr. Patrick preferred to start exams with easy items to encourage and motivate students as well.

Dr. Patrick: I like to start with the easy ones and then get harder. I like to encourage people first when they start taking the exam rather than have the hard ones at the beginning and then the easy ones at the end.

Interviewer: Okay, so by putting easier questions first that's encouraging students.

Dr. Patrick: That's what I think, yes.

Interviewer: Okay.

Dr. Patrick: But then also in the middle I put easy ones in to just ... if they have a low point and they go like, "I really have no idea what she's talking about" then they're encouraged again.

Key balancing, or the act of changing where the correct answer is located in the response set, was only discussed by one participant. Dr. Johnson said:

Another option that I put as option number B, and I but C as the correct answer. I know that there has been a study where students know that usually the right answer doesn't come as A, and so sometimes I do try to mix up the right answer because as instructors we're like oh, we don't want the first option to be the right option- [crosstalk 01:03:07] to know, and so they look at the last two options. And so it's just a study I read, and so I try to change up where the right answer may be, so that they don't, for some reason, just because of my bias, just because I want to put it as not the first option, that I may end up putting all of them in C or D, right.

4.5.2.4.1 Discussion

The fact that assessment strategies related to the exam overall were not discussed by many participants was interesting to note. The strategies that were discussed (item order and key balancing) are supported in the literature as appropriate to consider (Downing et al., 2010).

However, in regards to ordering items by difficulty, it has been found that placing too many difficult items next to each other can effect student performance on subsequent items (Schroeder, Murphy, & Holme, 2012). Thus, placing easy items near the beginning of the exam to encourage students may be appropriate, but instructors need to be careful not to load too many difficult items together at the end of an exam.

4.6 Knowledge of assessment interpretation and action taking

The ability to interpret and act on assessment data has been identified as an important aspect of assessment literacy (Abell & Siegel, 2011; Pellegrino, 2001). Although this study was not designed to probe specifically at this topic of assessment interpretation and action taking (AIAAT), some data provides limited insight into general chemistry instructors' knowledge of AIAAT.

Instructor comments fell into two main areas, their review of item statistics and their use of exam data. Example quotes will now be presented and discussed.

Regarding the review of exam statistics, the participating instructors mainly discussed using item difficulty and discrimination values in the MC exam creation or evaluation process. For example, while evaluating item four in the second phase of the interview (see Figure 3.3 in section 3.6.2). Dr. Sorenson connects the poor answer choices in the item to a potentially low discrimination value. She said:

Four is electronegativity, the electronegativity question and I mean I think, it's a decent question...I think the one about the pessimistic electrons...no one's going to pick that so that's kind of a give me they can eliminate that one, then they have four different things, and we do talk about it in a periodic trends lecture, so I don't think they are going to pick A. So really what they are deciding for, or deciding between is C, D, and E. So, I guess in that regard maybe I don't think it's a great question because if you can eliminate two of the answers right away then statistically speaking, they can know nothing and still do pretty well on it. So maybe it doesn't have a good discrimination or whatever as the [statistical] analysis tells

Furthermore, Dr. Crawford who has extensive training in educational measurement, discussed looking at item statistics of old exam items before including them into a current exam. Of all the participants who discussed item statistics, Dr. Crawford appeared the most confident. She said:

And then I'll usually look and see if there are items I've used before, I might look and see what were the statistics the last time we used them, was there anything that was funny where you had negative discrimination or something like that, difficulty scores above .9 or below .5 and see if there's something I can do to modify that.

Moreover, Dr. Patrick discussed using exam statistics to gauge how effective a previously administered item may be in an exam. In her comments, she mentions using the item discrimination value but is not confident about its use. She said:

Dr. Patrick: So first I'll go through the pool of multiple choice questions that have already shown to be effective, like in that sense I'm looking at the [statistic] report that we get, see how effective those questions were and how well the students answered those and how well they targeted the learning outcomes that I really wanted to get at.

Interviewer: Okay. Can you tell me more about how you use those reports to determine the quality in question?

Dr. Patrick: Yeah, so the reports have on the side, they have like an index indicator, like a number associated with each question. I truly forgot what the number has to look like in order to see if this is a good or bad question, but the indicator tells you. If it's above a certain point range, it means that many students did not get this question right or did get this question right. And I would look for the questions that was well distributed, where many got it right. Some did get it wrong, but most of the students got it to what I really ... or like, I could see that they got what I wanted to get at. And those would be the questions that I would select again.

Regarding the use of exam data, instructors' comments tended to fall into two main categories, namely, using exam data to change teaching practice, and using it to understand student misconceptions. Example quotes will now be presented and discussed. For example, Dr. Irvine discussed using exam data to change teaching practice. He said:

Dr. Irvine: The department has divided us into teaching groups, and then general chemistry teaching group meets together and we have created some standard questions that everybody has to give, so those ones are part of my question bank, and we report how the students do in each question, including what option they each selected. And so, the person over the teaching group gathers the data and produces a report, and lets us know how our students are doing compared to the average, and still we can see if we need to change any teaching.

Interviewer: And so, you really use that data to change how you're going to assess your students for the current semester, then?

Dr. Irvine: Yeah, yeah, and to determine if some modifications to happen to the teaching activities.

Furthermore, Dr. Deagen, discussed using exam data to remediate or change teaching regarding a VSEPR item. She said:

Then when I get the data back, then I know, "Where's this falling apart? Is it falling apart in the geometry or is it falling apart in the polarity?" Then you can go back and remediate that.

Moreover, Dr. Smith discussed how she decided to revisit a subject because many of her students missed an item on her exam. She said:

Other things, like for instance the question about sulfur dioxide and resonance, a bunch of my students missed this one as well. So, that meant when we met in lecture and in discussions the following week, we talked more about this and why one answer was right versus the other. It made me realize that we needed to revisit the subject.

The other main way that the participating instructors used exam data was to understand their student's misconceptions. For example, Dr. Irvine discussed how item three in the second phase of the interview (see Figure 3.3 in section 3.6.2) could be used to diagnose student misconceptions. He said:

I think this is one where selecting a particular answer would reflect the misconception that, if a student comes in for a review, then we can talk about that misconception.

Furthermore, Dr. Johnson mentioned how she used student responses to MC items to understand where they may have misconceptions. She said:

Dr. Johnson: The first, that it helps me to understand student misconceptions, and so I usually put some misconceptions as the other answer, and it helps me immediately identify that. And so, I grade my multiple choice through scantron and in our office, our instructional development, they give me a histogram with how many students got which question right and wrong and so I know okay, well this is a misconception I

absolutely need to address. And so, it helps me catch that. But I think it also helps students to think deeper about the topic, especially reflect on things that I've said in class, just all these little nuances and so that's why I like doing that.

Interviewer: Okay. That's great. And so, you mentioned that you get the statistic reports with your exams.

Dr. Johnson: Yeah.

Interviewer: And you said you use those to understand misconceptions?

Dr. Johnson: Yes. Yes.

Moreover, Dr. Smith mentioned that she likes using MC items because it is easy to use the data to see what misconceptions student may have. She said:

But then later I realized that oh no, wait. If you have a really good multiple-choice question and you have really good distractors, it can actually be really insightful. And it's also very easy then to tabulate the data and start to figure out where the misconceptions are. So, now I'm a pretty big fan.

4.6.1 Discussion

The fact that codes focused on assessment interpretation and action taking were sparse in this research is not surprising as the study did not probe directly into this area of assessment literacy. However, the comments that fit into this category were focused around important aspects of assessment design, namely, interpreting item statistics and using exam data to improve teaching or to understand students (Harshman & Yeziarski, 2017).

In a guide to developing quality MC exams, Towns described the importance of interpreting and using item analysis statistics such as difficulty and discrimination to improve item writing (Towns, 2014). Few of the participating instructors discussed using item level statistics, and when they did, they often felt unsure in their knowledge of these measures. For example, Dr. Patrick was unsure about how to interpret item discrimination values and could only talk about

them generally. This is in-line with literature that shows chemistry faculty being unfamiliar with item analysis terms (Raker & Holme, 2014). When it comes to analyzing exam data, tools have also been developed that can assist chemistry instructors in this effort (Brandriet & Holme, 2015). Contrarily to Dr. Patrick, Dr. Crawford, who had extensive educational measurement experience was confident in her use of item-level statistics. This may suggest that with further training confidence in using item-level statistics could increase.

Although the context of this study was focused on summative assessment, those who discussed the use of assessment data, discussed using it to alter teaching practice or to understand student misconceptions. Using assessment data in these ways (even from summative assessments) is encouraged in hopes to improve how chemistry is being taught (Holme et al., 2010).

Overall, the discussions of AIAAT by the instructors in this study were sparse, however those who did interpret or use exam data focused on item-level statistics and using assessment data in formative ways. Regarding the interpretation of exam data, there was uncertainty present in several of the instructors. Further training is recommended. Those who discussed their use of exam data discussed using it in formative ways. This is in-line with literature that recommends this type of data use for the improvement of chemistry courses (Holme et al., 2010).

CHAPTER 5. DEVELOPMENT AND USE OF A MULTIPLE-CHOICE ITEM WRITING FLAWS EVALUATION INSTRUMENT IN THE CONTEXT OF GENERAL CHEMISTRY

This chapter has been reproduced from (Breakall, Randles, & Tasker, 2019) with permission from The Royal Society of Chemistry.

5.1 Abstract

Multiple-choice (MC) exams are common in undergraduate general chemistry courses in the United States and are known for being difficult to construct. With their extensive use in the general chemistry classroom, it is important to ensure that these exams are valid measures of what chemistry students know and can do. One threat to MC exam validity is the presence of flaws, known as item writing flaws, that can falsely inflate or deflate a student's performance on an exam, independent of their chemistry knowledge. Such flaws can disadvantage (or falsely advantage) students in their exam performance. Additionally, these flaws can introduce unwanted noise into exam data. With the numerous possible flaws that can be made during MC exam creation, it can be difficult to recognize (and avoid) these flaws when creating MC general chemistry exams. In this study a rubric, known as the Item Writing Flaws Evaluation Instrument (IWFEI), has been created that can be used to identify item writing flaws in MC exams. The instrument was developed based on a review of the item writing literature and was tested for inter-rater reliability using general chemistry exam items. The instrument was found to have a high degree of inter-rater reliability with an overall percent agreement of 91.8 % and a Krippendorff Alpha of 0.836. Using the IWFEI in an analysis of 1,019 general chemistry MC exam items, it was found that 83% of items contained at least one item writing flaw with the most common flaw being the inclusion of implausible distractors. From the results of this study, an instrument has been developed that can

be used in both research and teaching settings. As the IWFEI is used in these settings we envision an improvement in MC exam development practice and quality.

5.2 Introduction

Assessing what students know and can do is an integral part of the teaching process and the responsibility of chemistry instructors (Bretz, 2012). This is because assessment informs instructional decisions, grade assignments, and can be used to alter curriculum (Pellegrino, 2001). Using assessment data in these ways is important for the improvement of how chemistry is being taught (Holme et al., 2010).

With the improvement of the teaching process being fundamentally connected with assessment, the quality of assessment has been an area of focus in the chemical education community since the 1920s (Bretz, 2013). Research into the assessment of student learning is diverse as represented by publications in Table 5.1. From this sizeable but not exhaustive list, we can see that improving assessment practice and quality has been a focus of those interested in bettering chemistry instruction.

Table 5.1. Recent work on improving assessment practices in chemistry courses

Research Topic	Reference
Investigating faculty familiarity with assessment terminology	(Raker, Emenike, & Holme, 2013; Raker & Holme, 2014)
Surveying chemistry faculty about their assessment efforts	(M. E. Emenike et al., 2013)
Developing two-tier multiple-choice instruments	(Chandrasegaran, Treagust, & Mocerino, 2007),
Analyzing ACS exam item performance	(Kendhammer et al., 2013; Schroeder et al., 2012)
Investigating teachers use of assessment data	(Harshman & Yeziarski, 2015)
Studying content coverage based on administered ACS exams	(Reed, Villafañe, Raker, Holme, & Murphy, 2017)
Creation of exam data analysis tools	(Brandriet & Holme, 2015)

Table 5.1 continued

Using Rasch modelling to understand student misconceptions and multiple-choice item quality	<i>(Herrmann-Abell & DeBoer, 2011)</i>
Development of instruments that can be used to measure the cognitive complexity of exam items	(Knaus, Murphy, Blecking, & Holme, 2011; Raker, Trate, Holme, & Murphy, 2013)
Aligning exam items with the Next Generation Science Standards	(M. M. Cooper, 2013; Lavery et al., 2016; Reed et al., 2016)
Creating exams that discourage guessing	(Campbell, 2015)
Outlining how to develop assessment plans	(Townes, 2010)

Self-constructed assessments, particularly multiple-choice (MC) assessments, are popular in undergraduate chemistry courses. In a recent survey, 93.2% of 1,282 chemistry instructors from institutions that confer bachelor's degrees in chemistry, reported creating their own assessments (Gibbons et al., 2018). Furthermore, in a survey of 2,750 biology, chemistry, and physics instructors from public and private US institutions, 56% of chemistry instructors reported using MC exams in some or all of their courses (Goubeaud, 2010). A practical reason for the popularity of MC assessments is their ease and reliability of grading for large groups of students (Thorndike & Thorndike-Christ, 2010).

Although MC exams tend to have high grading reliability, creating valid MC items that perform reliably is difficult and requires skill to do properly (Pellegrino, 2001). Against this background, chemistry instructors typically have little to no formal training in appropriate assessment practices (Lawrie et al., 2018). In addition to a lack of training, another reason creating MC assessments can be difficult is because there are numerous ways to lessen an exam's validity based on how it is designed. When designing MC exams, it is important to follow appropriate design recommendations to increase the exams validity. Many of these recommendations are outlined in the literature as item writing guidelines (Haladyna et al., 2010).

Disregarding these guidelines can introduce variance in student performance that stems from something other than the student's knowledge of what is being tested (Downing, 2002). This is known as construct irrelevant variance (CIV) ((Thorndike & Thorndike-Christ, 2010). Examples of CIV include students using test-taking strategies, teaching to the test, and cheating (Downing, 2002).

CIV in the form of testwiseness, which is defined as a student's capacity to utilize the characteristics and formats of a test and/or the test taking situation to receive a high score (Millman & Bishop, 1965), has been studied in students answering MC chemistry exam items (Towns & Robinson, 1993). Although it was found that students do use testwise strategies when answering exam items (Towns & Robinson, 1993), these exams can be written so as to reduce CIV in the form of testwiseness (Towns, 2014).

5.2.1 Multiple-choice item writing format and writing guidelines

A MC item typically consists of a stem, or problem statement, and a list of possible answer choices known as a response set. The response set usually contains one correct answer and incorrect options known as distractors. The ability to write MC items that test a desired construct and avoid CIV is difficult and takes time and practice to develop (Pellegrino, 2001).

MC items can be evaluated in several ways which can help a test creator make judgements about how an item functions and what performance on that item means about student learning. Two common quantitative measures of MC items are the difficulty and discrimination indices. Item difficulty, which is typically reported as a percentage, is the proportion of students who got an item correct over the total number of students who attempted the item. This can give an idea of how hard or easy an item is for the students

tested. Item discrimination, which can be measured in several ways, is a measure of how well an item separates students based on ability. The higher the discrimination value the better the item can differentiate between students of different ability levels. These measures, although they cannot tell you all things about an item's quality, can give an idea of how the item functions for the students tested.

Guidelines for how to write MC items have been published (Frey et al., 2005; Haladyna et al., 2010; Moreno & Marti, 2006; Towns, 2014). Following or disregarding these guidelines can affect how students perform on exams (Downing, 2005).

