

# **Innovative Uses of Assessments for Teaching and Research**



ACS SYMPOSIUM SERIES **1182**

# **Innovative Uses of Assessments for Teaching and Research**

**Lisa K. Kendhammer**, Editor  
*University of Georgia*  
*Athens, Georgia*

**Kristen L. Murphy**, Editor  
*University of Wisconsin-Milwaukee*  
*Milwaukee, Wisconsin*

**Sponsored by the  
ACS Division of Chemical Education**



American Chemical Society, Washington, DC

Distributed in print by Oxford University Press



## Library of Congress Cataloging-in-Publication Data

Innovative uses of assessments for teaching and research / Lisa K. Kendhammer, editor, University of Georgia, Athens, Georgia, Kristen L. Murphy, editor, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin ; sponsored by the ACS Division of Chemical Education.

pages cm. -- (ACS symposium series ; 1182)

Includes bibliographical references and index.

ISBN 978-0-8412-2986-0 (alk. paper)

1. Chemistry--Study and teaching--Evaluation. 2. Curriculum planning.  
I. Kendhammer, Lisa K., editor. II. Murphy, Kristen L., editor. III. American Chemical Society. Division of Chemical Education.

QD42.I56 2014

540.71--dc23

2014042388

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48n1984.

Copyright © 2014 American Chemical Society

Distributed in print by Oxford University Press

All Rights Reserved. Reprographic copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Act is allowed for internal use only, provided that a per-chapter fee of \$40.25 plus \$0.75 per page is paid to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. Republication or reproduction for sale of pages in this book is permitted only under license from ACS. Direct these and other permission requests to ACS Copyright Office, Publications Division, 1155 16th Street, N.W., Washington, DC 20036.

The citation of trade names and/or names of manufacturers in this publication is not to be construed as an endorsement or as approval by ACS of the commercial products or services referenced herein; nor should the mere reference herein to any drawing, specification, chemical process, or other data be regarded as a license or as a conveyance of any right or permission to the holder, reader, or any other person or corporation, to manufacture, reproduce, use, or sell any patented invention or copyrighted work that may in any way be related thereto. Registered names, trademarks, etc., used in this publication, even without specific indication thereof, are not to be considered unprotected by law.

PRINTED IN THE UNITED STATES OF AMERICA

# Foreword

The ACS Symposium Series was first published in 1974 to provide a mechanism for publishing symposia quickly in book form. The purpose of the series is to publish timely, comprehensive books developed from the ACS sponsored symposia based on current scientific research. Occasionally, books are developed from symposia sponsored by other organizations when the topic is of keen interest to the chemistry audience.

Before agreeing to publish a book, the proposed table of contents is reviewed for appropriate and comprehensive coverage and for interest to the audience. Some papers may be excluded to better focus the book; others may be added to provide comprehensiveness. When appropriate, overview or introductory chapters are added. Drafts of chapters are peer-reviewed prior to final acceptance or rejection, and manuscripts are prepared in camera-ready format.

As a rule, only original research papers and original review papers are included in the volumes. Verbatim reproductions of previous published papers are not accepted.

**ACS Books Department**

# Editors' Biographies

## **Lisa K. Kendhammer**

Lisa K. Kendhammer received her Ph.D. in 2013, in Chemical Education and Analytical Chemistry, at the University of Wisconsin-Milwaukee. Her dissertation project in Chemical Education under the direction of Kristen Murphy included identifying items that exhibited differential item functioning (DIF) on multiple-choice assessments and exploring the possible reasons why DIF occurs. Currently, she is a post-doctoral research associate at the University of Georgia under the direction of Norbert Pienta, where her main research project involves examining students' difficulty with general chemistry laboratory skills and techniques. Since 2011, she has had one publication and presented 3 invited talks on DIF.

## **Kristen L. Murphy**

Kristen L. Murphy received her Ph.D. in Physical Inorganic Chemistry at the University of Wisconsin-Milwaukee (UWM). She has served as the Associate Director for the American Chemical Society Examinations Institute since 2005 and as a faculty member in chemical education research at UWM since 2008. Her research projects include the investigation of a student's scale literacy and using differential item functioning to examine multiple-choice general chemistry exams for differential item performance by gender subgroups. She has been dedicated to the undergraduate program at UWM, teaching many large general chemistry courses, and received the UWM Faculty Distinguished Undergraduate Teaching Award in 2012.