Neglecting to adhere to item writing guidelines can affect exam performance in ways that are significant to both instructors and students. For example, in a 20 item MC general chemistry exam, each item is worth five percent of a student's exam score. Therefore, if even a few of the items are written in a way where a student either answers correctly or incorrectly based on features of the item instead of their understanding of the chemistry, this can unjustifiably inflate or deflate the student's exam grade by 10-20%. This has been demonstrated in a study where flawed items may have led to the incorrect classification of medical students as failed when they should have been classified as passing (Downing, 2005).

Many item writing guidelines exist, and they apply to five main aspects of MC exams including, preparing the exam, the overall item, the stem, the answer set, and the exam as a whole. These guidelines will now be outlined.

5.2.2 Preparing the exam

Linking multiple-choice items to one or more objectives of the course can help to ensure that the items are appropriate and valid for the exam being administered (Towns,

2010). This is often accomplished through the creation of a test blueprint. A test blueprint outlines the content and skill coverage for any given exam (Thorndike & Thorndike-Christ, 2010). It is important to note that the more objectives an item is testing the less the item can tell you about a student's knowledge of a specific objective.

5.2.3 The overall item

Multiple-choice items need to be written clearly (Haladyna & Rodriguez, 2013; Holsgrove & Elzubeir, 1998). This is because if an item is not clearly written, (i.e. can be interpreted incorrectly), students may select an incorrect answer choice even though they understand the chemistry being tested. It has been found that even small changes in the wording of an item can have significant effects on student performance (Schurmeier et al., 2010). For instance, using simpler vocabulary in chemistry exam items was found to significantly improve student performance when compared to the same items using more complex terminology (Cassels & Johnstone, 1984). This may be due to an increased demand on working memory capacity when more complex/unclear vocabulary is used (Cassels & Johnstone, 1984). Therefore, it is important that MC items are written using the simplest language possible, without sacrificing meaning.

Additionally, MC items need to be written succinctly (Haladyna et al., 2010). If an item is not succinctly written, unwanted CIV can be introduced. In a study of overly wordy accounting items, average student performance was worse when compared to more succinct versions of items testing the same content (Bergner, Filzen, & Simkin, 2016). This may be because overly wordy items can decrease reliability, make items more difficult, and increase the time it takes to complete an exam (Board & Whitney, 1972; Cassels & Johnstone, 1984; Rimland, 1960; Schrock & Mueller, 1982). This may be

because if an item is overly wordy, it can be testing reading ability or constructs other than the chemistry concepts being tested. Figure 5.1 shows examples of wordy and succinct items.

<p>What is the electron-pair geometry and molecular geometry of ICl_3?</p> <p>a. The electron-pair geometry is trigonal-planar and the molecular geometry is trigonal planar.</p> <p>b. The electron-pair geometry is trigonal-bipyramidal and the molecular geometry is trigonal planar.</p> <p>c. The electron-pair geometry is trigonal-bipyramidal and the molecular geometry is linear.</p> <p>d. The electron-pair geometry is linear and the molecular geometry is linear.</p>	<p>What is the electron-pair geometry and molecular geometry of ICl_3?</p> <table border="0"> <thead> <tr> <th style="text-align: left;">Electron-Pair Geometry</th> <th style="text-align: left;">Molecular Geometry</th> </tr> </thead> <tbody> <tr> <td>a. Trigonal-planar</td> <td>Trigonal-planar</td> </tr> <tr> <td>b. Trigonal-bipyramidal</td> <td>Trigonal-planar</td> </tr> <tr> <td>c. Trigonal-bipyramidal</td> <td>Linear</td> </tr> <tr> <td>d. Linear</td> <td>Linear</td> </tr> </tbody> </table>	Electron-Pair Geometry	Molecular Geometry	a. Trigonal-planar	Trigonal-planar	b. Trigonal-bipyramidal	Trigonal-planar	c. Trigonal-bipyramidal	Linear	d. Linear	Linear
Electron-Pair Geometry	Molecular Geometry										
a. Trigonal-planar	Trigonal-planar										
b. Trigonal-bipyramidal	Trigonal-planar										
c. Trigonal-bipyramidal	Linear										
d. Linear	Linear										

Figure 5.1. Example from (Towns, 2014) wordy item left; succinct item right. <https://pubs.acs.org/doi/abs/10.1021/ed500076x> Further permissions related to this figure should be directed to the ACS.

Items should be free from grammatical and phrasing cues (Haladyna & Downing, 1989). These are inconsistencies in the grammar or phrasing of an item that can influence how a student interprets, understands, and answers the item (Haladyna et al., 2010). These

<p>Grammatically Inconsistent</p> <p>Carbon has ____ protons.</p> <p>(a) One</p> <p>(b) Three</p> <p>(c) Six</p> <p>(d) Twelve</p>	<p>Grammatically Consistent</p> <p>Carbon has ____ proton(s).</p> <p>(a) One</p> <p>(b) Three</p> <p>(c) Six</p> <p>(d) Twelve</p>
---	---

Figure 5.2. Grammatically Inconsistent Item left; Grammatically consistent item right (Examples from appendix 1, see supplemental information)

can include associations between the stem and the answer choices that may cue a student to the correct answer. Grammatical inconsistencies can provide cues to students that can affect the psychometric properties of the exam item (Dunn & Goldstein, 1959; Plake, 1984; Weiten, 1984). See Figure 5.2 for examples of grammatically inconsistent and grammatically consistent items. In the grammatically inconsistent item in Figure 5.2, the answer choice (a) “one” is not consistent with the stem in that it would read, “Carbon has one protons”. This may cue students into the fact that this may not be the correct answer choice. Therefore, MC items should be written without grammatical or phrasing cues.

Items that contain multiple combinations of answer choices (known as K-type items) should be avoided when constructing MC exams (Albanese, 1993; Downing et al., 2010). An example of this format is shown in Figure 5.3. K-type items have been shown to contain cueing that decreases reliability and inflates test scores when compared to multiple true-false (select all that are true) items (Albanese, Kent, & Whitney, 1979; Harasym, Norris, & Lorscheider, 1980). This is likely due to the fact that examinees can easily eliminate answer choices by determining the validity of only a few responses (Albanese, 1993). Therefore, K-type items should be avoided as they can contain cues that decrease reliability and increase test scores.

Items should be kept to an appropriate level of cognitive demand. Although there are several ways to determine cognitive demand, including thorough and effective instruments such as the cognitive complexity rating tool (Knaus et al., 2011), one simple

- Which of the following molecules are polar?
1. O₂
 2. NH₃
 3. CO₂
 4. HCN
- a. 1, 2, 3 only
 - b. 1, 3 only
 - c. 2, 4 only
 - d. 4 only
 - e. All are correct

Figure 5.3. Example of k-type item format

guideline is that an item should not require more than six ‘thinking steps’ to answer. A thinking step is defined by Johnstone as a thought or process that must be activated to answer the question (Johnstone & El-Banna, 1986). If more than six thinking steps are involved in an item, student performance can decrease sharply due to working memory overload (Johnstone, 1991, 2006; Johnstone & El-Banna, 1986; Tsaparlis & Angelopoulos, 2000). This working memory overload may interfere with the student’s ability to demonstrate their understanding of the concept(s) being tested (Johnstone & El-Banna, 1986). Furthermore, in studies of M-demand (defined as the maximum number of steps that a subject must activate simultaneously in the course of executing a task), it has been shown that student performance decreases as the M-demand of an item increases (Hartman & Lin, 2011; Niaz, 1987, 1989). Although, the number of thinking steps in an item can be difficult to determine due to chunking, and varying ability levels, keeping the general number of

thinking steps in an item to six or less can help ensure that an item is assessing understanding of chemistry and not cognitive capabilities.

5.2.4 Stem creation guidelines

When writing a MC item, the central idea should be included in the stem (Haladyna et al., 2010; Towns, 2014). In other words, an item should be answerable without looking at the answer choices. Items with a stem that fails to include the central idea (unfocused stems), have been shown to increase item difficulty and decrease reliability (Board & Whitney, 1972; Dudycha & Carpenter, 1973). Thus, writing stems that include the central idea is recommended. See Figure 5.4 for examples of focused and unfocused stems.

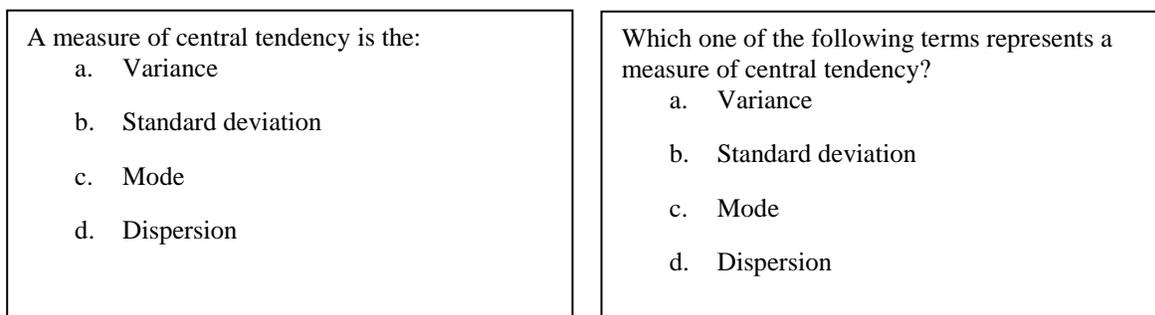


Figure 5.4. Unfocused (left) focused (right) Reproduced from (Dudycha & Carpenter, 1973) with permission

Positively worded items are recommended over negatively worded items (Haladyna & Rodriguez, 2013). This is because negative phrasing in an item can increase reading and reasoning difficulty, introduce cueing, and that testing a student's understanding of an exception is not always the same as testing their understanding of the actual learning objective (Cassels & Johnstone, 1984; Harasym et al., 1992; Thorndike & Thorndike-Christ, 2010). If an item must be worded negatively to test a desired skill or construct, the negative phrase should be highlighted (Haladyna et al., 2010). Negatively phrased items

were shown to be more difficult when the negative phrase was not emphasized (Casler, 1983). This may be because highlighting the negative phrase minimizes the risk that a student will miss the negative phrase while reading the item. Therefore, if negative phrasing must be used, it is recommended that it is emphasized.

5.2.5 Answer choice creation guidelines

The first guideline in creating answer choice sets is that all distractors should be plausible (Haladyna & Downing, 1989; Haladyna et al., 2010; Weitzman, 1970). This is because non-plausible distractors can be eliminated easily by students, decrease item difficulty, increase reading time, and lessen the number of items that can be included in an exam (Ascalon, Meyers, Davis, & Smits, 2007; Edwards, Arthur, & Bruce, 2012; Papenberg & Musch, 2017; Schneid et al., 2014; Tarrant & Ware, 2010). One way to attempt to create distractors are plausible, is to create them by using student errors or misconceptions (Case & Swanson, 2002; Gierl et al., 2017; Moreno & Marti, 2006; Tarrant et al., 2009). This strategy can help ensure that distractors are errors that students may make and therefore are more likely to choose (Gierl et al., 2017; Tarrant et al., 2009). Several studies have defined and operationalized implausible distractors as ones that fewer than 5% of students select (Tarrant & Ware, 2010; Tarrant et al., 2009; Wakefield, 1958)

The second answer choice set creation guideline is that ‘all of the above’ should be avoided (Downing et al., 2010; Haladyna et al., 2002b). This is because its use has been shown to significantly enhances student performance on MC items due to an inherent cueing effect (Harasym et al., 1998). This improvement in student performance is likely due to students being able to more easily eliminate or select ‘all of the above’ when compared to items in a ‘select all that are true’ format (Harasym et al., 1998).

Third, answer choices should be arranged in a logical order (Haladyna & Downing, 1989; Moreno & Marti, 2006). For example, in ascending or descending numerical order. In a study that found higher discrimination values on items with randomly ordered answer choices versus ones that were logically ordered, it was concluded that although answer choice order is not likely an influencing factor for higher-ability students, it may affect lower-ability students in their exam performance (Huntley & Welch, 1993). Additionally, arranging answer choices in logical or numerical order was found to be unanimously supported in a review of measurement textbooks (Haladyna et al., 2010).

The fourth recommendation is that the answer choices should be kept to approximately the same length/level of detail (Frey et al., 2005; Haladyna et al., 2010; Towns, 2014). This is because students can use the ‘choose the longest answer’ heuristic when taking an examination. Students may believe that if an answer choice is significantly longer or provides more detail than the other answer choices, then it is more likely to be correct. Research has shown that items with inconsistent length of answer choices are easier and less discriminating than items where the length of the answer choices are approximately equal (Dunn & Goldstein, 1959; Weiten, 1984).

5.2.6 Guidelines for the exam overall

Item placement in an exam relative to the other items can affect the psychometric properties of the item (Meyers, Miller, & Way, 2008). An exam creator should avoid placing three or more items that test the same cognitive tasks next to each other on an exam. This can lead to cueing effects from the previous items that can inflate performance on the target item (Schroeder et al., 2012).

Additionally, placing three or more difficult items next to each other is also poor practice in MC exam creation. This has been found to decrease performance on general chemistry items in ACS exams (Schroeder et al., 2012). Possible causes of this effect may be self-efficacy or exam fatigue (Galyon, Blondin, Yaw, Nalls, & Williams, 2012; Schroeder et al., 2012)

An exam should have an approximately even distribution of correct answer choices, a practice known as key balancing (Towns, 2014). This is because in unbalanced exams, it has been shown that test makers and test takers have a tendency to choose the middle options in a MC item for the correct response (Attali, 2003). This produces a bias that effects the psychometric properties of an exam by making items with middle keyed answer choices easier and less discriminating (Attali, 2003). Additionally, test takers tend to expect different answers on subsequent items when they see a “run” of answer choices even though this is not necessarily true (Lee, 2018). Thus, it is good practice to ensure that there is an approximately even distribution of correct answer choices.

An exam should not link performance on one item with performance on another (Haladyna et al., 2002b; Hogan & Murphy, 2007; Moreno & Martı, 2006). All items should be independent from each other so that a student has a fair chance to answer each item correctly.

With this said, linked “two-tier” multiple-choice items have been used to assess student reasoning by asking them to explain, or choose an explanation to, their answer choice in a previous item (Chandrasegaran et al., 2007; Tan, Goh, Chia, & Treagust, 2002). While this practice has merit when used with purpose, simply linking items that test chemistry content can disadvantage students who incorrectly answer the first item.

5.2.7 Item writing guideline violations in higher education

Violations of item writing guidelines have been studied across different disciplines in higher education for their frequency in examinations as shown in Table 5.2. The frequency of these violations has been noted as an area of concern and a focus for improvement in their respective papers.

Table 5.2. Item writing guideline violations in higher education

Discipline	Reference	Sample Size	Frequency of flawed items	Most common flaws
Nursing	(Tarrant et al., 2006)	2,770 MC Items	46.2%	Unclear stem 7.5% Negative stem 6.9% Implausible distractors 6.6%
	(Tarrant & Ware, 2008)	664 MC Items	47.3%	Unfocused stem 17.5% Negative Stem 13% Overly wordy 12.2%
Pharmacy	(Pate & Caldwell, 2014)	187 MC Items	51.8%	All of the above 12.8% Overly wordy 12.8% K-type 10.2%
Medical Education	(Stagnaro-Green & Downing, 2006)	40 MC Items	100%	Phrasing Cues 100% Unfocused stem 100% Overly wordy 73%

5.2.8 Purpose of the study

Since many guidelines exist for how to construct MC items, it can be difficult to remember these guidelines during the test creation or test evaluation process. Therefore, in this work an instrument has been developed that can assist instructors and researchers to assess if MC exams are adhering to accepted guidelines. The developed instrument should:

- Represent common item writing guidelines
- Be able to identify item writing guideline violations in MC items
- Be able to be used reliably between raters
- Be able to be used with a minimal amount of training

Additionally, although adherence to item writing guidelines has been a topic of research in other academic disciplines (Pate & Caldwell, 2014; Tarrant et al., 2006), there

have been few research studies that have addressed the extent that general chemistry exams adhere to MC item writing guidelines. Therefore, once developed, the instrument was used to evaluate the adherence of 43 general chemistry exams (1,019 items) to item writing guidelines.

5.3 Methodology

This study occurred in two phases: instrument development and item analysis. These two phases are outlined in Figure 5.5 and will be described here in detail.

5.3.1 Instrument Development

The instrument was developed based on a review of the guidelines and literature described in the introduction. An appendix of definitions and examples of each item writing guideline was also created to help guide the instrument's users (See Appendix B). The instrument, known as the Item Writing Flaws Evaluation Instrument (IWFEI), was revised in a pilot testing cycle with four chemistry education graduate students as shown in the cyclic portion of Figure 5.5.

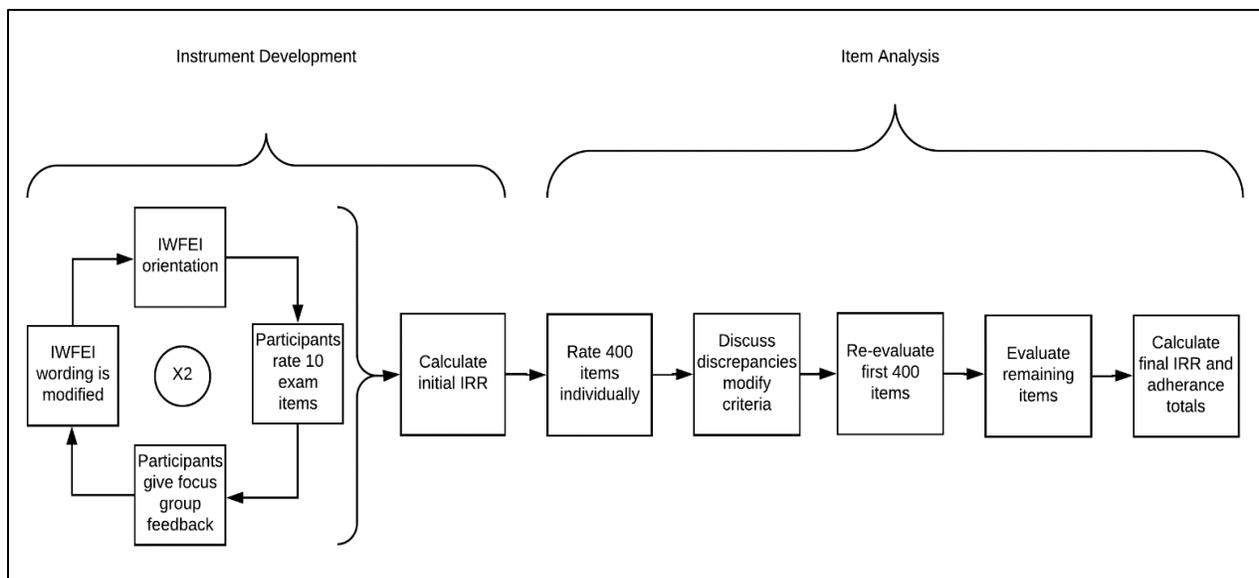


Figure 5.5. Flowchart of instrument development and item analysis

The participants were given a 20-minute orientation on how to use the IWFEI along with the IWFEI appendix. They then rated 10 general chemistry MC items individually and gave their feedback on the clarity of the IWFEI and definitions in a focus group setting. This cycle went through two iterations. The revised instrument was tested for initial inter-rater reliability among raters with various levels of teaching and MC exam writing experience (Table 5.3).

Raters 2 and 3 (Table 5.3) received a 20-minute orientation (from rater 1) on how to use the IWFEI and then raters 1-3 used the instrument to individually rate 10 general chemistry MC exam items. The 10 items were chosen because they contained a variety of item writing guideline violations that were within what the IWFEI intended to test. These items are found in Appendix C.

Table 5.3. Instrument development phase: Demographic information of inter-rater reliability raters

Rater	Position	Discipline	Years of experience teaching general chemistry	Number of general chemistry MC exams created
1	Graduate Student	Chemical Education	0-2	1-4
2	Assistant Professor	Inorganic	2-4	1-4
3	Instructor	Chemical Education	10+	20+

5.3.2 Item analysis of past exams

The resulting instrument was then applied to 43 1st semester general chemistry exams to evaluate their adherence to item writing guidelines and to further refine the instrument. The two raters have degrees in chemistry and are familiar with the course content. The 2nd rater received a 20-minute orientation from the 1st rater on how to use the IWFEI.

The chemistry exams analyzed were from a 1st semester general chemistry course for scientists and engineers at a public, R1 university in the midwestern United States. The exams were created by committees of instructors who taught different sections of the same course. A total of 33 unit exams and 10 final exams were analyzed. A unit exam is given during a semester and covers a subset of the course content. Typically, a unit exam covers approximately a fourth of the courses material. A final exam is typically cumulative and given at the end of the semester. The exams included topics common in first semester general chemistry such as: stoichiometry, gas laws, light and energy, radioactivity, periodic trends, thermochemistry, Lewis structures, polarity of molecules, intermolecular forces,

etc. The items contained various levels of representations including chemical formula, diagrams, molecular level representations, graphs, etc. If a representation in an item (such as chemical formulas) affected how to interpret a criterion in the IWFEI, it was noted in appendix 1 for the use of the instrument. The exams were administered between 2011 and 2016 and were composed of MC items exclusively. Unit and final exams contained 20 items and approximately 40 items, respectively. A total of 1,019 items were evaluated.

The two raters first evaluated 400 items individually. Then they discussed discrepancies in ratings among those 400 items. These discussions led to the modification of criteria wording and definitions in the IWFEI. Once modifications were made, the initial 400 items were re-evaluated using the updated instrument. Lastly, the remaining items in the data set were evaluated using the IWFEI.

Percent agreement and Krippendorff alpha statistics were then calculated between the two raters using the IRR package in R software. The raters then reached consensus on any disagreements. Once consensus was reached, percentages of how many items adhered to the various item writing guidelines in the instrument were calculated.

This study was approved by the institutional review board at the institution where the study took place.

5.4 Results

5.4.1 Item writing guideline evaluation instrument

The final version of the IWFEI (Figure 5.6) contains 11 criteria which apply to individual items and four criteria which apply to an exam overall. The IWFEI's format consists of a list of item writing guidelines, in the form of questions, and three choices of "yes", "no", or "not applicable". A "yes" suggests adherence to the item writing guideline

and a “no” suggests a violation of the guideline. A user of the IWFEI would rate *each* item in an exam with Criteria 1-11 and the exam as a whole using Criteria 12-15.

Criteria	Guideline	Yes	No	Not Applicable
1	Is the test item clear and succinct?			
2	If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?			
3	If the answer choices are numerical, are they listed in ascending or descending order?			
4	If the answer choices are verbal, are they all approximately the same length?			
5	Does the item avoid “all of the above” as a possible answer choice?			
6	Does the item avoid grammatical or phrasing cues?			
7	Could the item be answered without looking at the answer choices?			
8	Does the item avoid complex K-type item format?			
9	Is this item linked to one or more objectives of the course?			
10a	Are all answer choices plausible?			
10b	Are all answer choices plausible?			
11	Are there six or less thinking steps needed to solve this problem?			
12	Does the exam avoid placing three or more items that assess the same concept or skill next to each other?			
13	Does the exam avoid placing three or more difficult items next to each other?			
14	Is there an approximately even distribution of correct answer choices?			
15	Does the exam avoid linking performance on one item with performance on others?			

Figure 5.6. Item Writing Flaws Evaluation Instrument

Criterion 10, ‘Are all answer choices plausible?’ was given two definitions. Definition 10a being ‘All distractors need to be made with student errors or misconceptions.’ This definition may have utility as a reflective tool when the IWFEI is being used to analyze one’s own exam before it is administered, yet it was shown to have poor reliability when used to analyze exam items made by other instructors. Definition 10b is based on item statistics and defines an implausible distractor as one that fewer than 5% of students selected. This definition, (10b) should be used when using the IWFEI to evaluate historical exam items made by other instructors.

When the initial inter-rater reliability was calculated, a 78.2% agreement was found between the three raters with a 0.725 Krippendorff alpha. This suggests that there was a substantial level of agreement between the raters (Landis & Koch, 1977). Furthermore, when looking at the agreement between individual raters in Table 5.3, raters 1 and 2, 2 and 3, and 1 and 3 had a percent agreement of 83.3, 87.5, and 85.0%, respectively. Because all raters in Table 5.3 (less experienced and more experienced) agreed above an 80% level, we moved forward with the item analysis phase of the study. During that phase, when the final version of the instrument was used to rate the 1,019 items (10,458 total ratings) it had a 91.8% agreement and a 0.836 Krippendorff Alpha. As these reliability statistics were both above 80% and 0.80, respectively it was decided that an acceptable level of agreement had been reached.

Criterion 9 was used to analyze a subset of 96 items from the fall of 2016, instead of all 1,019 items, because of the lack of availability of learning objectives from previous semesters. Criterion 10 (using the 10a definition) had a consistently low level of reliability at 68.9%. Criteria 12-15 were rated 43 times, instead of 1,019 times, because they apply only once per exam being analyzed.

The inter-rater reliability statistics for the individual criteria of the IWFEI as used across the 1,019 items are found in Table 5.4.

Table 5.4. Inter-rater Reliability of the IWFEI Criteria

Criteria		Percent Agreement (%)
1	Is the test item clear and succinct?	93.9
2	If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded?	97.0
3	If the answer choices are numerical, are they listed in ascending or descending order?	95.9
4	If the answer choices are verbal, are they all approximately the same length?	90.8
5	Does the item avoid “all of the above” as a possible answer choice?	96.9
6	Does the item avoid grammatical and phrasing cues?	98.4
7	Could the item be answered without looking at the answer choices?	85.4
8	Does the item avoid complex K-type item format?	97.5
9	Is the item linked to one or more objectives of the course?	84.4
10a	Are all distractors plausible?	68.9
11	Are there six or less thinking steps needed to solve this problem?	94.5
12	Does the exam avoid placing three or more items that assess the same concept or skill next to each other?	95.3
13	Does the exam avoid placing three or more difficult items next to each other?	100.0
14	Is there an approximately even distribution of correct answer choices?	86.0
15	Does the exam avoid linking performance on one item with performance on others?	100.0

5.4.2 Item Analysis

An analysis of Criteria 1-11, when applied to individual items, revealed that on average, items contained 1.4 ± 0.8 violations per item. It was found that 80 items (7.9%) contained no violations, 505 items (49.6%) contained one violation, 347 items (34.1%) contained two violations, 73 items (7.2%) contained three violations, 11 items (1.1%)

contained four violations, and 3 items (0.3%) contained five violations of item writing guidelines. This is shown in Figure 5.7

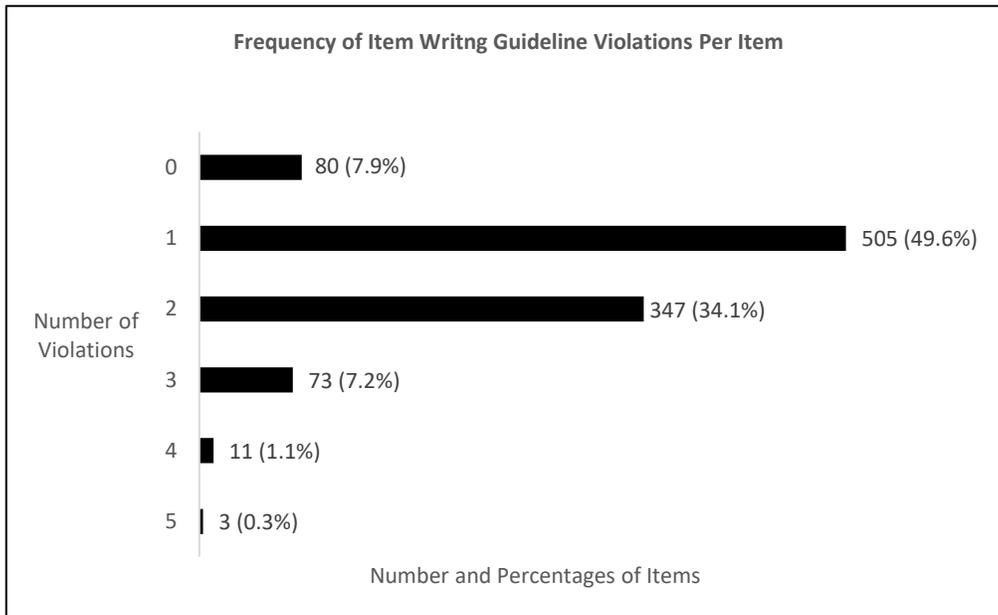


Figure 5.7. Frequency of item writing guideline violations per item

The number and percentages of items that adhered to and violated specific item writing guidelines (1-11), are found in Table 5.5. Boxes highlighted in blue represent guidelines classified as having a high level of adherence (above 90%) where those highlighted in red represent guidelines classified as having a low level of adherence (below 75%).

Table 5.5. Item writing guideline adherence and violation (per item)

	Abbreviated Criterion	Applicable items	Adhered (Yes)	Violated (No)	N/A	Adhered %	Violated %
1	Clear/Succinct	1019	960	59	-	94.2	5.8
2	Bolded negative phrase	77	65	12	942	84.4	15.6
3	Ascending/descending order of choices	311	226	85	708	72.7	27.3
4	Approx. even answer choice length	298	249	49	721	83.6	16.4
5	Avoid 'all of the above'	1019	973	46	-	95.5	4.5
6	Avoid grammar/phrasing cues	1019	1013	6	-	99.4	0.6
7	Answer without answer choices	1019	720	299	-	70.7	29.3
8	Avoid K-type	1019	963	55	-	94.5	5.4
9	Linked to objective	96	87	9	-	90.6	9.4
10a	Plausible distractors	1019	447	542	-	46.8	53.2
10b	Plausible distractors	1019	206	813	-	20.2	79.8
11	Six or less thinking steps	1019	972	44	-	95.4	4.3

Blue= above 90% adherence
Red= Below 75% adherence

Criteria 12-15, found in Table 5.6, apply to exams as a whole and not to individual items. Boxes highlighted in blue represent guidelines classified as having a high level of adherence (above 90%) where those highlighted in red represent guidelines classified as having a low level of adherence (below 75%).

Table 5.6. Item writing guideline adherence and violation (per exam)

	Abbreviated Criterion	Applicable exams	Adhered (Yes)	Violated (No)	N/A	Adhered %	Violated %
12	Avoid 3+ items testing the same concept next to each other	43	35	8	0	81.4	18.6
13	Avoid 3+ difficult items next to each other	43	42	1	0	97.7	2.3
14	Approx. even answer key distribution	43	17	26	0	39.5	60.5
15	Avoid linking items based on performance	43	43	0	0	100	0

Criteria two, three, and four only applied to a small portion of the items analyzed. See Table 4. Criterion two applied to 77 items with 84.4% of those items adhering to the guideline and 15.6% in violation. Criterion three applied to 311 items with 72.7% of those

items adhering to the guideline and 27.3% in violation. Criterion four applied to 298 items with 83.6% of those items adhering to the guideline and 16.4% in violation. The remaining criteria applied to all available items or exams.