## Chapter 1

# Innovative Uses of Assessments for Teaching and Research

Lisa K. Kendhammer<sup>1</sup> and Kristen L. Murphy<sup>\*,2</sup>

<sup>1</sup>Department of Chemistry, University of Georgia, 140 Cedar Street,  
Athens, Georgia 30602

<sup>2</sup>Department of Chemistry and Biochemistry,  
University of Wisconsin-Milwaukee, 3210 N. Cramer Street,  
Milwaukee, Wisconsin 53201

\*E-mail: [kmurphy@uwm.edu](mailto:kmurphy@uwm.edu).

Instruction and assessment are so common to teaching and learning that for many readers this may be second nature. There are certainly many kinds of instruction and assessment available to instructors, and these are chosen based on many factors. Where instruction may be more commonly discussed, assessments may be more guarded. Some may view assessments as any resource into understanding more about student learning, while others may view assessments in a narrower sense of hourly exams or final, summative exams. While these tests certainly do qualify as assessments and may have the necessity for being kept guarded (as some reuse tests or some use standardized tests), there are also other kinds of assessments that provide rich information about the efficacy of the instruction. Further, the results of assessments can be used to make decisions, such as course grades, and commonly may be associated with fulfilling that need. However, assessments can provide information to instructors and researchers about many other factors including students' prior knowledge, conceptual understanding, longitudinal progression of knowledge, and misconceptions. Finally, classroom assessments are valuable tools to reflect locally on instruction and globally to consider student content knowledge when reflected on the longitudinal performance of students and the implications for the program.

## **Introduction**

*What do students know? How do their prior knowledge and experiences shape this? What are their motivations for learning or their confidence in learning? How do we find this out?*

As instructors and researchers, these questions may be fundamental, but the manner in which these are answered is diverse and exciting. One method by which to answer these questions involves the use of assessments. Assessment may be viewed narrowly by some as course tests that are summative and formal. These classroom assessment techniques are important as they are used to judge what students know in terms of content knowledge and can contribute to the decisions of course grades. However, these assessments can be formative as they provide feedback to students about what they know and what they do not. Beyond this, content tests can be used to examine prior knowledge of students, and the feedback to instructors can guide future instruction and target the needs of the students. Other types of classroom assessments could include informal assessments for formative feedback to students, assessments specifically examining for student misconceptions, assessments built to examine aspects of the affective domain including self-efficacy or motivation, and assessments examining students' metacognition. These assessments can take many forms, from forced-response tests (multiple-choice tests) to open-ended questionnaires or even student interviews.

Where classroom assessment can provide implications for classroom instruction, programmatic assessment can provide implications for the collection of courses that constitute a program. Therefore, programmatic assessment may build on the same classroom assessment techniques, but these techniques are now considered in the context of the program. This could be considered for a single student or a cohort of students longitudinally, for a single course over an extended period of time, or commonly a collection of courses that build a program. Regardless, many different assessments can be used to reflect on the efficacy of a program, extending beyond summative final exams.

### **What Information Assessments Can Provide**

Logically, we expect to learn what students do or do not know about specific content areas from typical course assessments. We learn this through what students can or cannot do correctly on a test. This information can be valuable when assigning grades and providing more specific feedback to students about areas of strength or weakness. However, this can inform instructors when considering methods of instruction used for those specific content areas. Perhaps, the method of instruction was altered or more resources were provided: how did this affect the students' content knowledge? Content tests may provide some information about this.

Thinking beyond standard content tests, considering other, innovative assessments, we may enjoy a richer picture of what students know or understand by investigating prior knowledge, misconceptions, motivations, or self-concept. This can inform us as instructors so we can provide a better instructional



environment that can target areas of weakness. This can also inform us as researchers on different areas such as investigating how students learn and how prior experiences or perceptions affect learning.