5.5 Discussion

In this work, we developed the IWFEI based on accepted item writing guidelines and demonstrated that it can be used reliably to identify item writing guideline violations in 1st semester general chemistry multiple-choice exams. The development of the IWFEI addresses the lack of research literature on multiple-choice item writing guideline adherence in chemistry exams and provides a tool that can be used to continue research in this field. Its use has shown the frequency of item writing guideline violations in a sample of 1,019 1st semester chemistry exam items

5.5.1 Development

The development and refinement process used for the IWFEI has been used before in the creation of other instruments (Naeem, van der Vleuten, & Alfaris, 2012) and this process was used to improve wording and understandability of the instruments criteria. The criteria in the instrument were chosen based on accepted item writing guidelines that were discussed in the introduction section of this paper.

Additionally, the inter-rater reliability procedure of rating items individually, calculating agreement, and then coming to consensus on differences has been used in other studies successfully (Srinivasan et al., 2018).

Once developed, the instrument was tested for reliability and was shown to have a high level of reliability (91.8% agreement and 0.836 Krippendorff alpha) which is similar to other exam evaluation instruments such as the Cognitive Complexity Rating Instrument

(Knaus et al., 2011) and the Three Dimensional Learning Assessment Protocol (3D-LAP) (Lavery et al., 2016). It is important to note that the Cognitive Complexity Rating Instrument uses an interval rating scale versus the categorical scale used by the IWFEI. Although this makes it difficult to directly compare reliability values, they are at a similar high level.

The criterion that presented the greatest difficulty in using reliably was criterion 10 ‘Are all answer choices plausible?’. Initially, we began our analysis of the historical exam data using definition 10a which defines implausible distractors as being made with student errors or misconceptions. This proved to be an unreliable way to evaluate historical, non-self-constructed exam items with a percent agreement between raters of 68.9%. Although definition 10a may have utility when evaluating one’s own exam items before administration, it was not appropriate when evaluating items created by other instructors. Conversely, definition 10b which defines implausible distractors as ones that fewer than five percent of students select, would be appropriate for evaluating non-self-constructed exams, but would be impossible for analyzing exam items before administration or without exam statistics available. We see both definitions as having merit when used in appropriate situations. We foresee that having both definitions will increase the utility of the IWFEI and make it more useful to researchers and instructors alike.

5.5.2 Use of the Instrument

The IWFEI has been used in this study to identify items of concern in a sample of 43 general chemistry exams. For example, the item on the left-hand side of Figure 8 was identified by using the IWFEI as not being succinct (violation of criterion one). We notice that this item is overly wordy in the stem and the answer choices. This item would

take students a significant amount of time to read and this may disadvantage some students in their exam performance. In the literature, Towns described similar items as being overly wordy as shown on the right-hand side of Figure 5.8 (Towns, 2014).

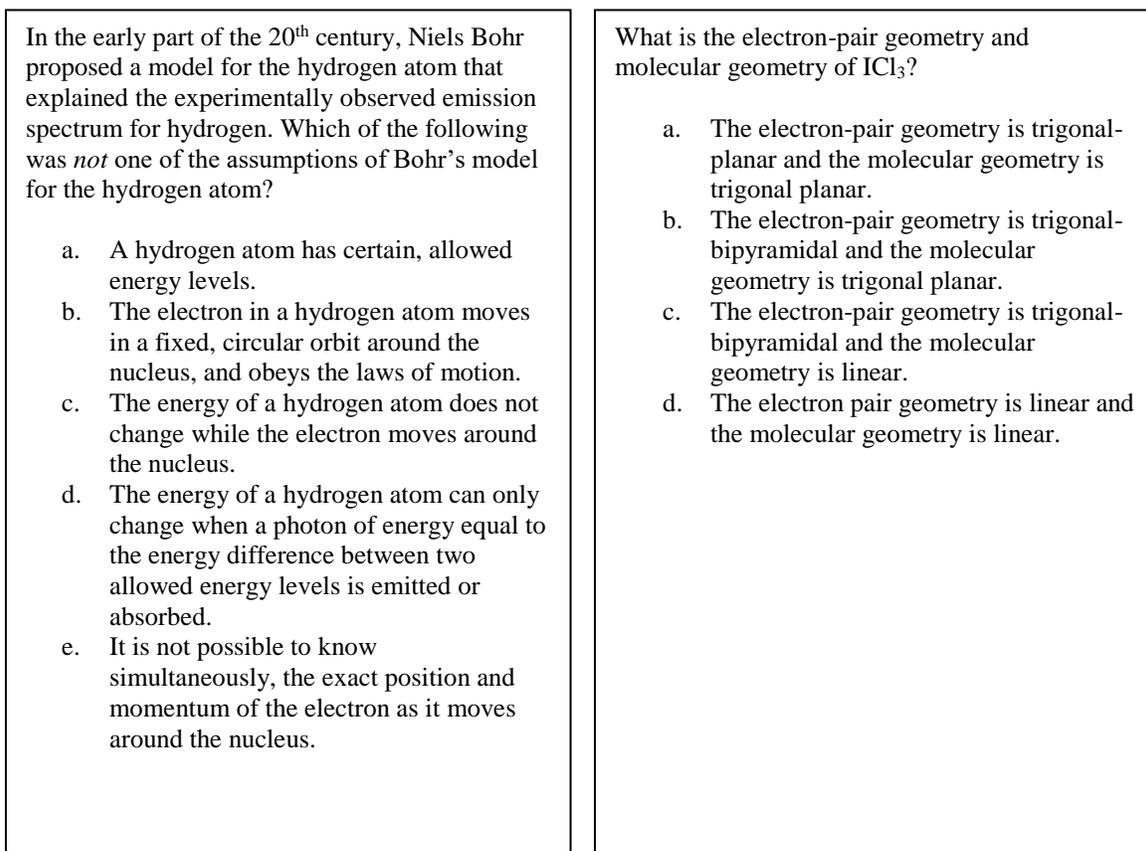


Figure 5.8. Overly wordy item identified using the IWFEI (left); Overly wordy item example from (Towns, 2014) (right) <https://pubs.acs.org/doi/abs/10.1021/ed500076x>
Further permissions related to the right side of this figure should be directed to the ACS

In another example, the item on the left-hand side of Figure 5.9 was identified as a K-type item by using the IWFEI. This item format has answer choices that can be easily eliminated based on analytic reasoning and thus may not provide the most valid data on

what students know. The item on the right in Figure 5.9 is an example of a K-type item as found in the literature.

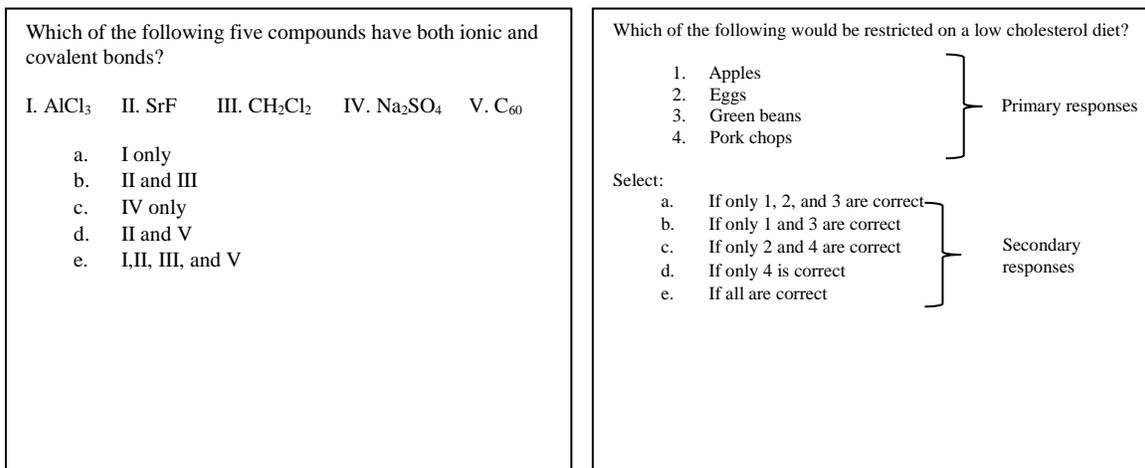


Figure 5.9 K-type items identified using the IWFEI (left) and found in the literature (right)

In a third example, the items in Figure 5.10 were identified as containing implausible distractors by using the IWFEI. In the literature, an implausible distractor has been given an operational definition as a distractor that fewer than 5% of students select. The percentage of students who selected each answer choice is indicated. From this, we see that the item on the left-hand side contains two implausible distractors, b and e, while the item on the right-hand side contains one implausible distractor, e. In the item on the right, e quintinary, is not a level of protein structure. This may indicate that the instructor included this as an answer choice solely for formatting reasons or a belief that in MC items more choices are better, despite their quality.

<p>The C=C double bond in ethylene, C₂H₄, is composed of:</p> <p>21% a) two pi bonds. 2% b) 3p³ hybridization. 13% c) two sigma bonds. 63% d) one pi bond and one sigma bond. 1% e) only unhybridized p orbitals.</p>	<p>Which of the levels of protein structure results from interactions between the backbone atoms of the polypeptide chain?</p> <p>17% a) primary 48% b) secondary 25% c) tertiary 9% d) quaternary 1% e) quintinary</p>
---	--

Figure 5.10. Items with implausible distractors identified using the IWFEI; Percentages of students choosing each answer choice is indicated

The three cases discussed above show examples of item writing flaws identified with the IWFEI and how they are comparable to what is described in the literature. This provides evidence that the IWFEI can be used in a valid way to identify items that contain flaws in their construction.

Once an item has been identified that contains violations of accepted guidelines, the user can then decide if/how they will modify the item. It is important to note that the IWFEI is not intended to identify *bad* items, but to identify items that may need to be modified based on the instructors or researcher's discretion.

To demonstrate this, a flawed item and its revision are shown in Figure 5.11. The original item in Figure 5.11 was identified as containing two flaws, an incomplete stem and implausible distractors (violations of criterion 7 and 10). In the revised version of the item, the stem was rewritten to contain a complete problem statement and the implausible distractors were removed.

Original	Revised
<p>The C=C double bond in ethylene, C₂H₄, is composed of:</p> <ul style="list-style-type: none"> a) two pi bonds. b) 3p³ hybridization. c) two sigma bonds. d) one pi bond and one sigma bond. e) only unhybridized p orbitals. 	<p>How many sigma and pi bonds are in the C=C double bond of ethylene, C₂H₄?</p> <ul style="list-style-type: none"> a) Two sigma, zero pi b) One sigma, one pi c) Zero sigma, two pi

Figure 5.11. Flawed item identified by using the IWFEI and a revised version

The most common flaw in the exams analyzed was the inclusion of implausible distractors at 79.8% of items. Although initially surprising, this percentage was similar to results in other studies where 90.2% and 100% of items, respectively contained implausible distractors (Haladyna & Downing, 1993; Tarrant & Ware, 2010).

The most common flaws found in the chemistry exams analyzed were different than the most common flaws found in nursing, pharmacy and medical examinations (Pate & Caldwell, 2014; Stagnaro-Green & Downing, 2006; Tarrant et al., 2006; Tarrant & Ware, 2008). In this study the most common flaws were: including implausible distractors (79.8%), uneven answer choice distribution (60.5%), and including incomplete stems (29.3%). When compared to the most common flaws in other studies (Table 5.2), the only overlapping most common flaw was including incomplete stems.

5.5.3 Limitations

The exams analysed as part of this study were not representative of all 1st semester chemistry exams and therefore generalizations about the quality of MC general chemistry exams cannot be made from this study.

Additionally, the IWFEI has been used and validated with a sample of 43 1st semester general chemistry exams, so it is unclear how it will perform with exams from other disciplines or content areas. With that said, we do foresee the IWFEI being able to be used for a wider variety of courses than introductory chemistry, it has just not been used in these contexts yet. However, the criteria or guidelines in the IWFEI have been used in other disciplines, so we do not envision further validation to be an issue. We invite those from other disciplines to use the IWFEI.

We recognize that there are many other facets to consider when creating MC assessments that are not included in this instrument. The IWFEI was not intended to address all aspects of MC assessment design, but to help identify common item writing guideline violations found in the literature. Using this instrument alone does not guarantee an item or exam will perform as desired, although we do foresee its use improving the quality of MC assessment in regards to item writing guideline adherence.

5.6 Conclusions and Implications

In this study, an instrument has been developed that can be used to analyze MC general chemistry exams for item writing guideline violations. The IWFEI can be used by researchers and practitioners to identify ways to improve the quality of MC assessments. We see the IWFEI being used in both research and teaching settings. In research settings, we see the IWFEI being used to evaluate MC chemistry exams for item writing guideline violations as demonstrated in this study. Additionally, we see the IWFEI being used in teaching settings in two ways. First, as a tool to help guide chemistry instructors in the development and revision of their own exams, and secondly, as a tool to be used in professional development settings to advise instructors about MC assessment design. As

the IWFEI is used in these three ways, we envision an improvement in multiple-choice assessment design practice and quality.

CHAPTER 6. CONCLUSIONS AND IMPLICATIONS

This research had two distinct parts that focused on characterizing multiple-choice assessment practices in general chemistry. These two parts included:

- the phenomenographic analysis of what general chemistry instructors consider when they are creating MC exams for their students.
- the development and use of an instrument that can detect item writing guideline violations in MC exams

The conclusions and implications from the development of the IWFEI were outlined at the end of Chapter 5 (see Section 5.6). In summary, the development of this instrument provides a tool for instructors and researchers to use to improve the quality of MC exams and to further study MC assessment in general chemistry.

The other part of this research was an exploratory phenomenographic study that produced data on general chemistry instructors' considerations when designing MC assessments. The results of this study have led to several general conclusions and can inform professional development opportunities and future research in assessment design for chemistry instructors. General conclusions, recommendations for professional development, and recommendations for future research will now be outlined.

6.1 General Conclusions

General chemistry instructors use many literature-supported assessment design practices

Through the analysis of the data, it quickly became apparent that the participating general chemistry instructors used many literature-supported assessment-design practices. Some of these literature-supported practices include:

- A desire to create exams that test a representative range of chosen content
- They value the assessment of data analysis, higher-order thinking skills, and knowledge that is applicable beyond their current course
- They want their items to be easily understood
- They collaborate with each other on assessment design
- They revise, edit, and trial exams and items
- Many use a combination of short answer and MC items
- They consider formatting, significant figures, and units
- They recognize that MC assessments are easy to grade and can be used to assess many concepts quickly
- They recognize that MC assessments are difficult and time-consuming to create

The fact that the instructors interviewed as part of this study considered many appropriate assessment design principles may be surprising in relation to literature that suggests that chemistry instructors are unfamiliar with assessment terminology and are often uncomfortable with educational assessment (Bretz, 2012; Raker & Holme, 2014). This finding that chemistry instructors do consider appropriate assessment design principles can serve as a foundation to be built upon in efforts to assist chemistry instructors to further develop their MC assessment design abilities.

General chemistry instructors are not fully knowledgeable about, nor are fully utilizing, all MC assessment design practices

Another general conclusion that became apparent though analyzing the data was that many of the participating instructors lacked knowledge about assessment design and were not using all the appropriate assessment practices that could benefit them. The instructors demonstrated a lack of understanding of MC assessment design regarding the reasons behind item writing guidelines. Furthermore, many of the instructors were not using test blueprints during their assessment design process, although they desired representative coverage in their exams.

This finding of a lack of understanding is not surprising because chemistry instructors tend to receive little to no formal educational training nor training in assessment (Lawrie et al., 2018). This lack of understanding of assessment design practices can serve as a starting point in the design of professional development and future research opportunities.

6.2 Recommendations for professional development

There are five recommendations for professional development (PD) that have emerged from this research. They will now be outlined.

Recommendation 1: Focus on the reasonings behind MC assessment design principles

As was shown by the data in this study, many of the participating instructors lacked an understanding of assessment design principles such as the number of answer choices to include in an item and the purpose behind item writing guidelines such as avoiding ‘all of the above.’ Because of this, the first recommendation is to focus PD opportunities on informing chemistry instructors of the reasonings behind MC assessment design principles. A deeper understanding of design principles such as only including as many distractors as are plausible, can save chemistry instructors valuable time in the assessment design process. Additionally, a deeper understanding of item writing guidelines may help general chemistry instructors to create valid MC assessments for their courses.

Recommendation 2: Focus on the effective use of test blueprints

The second recommendation for PD is to focus on why and how to use test blueprints during assessment design. Based on data from this study which shows that many of the instructors desired representative exam coverage, it is believed that PD focused on test

blueprint use may be welcomed by chemistry instructors. Additionally, very few of the participants mentioned using a test blueprint even though they desired representative coverage of their learning objectives. Furthermore, many of the instructors wanted to assess higher-order thinking skills in their exams; a desire that could be supported by test blueprint use. Therefore, training chemistry instructors on test blueprint use may aid them in obtaining the representative exam coverage and assessment of higher-order thinking skills that they desire.

Recommendation 3: Focus on how to create effective answer-choices in a time-efficient manner

Many of the instructors mentioned that creating MC exams is a difficult and time-consuming process with the most difficulty arising in the creation of effective answer choices. Because of this, the third recommendation is to focus PD opportunities on how to quickly create plausible answer-choices for MC items. Training on how to do so may be well received by chemistry instructors.

Recommendation 4: Focus on how to evaluate the quality/appropriateness of existing items

The fourth recommendation is to focus PD efforts on how chemistry instructors can evaluate the quality of existing items. This is because many of the instructors used old exams and textbooks as sources of content for their exams. However, some instructors demonstrated a lack of ability to effectively evaluate these items. Additionally, there was a pattern of modifying existing items before including them into exams for their courses. With this strategy being used, it is important to train instructors on how to effectively evaluate the quality of MC items, so they know

how to modify them before inclusion. Of course, “quality” can be defined in many ways including but not limited to: an items adherence to item writing guidelines, level of cognitive skill tested by the item, alignment with course learning objectives or alignment with the Next Generation Science Standards. PD opportunities that can train chemistry instructors to evaluate MC items at some or all of these levels of quality are recommended.

There are some tools that may be able to assist with this effort. The IWFEI, which was developed as part of this work (Chapter 5; Breakall et al., 2019), could be used to help chemistry instructors evaluate existing items for adherence to item writing guidelines. Additionally, the 3D-LAP is an instrument that can be used to check the alignment of test items to the NGSS (Lavery et al., 2016). Furthermore, the cognitive complexity rating tool could be used to evaluate the complexity of existing exam items (Knaus et al., 2011). These tools may be useful to introduce in PD situations and may help chemistry instructors to evaluate the quality of existing exam items.

Recommendation 5: Focus on how to design assessment items that assess higher-order, widely applicable skills may be well received.

The fifth recommendation for PD is to focus on how chemistry instructors can design assessment items that assess higher-order thinking skills and widely applicable knowledge. The participating instructors discussed higher-order thinking skills and widely applicable knowledge in relation to their assessment values, assessment strategies, and what they want to assess in their exams. This demonstrates that many of the participants are interested in and value the assessment of these two areas and thus, PD opportunities focused on the assessment of higher-order thinking skills and widely applicable knowledge may be well received by chemistry instructors.

6.3 Recommendations for future research

Several recommendations for future research emerged from this exploratory study. These recommendations will now be outlined.

Recommendation 1: Investigate the barriers and motivators to test blueprint use

The instructors in this study desired representative exam coverage yet were not using test blueprints during assessment development. Therefore, it is recommended to investigate why some instructors use blueprints while others do not. A qualitative study may be especially effective at investigating the barriers and motivators to test blueprint use. Such research may provide detailed insight into how to motivate chemistry instructors to use test blueprints during exam development and could inform professional development.

Recommendation 2: Use survey methods to gain a representative view of chemistry instructors assessment literacy

This current study has provided in-depth, qualitative data on what chemistry instructors consider when creating MC exams. However, it would also be useful to design a survey to obtain a representative view of chemistry instructors assessment literacy. Such a representative view could be used to validate these findings along with providing further insight into the assessment literacy of chemistry instructors.

Recommendation 3: Design tools that can decrease the time takes to create valid MC exams (especially the answer choices)

In this study many of the instructors mentioned that creating MC assessment (especially the answer choices) is time-consuming and difficult. Therefore, tools that could be used decrease the amount of time it takes to create MC exams may be well received and help chemistry instructors in their assessment design efforts.

Recommendation 4: Evaluate the effectiveness of professional development opportunities

This research has shown that the participating chemistry instructors lacked a deeper knowledge of assessment design practices and test blueprint use. Therefore, it is recommended to study the effectiveness of professional development opportunities on improving chemistry instructors understanding of assessment design and use of test blueprints.

REFERENCES

- Abell, S., & Siegel, M. (2011). Assessment Literacy: What Science Teachers Need to Know and Be Able to Do. In *The Professional Knowledge Base of Science Teaching* (pp. 205–221). <https://doi.org/10.1007/978-90-481-3927-9>
- AERA. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Albanese, M. (1993). Type K and Other Complex Multiple-Choice Items: An Analysis of Research and Item Properties. *Educational Measurement: Issues and Practice*, 12(1), 28–33. <https://doi.org/10.1111/j.1745-3992.1993.tb00521.x>
- Albanese, M., Kent, T., & Whitney, D. (1979). Cluing in Multiple-Choice Test Items with Combinations of Correct Responses. *Journal of Medical Education*, 54, 948–950.
- Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty. *Applied Measurement in Education*, 20(2), 153–170. <https://doi.org/10.1080/08957340701301272>
- Attali, Y. (2003). Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable. *Journal of Educational Measurement*, 40(2), 109–128.
- Austin, A., & Baldwin, R. (1991). *Facility Collaboration: Enhancing the Quality of Scholarship and Teaching*. Washington DC: George Washington University.
- Banerjee, A. C. (1991). Misconceptions of students and teachers in chemical equilibrium. *International Journal of Science Education*, 13(4), 487–494. <https://doi.org/10.1080/0950069910130411>
- Bergner, J., Filzen, J. J., & Simkin, M. G. (2016). Why use multiple choice questions with excess information? *Journal of Accounting Education*, 34, 1–12. <https://doi.org/10.1016/j.jaccedu.2015.11.008>
- Bhattacharyya, G., & Bodner, G. M. (2005). “ It Gets Me to the Product ”: How Students Propose Organic Mechanisms. *Journal of Chemical Education*, 82(9).
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives. Handbook 1: Cognitive Domain*. New York: Longmans Green.
- Board, C., & Whitney, D. R. (1972). The Effect of Selected Poor Item-Writing Practices on Test Difficulty, Reliability and Validity. *Journal of Educational Measurement*, 9(3), 225–233.
- Bodner, G., & Orgill, M. (2007). *Theoretical Frameworks for Research in Chemistry/Science Education* (G. M. Bodner & M. Orgill, Eds.).

- Boud, D. (1995). Assessment and Learning : Contradictory or Complementary? *Assessment for Learning in Higher Education*, 35–48.
- Brandriet, A., & Holme, T. (2015). Development of the Exams Data Analysis Spreadsheet as a Tool to Help Instructors Conduct Customizable Analyses of Student ACS Exam Data. *Journal of Chemical Education*, 92(12), 2054–2061. <https://doi.org/10.1021/acs.jchemed.5b00474>
- Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*, 20(2), 369–382. <https://doi.org/10.1039/c8rp00262b>
- Bretz, S. L. (2010). Faculty Perspectives of Undergraduate Chemistry Laboratory: Goals and Obstacles to Success. *Journal of Chemical Education*, 87(12).
- Bretz, S. L. (2012). Navigating the landscape of assessment. *Journal of Chemical Education*, 89(6), 689–691. <https://doi.org/10.1021/ed3001045>
- Bretz, S. L. (2013). A Chronology of Assessment in Chemistry Education. In *Trajectories of Chemistry Education Innovation and Reform* (pp. 145–153).
- Bretz, S. L. (2014). *Students ' Conceptual Knowledge of Chemistry*.
- Bretz, S. L. (2019). Evidence for the Importance of Laboratory Courses [Editorial]. *Journal of Chemical Education*, 96, 193–195. <https://doi.org/10.1021/acs.jchemed.8b00874>
- Butler, A. C. (2018). Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning? *Journal of Applied Research in Memory and Cognition*, 7(3), 323–331. <https://doi.org/10.1016/j.jarmac.2018.07.002>
- Campbell, M. L. (2015). Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed to Discourage Guessing. *Journal of Chemical Education*, 92(7), 1194–1200. <https://doi.org/10.1021/ed500465q>
- Case, S. M., & Swanson, D. B. (2002). Constructing Written Test Questions For the Basic and Clinical Sciences. In *National Board of Medical Examiners* (Vol. 27). Retrieved from http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf
- Casler, L. (1983). Emphasizing the negative: A note on " not" in multiple-choice questions. *Teaching of Psychology*, 10(1), 51. https://doi.org/10.1207/s15328023top1001_15
- Cassels, J., & Johnstone, A. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, 61(7), 613–615. <https://doi.org/10.1021/ed061p613>

- Chandrasegaran, a. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293. <https://doi.org/10.1039/b7rp90006f>
- Coderre, S., Woloschuk, W., & McLaughlin, K. (2009). Twelve tips for blueprinting. *Medical Teacher*, 31(4), 322–324. <https://doi.org/10.1080/01421590802225770>
- Cooper, M., & Klymkowsky, M. (2013). Chemistry, Life, the Universe and Everything (CLUE): A new approach to general chemistry, and a model for curriculum reform. *Journal of Chemical Education*, 90(9), 1116–1122. Retrieved from <http://gateway.webofknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcAuth=meke ntosj&SrcApp=Papers&DestLinkType=FullRecord&DestApp=WOS&KeyUT=000330097000004%5Cnpapers2://publication/doi/10.1021/ed300456y>
- Cooper, M. M. (2013). Chemistry and the next generation science standards. *Journal of Chemical Education*, 90(6), 679–680. <https://doi.org/10.1021/ed400284c>
- Dekorver, B. K., & Towns, M. (2015). General Chemistry Students' Goals for Chemistry Laboratory Coursework. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.5b00463>
- Dell, K. A., & Wantuch, G. A. (2017). How-to-guide for writing multiple choice questions for the pharmacy instructor. *Currents in Pharmacy Teaching and Learning*, 9(1), 137–144. <https://doi.org/10.1016/j.cptl.2016.08.036>
- DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28(3). <https://doi.org/10.1007/s11092-015-9233-6>
- Dills, C. (1998). The Table of Specifications: A Tool for Instructional Design and Development. *Educational Technology*, 38(3), 44–51.
- Downing, S. M. (2002). Construct-irrelevant Variance and Flawed Test Questions : Do Multiple-choice Item-writing Principles Make Any Difference? *Academic Medicine*, 77(10), 103–104.
- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10(2), 133–143. <https://doi.org/10.1007/s10459-004-4019-5>
- Downing, S. M., Haladyna, T. M., & Rodriguez, M. C. (2010). A Review of Multiple-Choice Item-Writing. *Applied Measurement in Education*, 15(3), 309–333. <https://doi.org/10.1207/S15324818AME1503>

- Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116–121.
<https://doi.org/10.1037/h0035197>
- Duffee, L., & Aikenhead, G. (1992). Curriculum change, student evaluation, and teacher practical knowledge. *Science Education*, *76*(5), 493–506.
<https://doi.org/10.1002/sce.3730760504>
- Duis, J. M. (2011). Organic chemistry educators' perspectives on fundamental concepts and misconceptions: An exploratory study. *Journal of Chemical Education*, *88*(3), 346–350.
<https://doi.org/10.1021/ed1007266>
- Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, *19*(2), 171–179. <https://doi.org/10.1177/001316445901900203>
- Eccles, J. (2013). Subjective task value and the Eccles et al. model of achievement related choices. In A. Elliot & C. Dweck (Eds.), *Handbook of Competence and Motivation* (1st ed.). Gilford Publications.
- Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The Three-option Format for Knowledge and Ability Multiple-choice Tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment*, *20*(1), 65–81.
<https://doi.org/10.1111/j.1468-2389.2012.00580.x>
- Emenike, M. E., Schroeder, J., Murphy, K., & Holme, T. (2013). Results from a national needs assessment survey: A view of assessment efforts within chemistry departments. *Journal of Chemical Education*, *90*(5), 561–567. <https://doi.org/10.1021/ed200632c>
- Emenike, M., Raker, J. R., & Holme, T. (2013). Validating chemistry faculty members' self-reported familiarity with assessment terminology. *Journal of Chemical Education*, *90*(9), 1130–1136. <https://doi.org/10.1021/ed400094j>
- Eubanks, D., & Eubanks, L. (1995). *Writing Tests and Interpreting Test Statistics: A Practical Guide*. ACS DivCHED Examinations Institute.
- Fensham, P. J., & Bellocchi, A. (2013). Higher order thinking in chemistry curriculum and its assessment. *Thinking Skills and Creativity*, *10*, 250–264.
<https://doi.org/10.1016/j.tsc.2013.06.003>
- Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research and Evaluation*, *18*(3), 1–7.
- Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357–364.
<https://doi.org/10.1016/j.tate.2005.01.008>

- Fuhrman, M. (1996). Developing Good Multiple-Choice Tests and Test Questions. *Journal of Geoscience Education*, 44(4), 379–384. Retrieved from papers://dee23da0-e34b-4588-b624-f878b46d7b3d/Paper/p524
- Galyon, C. E., Blondin, C. A., Yaw, J. S., Nalls, M. L., & Williams, R. L. (2012). The relationship of academic self-efficacy to class participation and exam performance. *Social Psychology of Education*, 15(2), 233–249. <https://doi.org/10.1007/s11218-011-9175-x>
- Gibbons, R. E., Reed, J. J., Srinivasan, S., Villafaña, S. M., Laga, E., Vega, J., ... Penn, J. D. (2018). Assessment in Postsecondary Chemistry Education: A Comparison of Course Types. *Assessment Update*, 30(3), 8–11. <https://doi.org/10.1002/au.30131>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gottheiner, D. M., & Siegel, M. A. (2012). Experienced Middle School Science Teachers' Assessment Literacy: Investigating Knowledge of Students' Conceptions in Genetics and Ways to Shape Instruction. *Journal of Science Teacher Education*, 23(5), 531–557.
- Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college biology, chemistry and physics. *Journal of Science Education and Technology*, 19(3), 237–245. <https://doi.org/10.1007/s10956-009-9196-9>
- Haladyna, T. M., & Downing, S. M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement In Education*, 2(1), 37–41. <https://doi.org/10.1207/s15324818ame0201>
- Haladyna, T. M., & Downing, S. M. (1993). How Many Options is Enough For a Multiple-Choice Test Item? *Educational and Psychological Measurement*, 53, 999–1010.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002a). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 7347(April 2011), 37–41. <https://doi.org/10.1207/S15324818AME1503>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002b). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 7347(April 2011), 37–41. <https://doi.org/10.1207/S15324818AME1503>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2010). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Http://Dx.Doi.Org/10.1207/S15324818AME1503_5.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York, NY: Routledge.