A word of caution about drawing conclusions based on information provided from assessments. The words “valid” and “reliable” are used commonly to describe assessment results and many researchers are careful in their conclusions based on limitations associated with establishing validity and reliability (1). As with any assessment or instrument, we must carefully consider what we are measuring and how we measure it. Validity checks should be considered routinely, particularly when developing new assessments (2). Innovation in developing and using assessments should not be hampered by adding validity and reliability checks, but rather strengthened because of this.

### **Purpose of This Book**

The purpose of this book is to provide a small collection of innovative ways that assessments have been used for classroom or programmatic assessment or for research investigations. This is by no means comprehensive, but rather a means to encourage innovation in other classrooms or in investigating other research. Therefore, this selection offers samples of assessments that have been developed or adapted, new assessment methods or techniques, new methods of providing feedback, or comparisons of methods for establishing test fairness. We hope this provides a spark or idea for innovative uses of assessment in other areas of teaching or research.

### **How To Use This Book**

There is no prescribed method for using this book. If an instructor or researcher has similar goals as described, then it is reasonable that the same or similar assessment may be used in the manner described. Even when using the same or similar population, one would expect that instructors or researchers would incorporate validity checks (2) in order to establish trust in the results and corresponding judgments made based on the assessments. It is also possible that instructors or researchers will consider the different types and uses of assessments presented and adapt these for different populations or testing environments or conditions. This new research would then add to the innovation initially presented here to further our collective knowledge of what students know. Finally, we also could expect that an instructor or researcher considering the work presented here could be invigorated to investigate new assessment pathways that would lead to new and innovative assessments for research and teaching.

The book is organized into four general sections as shown in Table 1. The first section describes the processes by which assessments are constructed and used. The second section focuses on what is learned from assessments in an informal environment, including the use of practice exams and feedback provided to help students reflect on their own learning. Formal classroom assessments and the decisions associated with different assessments and techniques comprises

the third section. The final section focuses on assessment goals and innovative investigations of student learning with descriptions of new assessments and new online tools for measuring student understanding.

**Table 1. Organization of the Chapters**

Section I	Chapters 2, 3, and 4	The Process By Which Assessments Are Developed And Evaluated And How This Can Facilitate Research
Section Ii	Chapters 5 and 6	Informal Classroom Assessments – Helping Students Reflect On Their Learning
Section Iii	Chapters 7 and 8	Formal Classroom Assessments – Gauging Student Learning
Section Iv	Chapters 9, 10, and 11	Assessment Goals And Innovative Methods For Investigating Student Learning

We hope that in whatever form you find this work useful, you are encouraged to investigate student learning that pushes us all to think about answering important questions such as, “What do my students know?”

## References

1. Barbera, J., VandenPlas, J. R. All Assessment Materials Are Not Created Equal: The Myths about Instrument Development, Validity, and Reliability. In *Investigating Classroom Myths through Research on Teaching and Learning*; Bunce, D. M., VandenPlas, J. R., Eds.; ACS Symposium Series 1074; American Chemical Society: Washington, DC, 2011; pp. 177–194.
2. Arjoon, J. A.; Xu, X. Y.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013**, *90*, 536–545.

## Chapter 2

# Matching the Evaluation Plan to the Question

**Diane M. Bunce\* and Regis Komperda**

**Chemistry Department, The Catholic University of America,  
620 Michigan Avenue, NE, Washington, DC 20064**

**\*E-mail: [Bunce@cua.edu](mailto:Bunce@cua.edu).**

There is a growing demand to evaluate the effectiveness of teaching pedagogies in the chemistry classroom and laboratory but there is often confusion on where to start and how to conduct the evaluation. This chapter will address key issues in defining the question of interest and choosing the research approach; establishing a theoretical framework; designing the methodology and tools to investigate it, analyzing the data, and presenting the results in a useful manner. The constant interplay between the question of interest and the evaluation plan to investigate and present it, is the main focus of this chapter.

### Introduction

Evaluation and assessment are terms that are often used interchangeably. In reality, they are similar but distinct. Assessment is usually thought of as formative and used diagnostically to identify ways to improve learning. Assessment is more appropriately used when describing an investigation within one's own classroom where the results are used by both teacher and students to improve their shared experience. Evaluation is summative and deals with a judgment on the effectiveness of an approach or product (*I*). Evaluations are also investigations of teaching and learning situations, but are more generalizable. Evaluations are more likely than assessments to be submitted for publication in peer-reviewed journals or reports.

This chapter will deal primarily with evaluations. There are four parts to the discussion of matching an evaluation plan to the question being addressed, all of which are interrelated. These four parts include the following: Defining

the question and selecting an appropriate research approach; Designing the methodology and tools to investigate the question; Analysis of the data; and Presentation of the data. We will address each of these components and discuss the interconnections among them.

## **Defining the Question and Selecting an Appropriate Research Approach**

Questions are the fuel for research. Without questions, there would be nothing to investigate. As a result, we as researchers need to spend a significant amount of effort on defining the question. Usually research questions start with an idea that has perplexed the researcher based either on observation, experience, readings, or theoretical issues. However, new pressures from state officials, school or university administrations, or funding agencies have led to a demand to know if, how, and why different teaching and/or learning approaches work. Regardless of where the idea is generated, it is the researcher who must develop the researchable question(s) from the idea presented. There are published resources (2) that can help with this development.

Essentially the research idea must first be formulated into a question. That question is best analyzed if stated in language that expresses the goal(s) of the researcher. An example of an idea that might be of interest is “Does teaching chemistry using clickers (personal response devices) increase student achievement?” or “How does the use of clickers affect student learning?” Once the idea is verbalized, it can be tested for components of constructing a good researchable question (2) the first of which is “Is the question worth asking?” There are many questions that could be asked, but not all of them are important to increasing our understanding of how student knowledge is supported; the effectiveness of the interaction between how we teach and how students learn; or if the time, effort and money to implement a new teaching approach will have a measureable effect on student learning. Because the investigation of a question involves effort, time, and money, the question should be one that is important to the people involved whether they are students, teachers, parents, administrators, elected officials, granting agencies, or other researchers.

Once the idea has been formulated as a question, the next step is to make the question operational. This means rephrasing the question so that it can be addressed best by a specific type of research design either qualitative, quantitative or a mixed-methods design. The type of question asked helps to determine the best research approach. For instance, in our example above regarding clickers, if we want to know if clickers increase achievement, a quantitative approach would be the most appropriate. Here we are asking a question that deals with the demonstration of an outcome, i.e. what is the effect on student achievement of using clickers. We could identify a dependent variable (student achievement) and an independent variable (use of clickers) and run statistical tests to address this question. If, on the other hand, we were interested in the second question posed above, namely, “How does the use of clickers affect student learning?”,

we are really asking a more fundamental and open-ended question that lends itself more to a qualitative approach. Here we are interested in understanding the character of the interaction of student use of clickers and learning, not just the outcome of use. Statistical analyses would not address this question adequately but interviews, observations, focus groups, and surveys of the students would be more appropriate approaches. There are several good resources devoted to understanding the difference between quantitative and qualitative research and how to plan and implement qualitative research (3, 4).

Mixed methods use a combination of qualitative and quantitative approaches to address a research question (5, 6). Sometimes such a method is an equal mix of quantitative and qualitative approaches and other times, it uses one approach more extensively than the other. For instance, to understand better the effectiveness of clicker use on student achievement, a smaller component of the research design might include the use of a survey or interviews of a subset of the students to investigate *how* they used clickers in the study and what their opinions are on its usefulness.

Choosing the research approach (qualitative, quantitative or mixed methods) is the first step in matching the evaluation plan to the question. The decision on an approach based on the question asked, will impact the development, execution, and analysis of the research.

## **Quantitative Research**

In quantitative research, the next step in the process is to operationalize the question. This process helps the researcher identify the pertinent variables that should be measured or controlled in the experimental design (2).