- Hansen, J. D., Dexter, L., & Hansen, J. D. (1997). Quality Multiple-Choice Test Questions : Item- -Writing Guidelines and an Analysis of Auditing Testbanks. *Journal of Education for Business*, 73(2), 94–97. <https://doi.org/10.1080/08832329709601623>
- Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of " all of the above" on the reliability and validity of multiple-choice test items. *Evaluation & The Health Professions*, 21(1), 120–133.
- Harasym, P. H., Norris, D. ., & Lorscheider, F. L. (1980). Evaluating Student Multiple-Choice Responses: Effects of Coded and Free Formats. *Evaluation & The Health Professions*, 3(1), 63–84.
- Harasym, Price, Brant, Violato, & Lorscheider. (1992). Evaluation of Negation in Stems of Multiple-Choice Items. *Evaluation & The Health Professions*, 15(2), 198–220.
- Harshman, J., & Yeziarski, E. (2015). Guiding teaching with assessments: high school chemistry teachers ' use of data-driven inquiry †. *Chemistry Education Research and Practice*, 16, 93–103. <https://doi.org/10.1039/C4RP00188E>
- Harshman, J., & Yeziarski, E. (2017). Assessment Data-driven Inquiry: A Review of How to Use Assessment. *Science Educator*, 25(2), 97–107. Retrieved from <http://eds.a.ebscohost.com.proxy.lib.miamioh.edu/eds/detail/detail?vid=1&sid=07dcb7bb-8657-4589-8778-205ee9933588%40sessionmgr4006&hid=4113&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3D%3D#db=eft&AN=121074982>
- Hartman, J. R., & Lin, S. (2011). Analysis of Student Performance on Multiple-Choice Questions in General Chemistry. *Journal of Chemical Education*, 88, 1223–1230.
- Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, 12(2), 184. <https://doi.org/10.1039/c1rp90023d>
- Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, 20(4), 427–441. <https://doi.org/10.1080/08957340701580736>
- Holme, T. (2003). Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education*, 80(6), 171–178. <https://doi.org/10.2174/156802611794863580>
- Holme, T. (2011). Assessment Data and Decision Making in Teaching. *Journal of Chemical Education*, 88(8), 1017–1017. <https://doi.org/10.1021/ed200350w>
- Holme, T., Bretz, S. L., Cooper, M., Lewis, J., Paek, P., Pienta, N., ... Towns, M. (2010). Enhancing the Role of Assessment in Curriculum Reform in Chemistry. *Chemistry Education Research and Practice*, 11(2), 92–97. <https://doi.org/10.1039/c005352j>

- Holsgrove, G., & Elzubeir, M. (1998). Imprecise terms in UK medical multiple-choice questions: What examiners think they mean. *Medical Education*, *32*(4), 343–350. <https://doi.org/10.1046/j.1365-2923.1998.00203.x>
- Huntley, R., & Welch, C. (1993). Numerical Answer Options: Logical or Random Order? *Paper Presented at the Annual of Meeting of the American Educational Research Association, Atlanta, GA.*
- Jacobs, L., & Chase, C. (1992). *Developing and Using Tests Effectively: A Guide for Faculty* (1st Editio). San Francisco: Jossey-Bass Inc.
- Johnstone, A. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning*, *7*(2), 75–83. <https://doi.org/10.1111/j.1365-2729.1991.tb00230.x>
- Johnstone, A. (2006). Chemical education research in Glasgow in perspective. *Chemistry Education Research and Practice*, *7*(2), 49. <https://doi.org/10.1039/b5rp90021b>
- Johnstone, A., & El-Banna, H. (1986). Capacities, demands and processes - a predictive model for science education. *Education in Chemistry*, *23*, 80–84.
- Karthikeyan, S., O'Connor, E., & Hu, W. (2019). Barriers and facilitators to writing quality items for medical school assessments - A scoping review. *BMC Medical Education*, *19*(1), 14–17. <https://doi.org/10.1186/s12909-019-1544-8>
- Kendhammer, L., Holme, T., & Murphy, K. (2013). Identifying Differential Performance in General Chemistry: Differential Item Functioning Analysis of ACS General Chemistry Trial Tests. *Journal of Chemical Education*, *90*, 846–853.
- Knaus, K., Murphy, K., Blecking, A., & Holme, T. (2011). A Valid and Reliable Instrument for Cognitive Complexity Rating Assignment of Chemistry Exam Items. *Journal of Chemical Education*, 554–560.
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometrics Society*, *33*(1), 159–174. <https://doi.org/10.2307/2529310>
- Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., ... Cooper, M. M. (2016). Characterizing college science assessments: The three-dimensional learning assessment protocol. *PLoS ONE*, *11*(9), 1–21. <https://doi.org/10.1371/journal.pone.0162333>
- Lawrie, G. A., Schultz, M., Bailey, C. H., & Dargaville, B. L. (2018). Personal journeys of teachers: an investigation of the development of teacher professional knowledge and skill by expert tertiary chemistry teachers. *Chemistry Education Research and Practice*. <https://doi.org/10.1039/C8RP00187A>

- Lee, C. J. (2018). The test taker's fallacy: How students guess answers on multiple-choice tests. *Journal of Behavioral Decision Making*, (July 2018), 1–12. <https://doi.org/10.1002/bdm.2101>
- Lincon, Y. ., & Guba, E. . (1985). *Naturalistic Inquiry*. Beverly Hills CA: Sage.
- Lord, B. T., & Bavisar, S. (2007). Moving students from information recitation to information understanding: Exploiting Bloom's Taxonomy in creating science questions. *Journal of College Science Teaching*.
- Magnusson, S, Krajcik, J., & Borko, H. (1999). Nature, sources and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. Lederman (Eds.), *Examining Pedagogical Content Knowledge* (pp. 95–132). Kluwer Academic Publishers.
- Magnusson, Shirley, Krajcik, J., & Borko, H. (1999). Nature, Sources, and Development of Pedagogical Content Knowledge for Science Teaching. In *PCK and Science Education* (pp. 95–132).
- Marton, F. (1981). Phenomenography - Describing Conceptions of the World Around Us. *Instructional Science*, 10, 177–200.
- Marton, F. (1986). Phenomenography: A Research Approach to Investigating Different Understandings of Reality. *Journal of Thought*, 21(3), 28–49.
- McClary, L. M., & Bretz, S. L. (2012). Development and Assessment of A Diagnostic Tool to Identify Organic Chemistry Students' Alternative Conceptions Related to Acid Strength. *International Journal of Science Education*, 34(15), 2317–2341. <https://doi.org/10.1080/09500693.2012.684433>
- Mcmorris, R. F., Boothroyd, R. A., & Pietrangelo, D. J. (1997). Humor in Educational Testing : A Review and Discussion. *Applied Measurement In Education*, 10(3), 269–297. <https://doi.org/10.1207/s15324818ame1003>
- Mertler, C. A. (1999). Assessing Student Performance: A Descriptive Study of the Classroom Assesment Practices of Ohio Teachers. *Education*, 120(2), 285–296.
- Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Applied Measurement in Education*, 22(1), 38–60. <https://doi.org/10.1080/08957340802558342>
- Millman, J., & Bishop, C. H. (1965). An Analysis of Test-Wiseness. *Educational and Psychological Measurement*, XXV(3), 707–726.
- Moreno, R., & Marti, R. J. (2006). New Guidelines for Developing Multiple-Choice Items. *Methodology European Journal of Research Methods for the Behavioral & Social Sciences*, 2(2), 65–72. <https://doi.org/10.1027/1614-1881.2.2.65>

- Murphy, K., Holme, T., Zenisky, A., Caruthers, H., & Knaus, K. (2012). Building the ACS exams anchoring concept content map for undergraduate chemistry. *Journal of Chemical Education*, 89(6), 715–720. <https://doi.org/10.1021/ed300049w>
- Naeem, N., van der Vleuten, C., & Alfari, E. A. (2012). Faculty development on item writing substantially improves item quality. *Advances in Health Sciences Education*, 17(3), 369–376. <https://doi.org/10.1007/s10459-011-9315-2>
- Nakhleh, M. M. B. (1993). Are our students conceptual thinkers or algorithmic problem solvers? Identifying conceptual students in general chemistry. *Journal of Chemical Education*, 70(1), 52–55. <https://doi.org/10.1021/ed070p52>
- Next Generation Science Standards : For States, By States.* (2013). National Academies Press.
- Next Generation Science Standards: For States, By States.* (2013). Washington DC: National Academy of Sciences.
- Niaz, M. (1987). Relation between M-Space of Students and M-Demand of Different Items of General-Chemistry and Its Interpretation Based Upon the Neo-Piagetian Theory of Pascual-Leone. *Journal of Chemical Education*, 64(6), 502–505. <https://doi.org/10.1021/ed064p502>
- Niaz, M. (1989). The relationship between M-demand, algorithms, and problem solving: A neo-Piagetian analysis. *Journal of Chemical Education*, 66(5), 422. <https://doi.org/10.1021/ed066p422>
- Orgill, M., & Sutherland, A. (2008). Undergraduate chemistry students ' perceptions of and misconceptions about buffers and buffer problems. *Chemistry Education Research and Practice*, 131–143.
- Papenberg, M., & Musch, J. (2017). Of Small Beauties and Large Beasts: The Quality of Distractors on Multiple-Choice Tests Is More Important Than Their Quantity. *Applied Measurement in Education*, 30(4), 273–286. <https://doi.org/10.1080/08957347.2017.1353987>
- Pate, A., & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning*, 6(1), 130–134. <https://doi.org/10.1016/j.cptl.2013.09.003>
- Patton, M. Q. (2015). *Qualitative Research and Evaluation Methods* (4th ed.). Thousand Oaks, California: Sage.
- Pellegrino, J. W. (2001). Knowing What Students Know. In *National Academy of the Sciences*. <https://doi.org/10.17226/10019>
- Plake, B. (1984). Can Relevant Grammatical Cues Result In Invalid Test Items. *Educational and Psychological Measurement*.

- Presley, M., & Hanuscin, D. (2015). *INVESTIGATING HOW PARTICIPATORY ACTION RESEARCH AND THE USE OF ASSESSMENT INSTRUMENTS CAN SUPPORT COLLEGE INSTRUCTORS' SCIENCE ASSESSMENT LITERACY*. University of Missouri.
- Raker, J. R., Emenike, M. E., & Holme, T. A. (2013). Using structural equation modeling to understand chemistry faculty familiarity of assessment terminology: Results from a national survey. *Journal of Chemical Education*, *90*(8), 981–987. <https://doi.org/10.1021/ed300636m>
- Raker, J. R., & Holme, T. A. (2014). Investigating faculty familiarity with assessment terminology by applying cluster analysis to interpret survey data. *Journal of Chemical Education*, *91*(8), 1145–1151. <https://doi.org/10.1021/ed500075e>
- Raker, J. R., Trate, J. M., Holme, T. A., & Murphy, K. (2013). Adaptation of an Instrument for Measuring the Cognitive Complexity of Organic Chemistry Exam Items. *Journal of Chemical Education*, 130918144937002. <https://doi.org/10.1021/ed400373c>
- Reed, J. J., Brandriet, A. R., & Holme, T. A. (2016). Analyzing the Role of Science Practices in ACS Exam Items. *Journal of Chemical Education*, acs.jchemed.6b00659. <https://doi.org/10.1021/acs.jchemed.6b00659>
- Reed, J. J., Villafañe, S. M., Raker, J. R., Holme, T. A., & Murphy, K. L. (2017). What We Don't Test: What an Analysis of Unreleased ACS Exam Items Reveals about Content Coverage in General Chemistry Assessments. *Journal of Chemical Education*, acs.jchemed.6b00863. <https://doi.org/10.1021/acs.jchemed.6b00863>
- Regan, T. (2015). Item-Writing Guidelines for Physics. *The Physics Teacher*, *53*(8), 485–487. <https://doi.org/10.1119/1.4933152>
- Rimland, B. (1960). The Effect of Including Extraneous Numerical Information in a Test of Arithmetic Reasoning. *Educational and Psychological Measurement*, *(4)*, 787–794.
- Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues & Practice*, *24*(2), 3–13.
- Saldana, J. (2016). *The Coding Manual for Qualitative Researchers* (3rd ed.; J. Seaman, Ed.). SAGE.
- Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving Teachers' Assessment Practices Through Professional Development: The Case of National Board Certification. *American Educational Research Journal*, *45*(3), 669–700. <https://doi.org/10.3102/0002831208316955>
- Schafer, W. D. (1991). Essential Assessment Skills in Professional Education of Teachers. *Educational Measurement: Issues and Practice*, *10*(1), 3–6. <https://doi.org/10.1111/j.1745-3992.1991.tb00170.x>

- Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics and standard setting. *Medical Education*, *48*(10), 1020–1027. <https://doi.org/10.1111/medu.12525>
- Schrock, T. J., & Mueller, D. J. (1982). Effects of Violating Three Multiple-Choice Item Construction Principles. *The Journal of Educational Research*, *75*(5), 314–318. <https://doi.org/10.1080/00220671.1982.10885401>
- Schroeder, J., Murphy, K. L., & Holme, T. A. (2012). Investigating factors that influence item performance on ACS exams. *Journal of Chemical Education*, *89*(3), 346–350. <https://doi.org/10.1021/ed101175f>
- Schurmeier, K. D., Atwood, C. H., Shepler, C. G., & Lautenschlager, G. J. (2010). Using item response theory to assess changes in student performance based on changes in question wording. *Journal of Chemical Education*, *87*(11), 1268–1272. <https://doi.org/10.1021/ed100422c>
- Siegel, M. A., & Wissehr, C. (2011). Preparing for the Plunge : Preservice Teachers ' Assessment Literacy. *Journal of Science Teacher Education*, *22*(4), 371–391. <https://doi.org/10.1007/s10972-011-9231-6>
- Smith, K. C., Nakhleh, M. B., & Bretz, S. L. (2010). An expanded framework for analyzing general chemistry exams. *Chem. Educ. Res. Pract.*, *11*(3), 147–153. <https://doi.org/10.1039/C005463C>
- Srinivasan, S., Reisner, B. A., Smith, S. R., Stewart, J. L., Johnson, A. R., Lin, S., ... Raker, J. R. (2018). Historical Analysis of the Inorganic Chemistry Curriculum Using ACS Examinations as Artifacts. *Journal of Chemical Education*, *95*(5), 726–733. <https://doi.org/10.1021/acs.jchemed.7b00803>
- Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher*, *28*(6), 566–568. <https://doi.org/10.1080/01421590600711153>
- Stark, L. A. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology. *CBE Life Sciences Education*, *8*(1), 1–6. <https://doi.org/10.1187/cbe.08>
- Stiggins, R. (1991). Asssment Literacy. *Phi Delta Kappa International*, *72*(7), 534–539.
- Svensson, L. (1997). Theoretical Foundations of Phenomenography Theoretical Foundations of Phenomenography. *Higher Education Reserach & Development*, *16*(2), 159–171. <https://doi.org/10.1080/0729436970160204>
- Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation*, *19*(3), 311–325. [https://doi.org/10.1016/S0191-491X\(05\)80013-6](https://doi.org/10.1016/S0191-491X(05)80013-6)

- Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching*, 39(4), 283–301. <https://doi.org/10.1002/tea.10023>
- Tarrant, Knierim, Hayes, & Ware. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 26(8), 662–671. <https://doi.org/10.1016/j.nedt.2006.07.006>
- Tarrant, & Ware. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2), 198–206. <https://doi.org/10.1111/j.1365-2923.2007.02957.x>
- Tarrant, & Ware. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today*, 30(6), 539–543. <https://doi.org/10.1016/j.nedt.2009.11.002>
- Tarrant, Ware, & Mohammed. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, 9, 40. <https://doi.org/10.1186/1472-6920-9-40>
- Thompson, A. R., & O'Loughlin, V. D. (2015). The Blooming Anatomy Tool (BAT): A discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. *Anatomical Sciences Education*, 8(6), 493–501. <https://doi.org/10.1002/ase.1507>
- Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and Evaluation in Psychology and Education* (8th Editio). Pearson.
- Tomanek, D., Talanquer, V., & Novodvorsky, I. (2008). What do science teachers consider when selecting formative assessment tasks? *Journal of Research in Science Teaching*, 45(10), 1113–1130. <https://doi.org/10.1002/tea.20247>
- Towns, M. (2010). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education*, 87(1), 91–96. <https://doi.org/10.1021/ed8000039>
- Towns, M. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, 91(9), 1426–1431. <https://doi.org/10.1021/ed500076x>
- Towns, M., & Robinson, W. R. (1993). Student Use of Test-Wiseness Strategies in Solving Multiple-Choice Chemistry Examinations. *Journal of Research in Science Teaching*, 30(7), 709–722. <https://doi.org/10.1002/tea.3660300709>
- Trigwell, K. (2000). Phenomenography: Variation and Discernment. *Intertational Symposium Oxford Centre for Staff and Learning Development*, 75–85. Oxford UK.