One way to think about the process of operationalizing a research question is to look at the nouns and verbs in the question you want to ask. For instance, in our example of a research question (Does teaching chemistry using clickers increase student achievement?), there are several parts of this question that need defining. For instance, what does the term “teaching chemistry using clickers”, mean? Will clickers be used in class every day? How many times will they be used per class? At what point in the class will clickers be used-- at the beginning of the class as a review of the previous class’ presentation, before a topic is introduced, during the presentation of a topic, immediately after a topic is taught, or at the end of the class as a summary? Similarly, the next phrase in the research idea is equally vague. What does “increase student achievement” mean? Does this mean a statistically significant increase in student achievement scores on the same standardized exam used this year compared to previous years’ classes’ scores? Or does it refer to a statistically significant increase in student achievement scores on the same standardized exam compared to an equivalent class taught (with or without the same instructor) during the same time frame that does not use clickers? Does increased student achievement refer to scores on a test, either multiple choice or open ended, or does it mean something else like students’ ability to explain the chemical theory behind an answer? Obviously, being explicit about how the researcher interprets the research question will help define the evaluation plan. It

is, therefore, appropriate at this point in the research planning to replace the vague phrases in the research question with more specific descriptions that will help direct the resulting methodology. This is what is meant by operationalizing the question.

Redefining a research question into an operational question results in both major and minor decisions that affect the evaluation plan. These questions include additional components of constructing good research questions (2), including who will be studied, in how many different sections, courses, or institutions, when measurements will be taken during the semester and the class period as well as for how long or with how many repetitions such measurements are needed. Defining who will be studied is thus important to developing the evaluation plan.

At this point in the process of matching an evaluation plan to the question, it is important for the researcher to take a step back and be explicit about what the “take home” message of the research will be. In our example about the effectiveness of clickers, the take home message might deal with whether the use of clickers in class is worth the time, effort and money to implement them. The purpose of the research question then becomes--Is there a change in student learning as a result of using clickers, and if so, is the change statistically significant and meaningful? There is always room in the analysis of a study for unplanned insights to manifest themselves but using the main take home message to make sure that the methodology addresses the pertinent variables is an important step.

The next step in the process of defining the research question is to examine the more explicit meaning of the phrases in the current iteration of the research question and ask what type of data would address the question as currently asked. For instance, to examine “increased achievement”, scores on standardized exams for the current class is one approach. However, if the researcher wants to compare achievement on a standardized exam either used historically or concurrently by different teachers, some things must either already be in place or be possible to implement. These include whether such an exam is currently available, whether other teachers would agree to administer it, if historical data on students’ achievement on this exam exists and is accessible under Institutional Review Board (IRB) and Family Educational Rights and Privacy Act (FERPA) guidelines. If the researcher is interested in students’ ability to explain the underlying chemical theory for a given question, then either open-ended questions or interviews might be appropriate. Thus, including a qualitative component in the research design and redefining this proposed study as a mixed-methods approach might be called for. In these situations, the validity and reliability of the questions used in both situations must be established before they are implemented in the evaluation plan.

## **Qualitative Research**

Qualitative research operates according to a set of rules and expectations just as quantitative research does. It is an approach that starts with observations and questioning the key participants along with examination of the materials that help define the question of interest. Qualitative research is not one single approach but rather involves the choice of qualitative approach from the following five

general categories: ethnography, narrative, phenomenology, grounded theory, and case study. The approach chosen depends on both the question asked and the philosophical beliefs of the researcher (4)

In the example of clickers used here, if we wanted to understand how clickers affected student learning, one way to start would be to interview students. The purpose of the interviews would be to hear in students' own words, the effect they thought the use of clickers had on their learning. To pursue this question, the researcher might present the student with one or two clicker questions that were previously used in class and then ask the student to explain the underlying chemical concept presented in the clicker questions. This could be followed by asking students probing questions regarding how the specific clicker question or the discussion that ensued helped the student with the understanding needed to explain the concept during the interview or to address a pertinent question on a previously completed class test. When a subset of the students in the class has been interviewed, the student comments can be transcribed and analyzed using software designed for this purpose (7). The analysis of student comments should converge on the elucidation of key ways students use or do not use clickers to affect their understanding of the underlying concepts. This would define the qualitative study as phenomenological. The important point with this type of research is that the reasons and evidence come directly from the students. The data actually do "speak for themselves" in this type of research approach. The role of the researcher is to facilitate the conveyance and analysis of the data provided by students. This is not necessarily an easy process but the wealth of authentic knowledge gleaned from it, is rich, in-depth, and provides an understanding that may not be accessible in any other format.

Qualitative research in this example should also address the bigger picture of our research question. In the case of the hypothetical clicker study, another key player in how the use of clickers affects student learning, is the teacher who develops or chooses both the clicker questions and the corresponding test questions or other measures of student understanding. To understand the larger picture of how clickers affect student learning, the teachers should be interviewed and their responses analyzed and compared to those of the interviewed students. For instance, how do the teachers think the use of a specific clicker question (the same one shown to students in their interviews) relates to the instrument of student learning used to measure that content? The researcher can analyze the teachers' responses and compare them to those of the students. If there is not a good correlation between the two, this might signal a disconnect between what the teacher thinks is being conveyed to students and what actually is being conveyed. Thus the impact or lack thereof for clicker use might be a result of a disconnect between the teacher's view of what is happening and that of the student, rather than the actual effect of clicker use.

### **Mixed-Methods Research**

In mixed-methods research, we use the strengths of each research approach (quantitative and qualitative) to understand more fully the question we want to ask. As was mentioned previously, the relative contribution of each separate

research approach to the mixed methods is a decision for the researcher based upon the ultimate goal of the research. For instance, in our example of the effect of clicker use on student achievement, it might be beneficial to interview students asking them to describe how they used the information they gained during clicker use when they solved a corresponding achievement question. This use of qualitative methods is similar to that described above for a qualitative approach. Here, qualitative methods would be used to help explain a significant or non-significant effect as measured by the achievement grade. Another example of how mixed methods might work in this example would be the development, validation and use of a survey instrument to glean information on how students view both the use and effect of clickers on achievement.

The mixed-methods designs described here might be considered *convergent parallel* mixed methods because both qualitative and quantitative tools are being administered at the same time in the research design. *Explanatory sequential* mixed methods could also be used if the interviews or survey of students were administered at some time after the final achievement measure was completed and students were aware of their results on the achievement before being asked to explain why they these results had occurred (4).

## Theoretical Frameworks

In order for the proposed research to have an impact on the field, it must have a theoretical framework. In other words, the researchable question must advance our understanding of what we already know about teaching and learning chemistry. If the proposed research does not have a theoretical framework, it then runs the risk of standing as an isolated fact in a collection of isolated facts and its generalizability to other classes will be limited (quantitative research) or the analyzed data might not coalesce into an understandable, generalized model of understanding (qualitative research). Thus, use of a theoretical framework may differ between quantitative and qualitative research approaches but in both approaches, the theoretical framework helps the researcher develop a model of the data in the study. In quantitative research, the identification of a theoretical framework used to create a model for understanding occurs at the beginning of the study to help guide the experimental procedure and at the end for use in interpreting the results. In a qualitative approach, the researcher must be aware of the existing appropriate theoretical frameworks in the literature but it is the data itself that help develop the model used to interpret the results in the particular study.

Guidance in selecting and using theoretical frameworks in chemical education studies is addressed in other resources (8, 9). Choosing a theoretical framework is different from citing similar studies. Citing similar studies is important but as been stated elsewhere (2), these references to other studies are an *application framework* that serve to inform the researcher of both what has already been done to investigate this topic and how it has been done. A theoretical framework, on the other hand, is at a higher-level of abstraction and results in the development of a model to interpret learning as a cognitive, educational, sociological or human-computer interface process best explained through the tenets of these



fields. A theoretical, as well as an application framework, are important to understand more deeply the research question and what effects are likely to be observed, documented or measured in the evaluation.