- Tsaparlis, G., & Angelopoulos, V. (2000). A model of problem solving: Its operation, validity, and usefulness in the case of organic-synthesis problems. *Science Education*, 84(2), 131–153. [https://doi.org/10.1002/\(SICI\)1098-237X\(200003\)84:2<131::AID-SCE1>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1098-237X(200003)84:2<131::AID-SCE1>3.0.CO;2-4)
- Vanmali, B., & Siegel, M. (2012). *ASSESSING ASSESSMENT : HOW USE OF THE CONCEPT INVENTORY OF NATURAL SELECTION INFLUENCES THE INSTRUCTIONAL PRACTICES OF AN EXPERIENCED BIOLOGY PROFESSOR AND SUPPLEMENTAL Instruction Leader*. University of Missouri-Columbia.
- Viera, A. J., & Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, (May), 360–363.
- Wakefield, J. A. (1958). Does the fifth choice strengthen a test item? *Public Personnel Review*, 19, 44–48.
- Weiten, W. (1984). Violation of selected item construction principles in educational measurement. *The Journal of Experimental Educational*, 52(3), 174–178. <https://doi.org/10.1080/00220973.1984.11011889>
- Weitzman, R. A. (1970). Ideal Multiple-Choice Items. *Journal of the American Statistical Association*, 65(329), 71–89.
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice : A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Young, K., Lashley, S., & Murray, S. (2019). Influence of Exam Blueprint Distribution on Student Perceptions and Performance in an Inorganic Chemistry Course. *Journal of Chemical Education*. <https://doi.org/10.1039/C9RP00112C>
- Zhicheng, Z., & Burry-Stock, J. (2003). Classroom Assessment Practices and Teachers' Self-Perceived Assessment Skills. *Applied Measurement in Education*, 16, 323–342. <https://doi.org/10.1207/S15324818AME1604>

APPENDIX A

Coding Scheme

Name	Description
Assessment Values and Principles	Statements about overarching ideas and beliefs that guide decisions for chemistry assessments. Usually accompanied by “because” or a reason for the practice. “ I do this because...”
Algorithmic questions should be free response; conceptual should be multiple-choice	Statements that reflect a belief that in an exam algorithmic/mathematical questions should be free-response and conceptual questions should be multiple-choice
Assessment as a motivator	Statements that reflect a belief that assessments can be used to motivate students to learn
Equitable items for all learners	Statements that reflect a philosophy whereby all items should be equitable to all learners
Exam coverage	Statements that reflect a belief about the range of content that should be tested on an exam
Exam Security	Statements that reflect a belief that effort should be taken to minimize cheating
Familiar Item Format	Statements that reflect a belief that items should be presented in a way that students are accustomed to. This includes the use of familiar vocabulary.
Item Validity	Statements that reflect a belief that items should assess only the desired skill and should avoid assessing extraneous skills.
Learning Objectives	Statements that reflect a belief that learning objectives are a main guide to creating assessments

Name	Description
Limitations of multiple-choice exams	Statements that reflect a belief that multiple-choice exams have limiting factors
Models of Learning	Statements that refer to using a model of learning to guide decisions on making assessments or assessment items. Models of learning could include: Johnstone's triangle model of learning. This model incorporates symbolic, macroscopic, and sub-microscopic considerations. Often referred to as the chemistry triplet or Johnstone's triangle. The three-dimensional learning model proposed by the (NGSS). The three dimensions include, crosscutting concepts, science/engineering practices, and disciplinary core ideas.
Number of concepts an item should test	Statements that reflect a belief about the number of concepts an item should test. Often accompanied by "because". Ex. I wrote the item "because" it tested X number of concepts.
Obviously implausible distractors	Statements that reflect a belief about including obviously wrong answer choices in an exam
Students can learn from assessment	Statements that reflect a belief that students can learn from assessments.
Students should analyze data	Statements that reflect a belief that students should analyse data during an exam
Testing levels of understanding	Statements that reflect a belief that assessing different/more complex levels of understanding is important in an exam.
Trick questions and or testing exceptions	Statements that reflect a belief about trick questions or questions that test exceptions to a rule
Widely Applicable	Statements that reflect a belief that assessments should test skills and knowledge that are useful beyond the current course.

Name	Description
External Influences	Factors that influence how an instructor chooses to assess their students that are outside of themselves. For example, departmental regulations, relationships, or traditions, institutional regulations, relationships, or traditions.
Knowledge of Assessment Interpretation and Action-Taking	What an instructor knows about interpreting and acting upon assessment data
Review item statistics	Statements that refer to reviewing item statistics to assess student progress or how the exam performed. This could include statistics such as item difficulty or discrimination.
Use of exam data	Statements that refer to the use of exam data. This would include using exam data to change teaching practices or to understand student misconceptions.
Knowledge of Assessment Purposes	Why an instructor chooses to assess students
Use MC assessment data to shape instructional decisions	Statements referring to an instructor's belief that the purpose of assessment should include informing instructional decisions
Knowledge of Assessment Strategies	Statements that reflect the way an instructor assesses student learning or designs assessment tasks
Algorithmic questions should be free response; conceptual should be multiple-choice	Statements that refer to using free-response format for algorithmic/mathematical questions and/or multiple-choice format for conceptual questions
Answering time	Statements that reflect considerations for how long it takes students to answer an item or an exam as a whole

Name	Description
Assessment training	Statements that refer to training or mentoring opportunities about how to create multiple-choice exams
Benefits and Disadvantages of MC Assessment	Statements that refer to the benefits and disadvantages of multiple-choice assessment. This may include things such as their ability to address many concepts in an exam, grading, reliability, the difficult and time consuming nature of creating them, etc.
Collaboration	Statements that refer to working with others (i.e. colleagues) to create exams
Consider the types of cognitive skills tested	Statements that reflect the cognitive actions required by students to complete an item. For example, an instructor may refer to an item as being simple recall, higher-order, etc.
Creating multiple exam or item versions	Statements that refer to creating multiple versions of an exam or item
Exam item properties	
Exam Overall	
Item order	Statements about the order of the items in an exam. For instance, ordering items from easy to hard, grouping similar items together, etc.
Key balancing	Statements that refer to varying where the correct answer is keyed in a multiple choice exam.
Item Overall	
Amount of Information in an Item (Conciseness)	Statements about the length/wordiness of an item. For example, statements about including extra information to provide more context or statements about there being too much information in an item.

Name	Description
Formatting consistencies	Statements that reflect the need for consistent formatting, phrasing, and stylistic components of items
Grammatical and phrasing cues	Statements about how inconsistencies in the grammar or phrasing of item test could clue students to improved performance.
Homework questions influence test item	Statements about how homework questions/assignments influence test items
Item Clarity	Statements about how easy/difficult it is to understand the phrasing of an item or what the item is asking the student to do.
Item format	Statements that refer to the type or format of a multiple choice item. Item formats may include, k-type items, select the one correct answer type, select all that are correct type, two-tier MC items, etc. K-type items have a format where there are primary choices of say I, II, III, and IV. Then students are asked to choose between combinations of those answer choices. For example, A. I and II B. III and IV C. II only D. I, II, and IV
Number of concepts tested in an item	Statements that discuss the number of concepts that should be tested by an item
Number of thinking steps	Statements that refer to the number of cognitive steps needed to answer an item. For example, If a student needed to find the molar mass and then calculate the number of moles, that could be viewed as two thinking steps.
Only test the targeted skill or objective	Statements that refer to testing only the desired skill or objective. This also refers to eliminating skills or objectives from items that aren't desired to be tested.

Name	Description
Significant figures	Statements that refer to significant figures
Units	Statements referring to the use of units in multiple-choice items
Response Set	
All_None of the Above	Statements about using 'all of the above' or 'none of the above' as an answer choice
Answer choice order	Statements about the order of answer choices in a multiple choice item. For example, ordering them from least to greatest or in a logical order.
Distractors	Statements that refer to the distractors in a multiple-choice item. This can include statements about: the number of distractors, the plausibility of distractors, making distractors using errors or misconceptions, making distractors from responses to free-response items, etc.
Length of answer choices	Statements about the respective length of the answer choices
Symmetry and Similarity	Statements that refer to symmetry or similarity in the answer choices.
Stem	
Complete problem statement in the stem	Statements about the completeness of the stem. Does the stem contain an understandable question or problem?
Negative phrasing	Statements about the inclusion or exclusion of negative phrasing in an item
Internet	Statements that refer to the instructor using internet resources to help design items or to check content accuracy.

Name	Description
Item based on how a topic was taught	Statements that reflect basing an item not just on what was taught but the way it was taught
Item difficulty	Statements about considering item difficulty when designing an exam
Item exam preparation timing	Statements about when an instructor prepares an exam or exam items for their course
Lack of knowledge of MC assessment design	Statements that reflect an instructors lack of understanding of the design of multiple-choice items and exams.
Novel content	Statements or actions that refer to using problems that are unfamiliar to students in exam items. Unfamiliar refers to problems and not concepts.
Revise, edit, and trial	Statements about revising, editing, and trialing exams or exam items before administering it to the students
Students prior knowledge and expectations	Considerations and/or expectations of what students should know to be able to answer an item
Using multiple types of assessment items	Statements that refer to the use of different types of items on the same exam or in the same course. (Short answer vs. multiple-choice)
Knowledge of What to Assess	How/what an instructor determines to assess
Determining what to assess	Statements about how an instructor determines what to put on an assessment. This could include considerations of: homework, learning objectives, lecture or lecture notes, level understanding tested, number of thinking steps in an item, previous item statistics, topic importance, what students need to demonstrate competency of, etc.

Name	Description
Finding the items	Statements about where an instructor finds items for an exam. This could include: adapting previously administered items, creating items personally, using previously administered items directly, adapting items from standardized exams such as the MCAT or ACS exams, finding or adapting items from a textbook, finding or adapting items from the internet, using items from when the instructor was a student, etc.
What to assess	Statements about what an instructor chooses to assess in an exam. These may include choosing to assess: conceptual understanding, data analysis skills, lab skills, misconceptions, unfamiliar content, understanding of different representations, the "why" behind a topic, widely applicable skills or knowledge, representative coverage of the material, etc.
Views of Learning	How an instructor views student learning

APPENDIX B

Table of Contents

Introduction to Multiple-Choice Item Format	Page 2
Is the Test Item Clear and Succinct?	Page 3
If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded?	Page 4
If the answer choices are numerical: Are they listed in ascending or descending order?	Page 5
If the answer choices are verbal: Are the answer choices all approximately the same length?	Page 6
Does the item avoid “all of the above” as a possible answer choice?	Page 7
Does the item avoid grammatical and phrasing cues?	Page 8
Could the item be answered without looking at the answer choices?	Page 9
Does the item avoid complex K-type item format?	Page 10
Is this item linked to one or more objectives of the course?	Page 11
Are all answer choices plausible?	Page 12
Are there six or less thinking steps needed to solve this problem?	Page 13
Does the exam avoid placing three or more items that assess the same concept next to each other?	Page 14
Does the exam avoid placing three or more difficult items next to each other?	Page 14
Is there an even distribution of correct answer choices?	Page 14
Does the exam avoid linking performance on one item with performance on others?	Page 15
References	Page 16

Introduction to Multiple Choice Item Format

_____ 9. How many moles of K^+ ions are in 30 mL of 0.60 M K_3PO_4 ?

(a) **0.054**

(b) 0.042

(c) 0.036

(d) 0.018

(e) 0.006

Labels and arrows:

- Item: points to the entire question box.
- Stem of the Item: points to the question text.
- Answer choices: points to the list of options (a) through (e).
- Correct response (keyed answer): points to option (a).
- Distractors: points to options (b) through (e).

Is the test item clear and succinct?

- The stem can only be interpreted as having one meaning.
- The stem doesn't include any extra information or wording (**Needed context is appropriate**).
- The answer choices don't include any extra information or wording
- There is clearly only one correct answer choice

Good Example

- _____ 1. The atomic weight of silicon is 28.0855. Round this number to 4 significant figures.
- (a) 28.0
(b) 28.08
(c) 28.09
(d) 28.086
(e) 28.1

The stem has only one interpretation with no extra information.

The answer choices don't include any extra information and only one answer choice can be interpreted as correct.

Poor Example

- _____ 6. It takes 19 days for a particular nuclide to decay 30% of its original activity. What is the half-life of this nuclide?
- (a) It would take 0.44 days
(b) It would take 11 days
(c) It would take 16 days
(d) It would take 27 days
(e) It would take 37 days

The stem could be interpreted as decaying from 100% to 70% or as decaying from 100% to 30%. This makes the question unclear.

The answer choices are not as succinct as possible. They include extra information/wording. "It would take" could be removed.

Poor Example

- _____ 3. Aspirin is a pain killer that has a density of 1.40 g/cm³. What is the amount (in moles) of aspirin, C₉H₈O₄, in a 325 mg tablet that is 100% aspirin?
- (a) 0.00180 mol
(b) 0.00325 mol
(c) 0.467 mol
(d) 1.80 mol
(e) 2.80 mol

The density of Aspirin is extra information that is not needed to solve the problem. This introduces student ability to determine needed information as a variable in student performance. The question is no longer just testing the intended chemistry content.

If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?

- The words “not” or “except” should be bolded or capitalized if included in the item.
- Avoiding “not” or “except” is ideal in most cases.

Good Example

_____ 1. Which of the following contains a triple bond?

- ethylene
- ethane
- propene
- benzene
- propyne

← The stem of this question doesn't contain negative phrasing. This is ideal for most items.

Good Example

_____ 1. Which of the following does **NOT** contain a triple bond?

- Butyne
- Pentyne
- Hexyne
- Benzene
- Propyne

← The negative phrase is capitalized.

Poor Example

_____ 1. All of the following processes are exothermic **except**:

- Combustion of propane
- Rusting of iron
- Freezing of water
- Melting of ice

← The word 'except' is not bolded or capitalized.

If the answer choices are numerical:

Are they listed in ascending or descending order?

- Numerical answer choices should be listed in ascending or descending order. For example: 1,2,3 vs. 2,1,3

Good Example

_____ 12. What is the molality of a solution prepared by mixing 12.0 g benzene (C_6H_6) with 38.0 g CCl_4 ?

- a. 0.240 *m*
- b. 0.316 *m*
- c. 0.508 *m*
- d. 0.622 *m*
- e. 4.05 *m*

The answer choices are written in ascending numerical order.

Poor Example

_____ 12. What is the molality of a solution prepared by mixing 12.0 g benzene (C_6H_6) with 38.0 g CCl_4 ?

- a. 4.05 *m*
- b. 0.240 *m*
- c. 0.622 *m*
- d. 0.316 *m*
- e. 0.508 *m*

The answer choices are not written in ascending or descending numerical order.

This is *Not Applicable* if an item is K-type

Symbolic answer choices, such as electron configurations or chemical formulas are NOT considered numerical.

This criterion would be Not Applicable.

If the answer choices are verbal:

Are the answer choices all approximately the same length?

- An answer choice should **not** be substantially longer or shorter than any of the other choices. This may cue students to an answer without consideration of the item content.

Good Example

_____ 19. What is the purpose of standardizing a solution?

- a. To determine its purity.
- b. To determine its concentration.
- c. To measure its volume.
- d. To determine its molecular formula.
- e. To determine the endpoint.

This item keeps all answer choices approximately the same length.

Poor Example

_____ 19. What is the purpose of standardizing a solution?

- a. To determine its purity.
- b. The purpose is to determine the concentration of the solution
- c. To measure its volume.
- d. To determine its molecular formula.
- e. To determine the endpoint.

One answer choice is significantly longer than the others.

(This item includes phrasing cues as well (see page 8))

This is *Not Applicable* if an item is K-type

Symbolic answer choices, such as electron configurations or chemical formulas are NOT considered verbal.

This criterion would be *Not Applicable* if answer choices are symbolic.

Does the item avoid “all of the above” as a possible answer choice?

- Using “all of the above” as an answer choice can cue students to eliminate distractors.

Good Example

- _____ 1. Which of the following contains a triple bond?
- a. ethylene
 - b. ethane
 - c. propene
 - d. benzene
 - e. propyne

This item doesn't use all of the above or none of the above as answer choices

Poor Example

- _____ 1. Which of the following contains a triple bond?
- a. ethylene
 - b. ethane
 - c. propene
 - d. propyne
 - e. all of the above

The use of 'all of the above' is quickly eliminated when a student recognizes any molecule that doesn't contain a triple bond.

For K-type items, if an answer choice includes all of the possibilities, then it violates this guideline.

Does the item avoid grammatical and phrasing cues?

- A cue leads a student to the right answer or to eliminating a distractor.
- A grammatical cue is a difference in grammar between the stem and the answer choices or between answer choices.
- A phrasing cue is where a phrase from the stem is used in one distractor or in the correct answer.

Good Example (Grammatical cuing)

__ 19. Carbon has ____ proton(s).

(a) One
(b) Three
(c) Six
(d) Twelve

← This item keeps the grammar of the stem consistent with the answer choices.

Poor Example (Grammatical cuing)

__ 19. Carbon has ____ protons?

(a) One
(b) Three
(c) Six
(d) Twelve

← Answer choice A does not fit the grammatical structure of the stem. This may cue students to it being the incorrect answer.

Good Example (Phrasing cues)

__ 19. How many proton(s) does Carbon have?

(a) One
(b) Three
(c) Six
(d) Twelve

← This item gives no phrasing cues to the correct answer.

Poor Example (Phrasing cues)

__ 19. How many proton(s) does Carbon have?

(a) One
(b) Three
(c) Six protons
(d) Twelve

← This item using a phrase (protons) from the stem in the correct answer choice. This may cue students to choose this answer.

Could the question be answered without looking at the answer choices?

- It is important to write the stem of an item in a way that it could be answered without looking at the answer choices. This ensures that the central idea is included in the stem.

Good Example

__ 19. Hydrogen can have how many protons?

(a) 0 or 1
(b) 1
(c) 1 or 2
(d) 2

This item contains the central idea in the stem and can be answered without the answer choices.

Poor Example

__ 19. Hydrogen:

(a) can have 0 or 1 protons
(b) can only have 1 proton
(c) can have 1 or 2 protons
(d) can only have 2 protons

This item cannot be answered without looking at the answer choices. The stem doesn't contain the central idea.

Does the item avoid complex K-type item format?

- K-Type items have answer choices that contain combinations of other answer choices.
- K-Type Items have been shown to cue students to the correct answer
- Ordering items, such as the ordering of ion-size, are not considered to be K-type.

Good Example

- _____ 12. What is the molality of a solution prepared by mixing 12.0 g benzene (C_6H_6) with 38.0 g CCl_4 ?
- 4.05 *m*
 - 0.240 *m*
 - 0.622 *m*
 - 0.316 *m*
 - 0.508 *m*

This item avoids K-type format

Poor Example

- _____ 5. Which of the following properties influence the frequency of a molecular vibration, seen in infrared absorption spectra?
- Size (radius) of the atoms on each side of the bond
 - Strength of the bond between atoms
 - Mass of the atoms on each side of the bond
- i only
 - ii only
 - iii only
 - i and ii
 - ii and iii

This is an example of a K-type question

Good Example

- _____ 17. A double bond is composed of _____ bond(s) and _____ rotate.
- Two sigma; cannot
 - Two pi; cannot
 - One sigma and one pi; cannot
 - One sigma and one pi; can
 - Two sigma; can

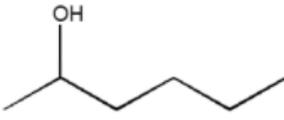
This is a fill-in-the-blank item.
This is NOT in k-type format.

Is this item linked to one or more objectives of the course?

- Test items should test one or more objectives of the course.

Good Example:

_____ 3. What is the correct formula for the organic molecule shown below?



a. $C_7H_{14}O$
b. $C_6H_{14}O$
c. $C_7H_{13}O$
d. $C_6H_{13}O$
e. $C_4H_{10}O$

Hypothetical Course Objectives

Students Should Be Able to:

1. Interconvert between skeleton structures and chemical formulas.
2. Determine the number of atoms in a molecule based on various representations.
3. Draw Lewis Dot Diagrams from molecular formulas.

This item directly assesses course objective 1 and indirectly assesses objective 2.

Poor Example:

_____ 2. Alkenes by definition contain a _____.

- a. C=C bond
- b. C≡C bond
- c. C-C bond
- d. C=H bond
- e. C≡H bond

This item doesn't assess any of listed the course objectives.

Are there six or less thinking steps needed to solve this problem?

- A thinking step is a small cognitive process that must be taken to solve a problem (Johnstone & El-Banna, 1986).
- The thinking steps should be based on the average student taking the exam.

The following is an example of the thinking steps that may exist in an item. Reproduced from (Johnstone & El-Banna, 1989) with permission from the Royal Society of Chemistry.

'What volume of molar hydrochloric acid would be exactly neutralized by ten grams of chalk?'

Thinking Steps:

1. chalk---calcium carbonate (recall)
2. calcium carbonate = CaCO_3 (recall or deduce)
3. Formula weight of CaCO_3 = 100 g (calculate)
4. When it reacts with hydrochloric acid, what are the products? (recall)
5. Write a balanced equation (transformation)
6. Recognize that 1 mole CaCO_3 ~ 2 moles HCl (deduce)
7. = 2 litres of molar HCl (recall)
8. 10 g CaCO_3 ~ 1/10 mole ~ 1/5 mole HCl (deduce)
9. ~ 1/5 litre molar HCl (recall) = 200 ml molar HCl

Because this item can be viewed as having nine thinking steps, it may be measuring working memory capacity along with the students understanding of chemistry. This negatively effects the validity of the item.

Does the exam avoid placing three or more items that assess the same concept or skill next to each other?

- Placing three or more similar questions next to each other may cue students to what the correct answer may be.
- A concept or skill is defined as the same learning objective.

Does the exam avoid placing three or more difficult items next to each other?

- A difficult item is defined as an item that you believe less than 50% of students will get correct.

Is there an approximately even distribution of correct answer choices?

- Correct answer choices should be approximately evenly distributed. No two distractors should have a difference of greater than two in frequency appearing in the key.

$$\text{Even Distribution} = \frac{i_t}{a} \pm 1 \text{ (for each answer choice)}$$

i_t = Total number of items in the exam

a = Answer choices per item

Good Example

Answer Key 4 choices per item

1) A

2) C

3) C

4) D

Poor Example

Answer Key 4 choices per item

1) A

2) C

3) C

4) C

Does the exam avoid linking performance on one item with performance on others?

- Items should be independent of one another on an exam.

Good Example

1. What is the molar mass of $C_6H_{12}O_6$?
 - a) 168.2
 - b) 180.2
 - c) 200.2
2. How many moles of water are there in a 14.0 gram sample of H_2O ?
 - a) 1.00
 - b) 0.780
 - c) 0.550

Students can do well on each item independent of each other.

Poor Example

1. What is the molar mass of $C_6H_{12}O_6$?
 - a) 168.2
 - b) 180.2
 - c) 200.2
2. Based your answer to question one, would a 0.5 mole sample of $C_6H_{12}O_6$ weigh more or less than 90.0 grams?
 - a) More
 - b) Less
 - c) Not enough information to tell

Students **cannot** succeed on item two if they don't succeed on item one.

Are all answer choices plausible?

- All distractors should be made by using common student errors or misconceptions. Even if only one distractor is not, then the item is in violation of this guideline.
- Each distractor should have been chosen by more than 5% of the students tested.

Good Example

_____ 6. What kind of electromagnetic radiation is able to break bonds?

- a. **Ultraviolet**
- b. Infrared
- c. Visible
- d. Microwave
- e. Radiowaves

All the answer choices are likely to be chosen. They are all viable forms of electromagnetic radiation

Poor Example

_____ 6. What kind of electromagnetic radiation is able to break bonds?

- a. **Ultraviolet**
- b. Infrared
- c. Visible
- d. Microwave
- e. The bonds of friendship are too strong to break.

Answer choice E is not plausible.

APPENDIX C

_____ 1. A wooden boat discovered just south of the Great Pyramid in Egypt has 72.5% of the original carbon-14 expected. The half-life of carbon-14 is 5,730 years. How old is the boat?

- (a) 4,154 years
- (b) 1,576 years
- (c) 10,672 years
- (d) 3,541 years
- (e) 2,658 years

_____ 2. Which of the following are key differences between chemical and nuclear reactions?

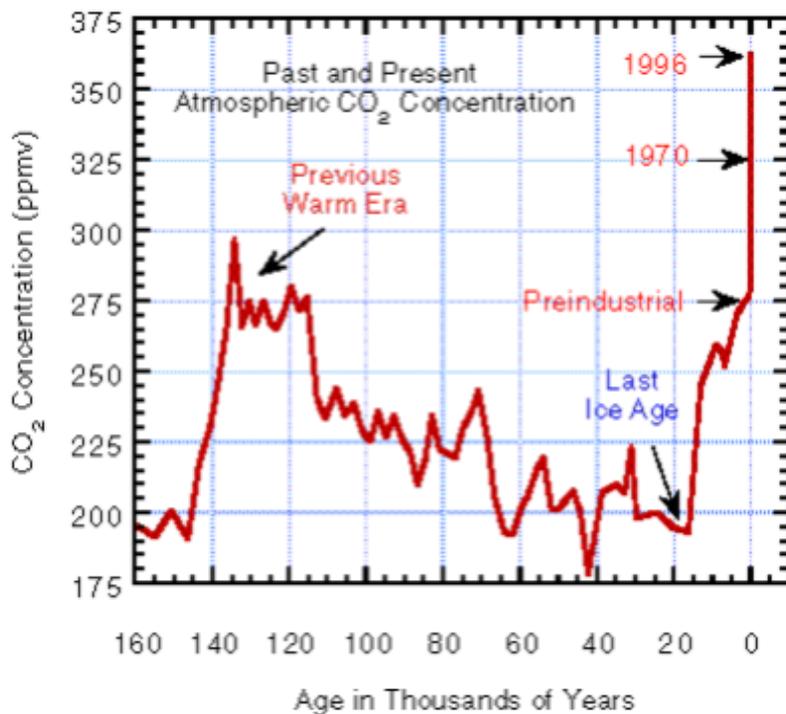
- I. Atoms do not change identity in chemical reactions, whereas in nuclear reactions they do
- II. Nuclear reactions release a greater amount of energy than chemical reactions
- III. Nuclear reactions have rates that depend on temperature, concentration, and catalysts, whereas chemical reactions do not

- (a) I
- (b) I, II
- (c) II, III
- (d) I, III
- (e) I,II,III

_____ 3. What is the molarity of a 35.0 mL solution of 9.00 *M* H₂SO₄ diluted to 0.500 L?

- (a) 6.30 *M*
- (b) 0.624 *M*
- (c) 61.1 *M*
- (d) 630. *M*
- (e) 0.630 *M*

4. Answer the following question based on the graph provided.



Which of the following **cannot** be determined based on the provided graph?

- (a) 42,000 years ago the CO₂ concentration was at an all time low
- (b) CO₂ levels increase during warm periods
- (c) There was a sharp increase in the CO₂ concentration after the last ice age
- (d) There were no warm eras prior to 160,000 years ago
- (e) 135,000 years ago the CO₂ level was about 33% higher than it was 35,000 years ago

5. A solution of caffeine (C₈H₁₀N₄O₂, 194.20 g/mol) in chloroform (CHCl₃, 119.37 g/mol) as a solvent has a concentration of 0.500 *m*. Calculate the percent caffeine by mass.

- a. 33.3%
- b. 16.3%
- c. 5.63%
- d. 8.85%
- e. 31.0%

- _____ 6. Why are molecular oxygen and molecular nitrogen **not** considered to be greenhouse gases?
- (a) The atoms are so light that the bond vibrations absorb in the UV.
 - (b) They have only two atoms and therefore cannot undergo asymmetric stretching.
 - (c) They lack a dipole moment.
 - (d) They are too dilute in the stratosphere, where the greenhouse effect takes place.
 - (e) The ozone layer filters radiation from these gases.

- _____ 7. Electronegativity:
- (a) has no periodic trends.
 - (f) is generally greatest for the transition metals.
 - (g) generally decreases left to right across a period and increases down a group.
 - (h) generally increases left to right across a period and decreases down a group.
 - (i) is the term for a common attitude among pessimistic electrons.

- _____ 8. The electronic configuration of Ca^{+2} in its ground state is:

- (a) $1s^2 2s^2 2p^6 2d^{10}$
- (b) $1s^2 2s^2 2p^6 3s^2 3p^6 3d^2$
- (c) $1s^2 2s^2 2p^6 3s^2 3p^6 4s^2$
- (d) $1s^2 2s^2 2p^6 3s^2 3p^6$
- (e) $1s^2 2s^2 2p^8 3s^2 3p^4$

- _____ 9. Which compound has a higher lattice energy, LiCl or CsCl? Why?
- (a) LiCl because it is more soluble than CsCl.
 - (b) LiCl because Li has a smaller ionic charge than Cs.
 - (c) LiCl because it has a smaller internuclear distance than CsCl.
 - (d) CsCl because it has a smaller internuclear distance than LiCl.
 - (e) CsCl because Cs has a smaller first ionization energy than Li.

- _____ 10. Carbon dioxide gas and methane gas are often called “greenhouse gases”.
Greenhouse gases
- (a) are the primary cause of acid rain.
 - (b) catalyze the destruction of the earth’s ozone layer.
 - (c) are the primary constituents of what is called “smog”.
 - (d) are linked to global warming by many models.
 - (e) None of the above statements is correct.

Key

- 1. E
- 2. B
- 3. E
- 4. A or D
- 5. D
- 6. C
- 7. D
- 8. D
- 9. C
- 10. D