## **Designing the Methodology and Tools (Evaluation Plan)**

### **Quantitative Research**

As the process of question development proceeds in quantitative studies, it may become advantageous to divide the original question into subquestions, each one of which can be accounted for in the evaluation plan with its own data collection and analysis. A series of subquestions allows the researcher to test for the influence of numerous factors such as mathematical aptitude, previous experience and/or level of success in math and science courses, etc. on the dependent measure, which in the case of our clicker example, is achievement. As the list of subquestions develops, it may become obvious that the research question as written is too large to be addressed in a single research project. If this is the case, then investigation of one or more of the subquestions may be adequate for a single research project. The development of subquestions also provides some protection against a simplistic ‘yes or no’ answer to the multi-variate question of learning. Time spent on the process of defining the research question and the initial development of the evaluation plan can serve to help define the boundaries of a research question that can be realistically addressed in the researcher’s current availability of time, effort, funds, and sample of the population. This analysis can also increase the chances that the evaluation will result in valid and reliable results (2).

At this point in a quantitative question and evaluation plan development, it may become necessary to identify possible intervening variables that could mask or skew the results of the planned evaluation. In the example of the effect of clickers on achievement, differing aptitude levels (such as SAT scores) or students’ experience and/or level of success in previous math and science courses may unduly affect the results of their achievement in the proposed evaluation. If such intervening variables can be controlled or measured, the results of the research can be analyzed by accounting for the part of observable or non-observable differences due to them, thus facilitating a more accurate assessment of the results of the evaluation.

In order for the evaluation plan to be effective in addressing a quantitative research question, it is necessary to design the plan with as much detail as possible and then check the plan for its appropriateness in terms of collecting data that are directly tied to each part of the research question. Designing a table listing each part of the research question or subquestion together with the tool used to collect appropriate data and the type of data expected to be generated is a mechanism for checking that all pertinent variables have been addressed before the evaluation plan is started.

Collecting data that is easy to collect but is not needed to investigate the research question is counterproductive. Just because you can collect certain data, doesn't mean that you should. Only data that can be linked to the theoretical framework through the research question or subquestions should be collected. Additional data can be collected if there is an identifiable reason for doing so. For instance, in our clicker question research, it would be inappropriate for us as researchers to ask and collect data on whether the parents of the participating students were divorced or separated. Our theoretical model does not include this variable in the relationship between clicker use and achievement (quantitative) or learning (qualitative), therefore, it is inappropriate for use to ask and/or collect data on parental marital status.

In addition to the development of a table of research question(s) and tools to collect data, it is wise to revisit the main take home message to check that the data collected will form a convincing argument to address the research question.

### *Quantitative Research Tools*

When the overall evaluation plan has been developed, it is time to examine more closely the tools that will be used to collect data. It is often at this point that the theoretical framework helps in the selection, modification or development of appropriate tools. If in our example of clicker use, we included the effect and quality of peer interaction in the discussion of the answers to clicker questions, it might be appropriate to record and analyze the discussion among students during the selection of an answer to the clicker questions. The analysis might involve the quantity and quality of student use of concepts to make logical arguments for selecting a specific answer. Tools available to measure this type of interaction could include analyzing the discussion using a rubric developed by the researchers or discourse analysis of the warrants and proofs evident in the students' conversation. Resources are available that offer suggestions on both the use of observations (10) and discourse analysis (11).

The selection of tools or instruments used in the evaluation plan to generate data needed to address the question asked is a daunting process. Sometimes the tools needed already exist (such as standardized exams) while other times they can be modified from existing tools (such as a publisher's test bank or published survey questions). However, there are often situations where the tools must be created by the researcher to match better the research question being asked. In all cases, the validity and reliability of the tools used must be established. Without documenting the validity and reliability of tools, the data generated is suspect. Validity and reliability help establish that the tools actually measure what the researcher claims they measure.

To examine further the connection between research question and evaluation plan with subsequent selection of tools, we turn to examples from published research. Three articles published in science or chemistry education research journals all posed questions dealing with student learning including whether a new lab program helped students reach previously established learning goals (12); whether student problem solving strategies improve through collaborative

learning (13); and whether the use of either clicker questions in class or online quizzes increased student achievement (14). These research questions are all based in experiences the researchers had within the teaching environment. The evaluation plans stressed the use of laboratory accuracy and precision data normally collected in lab (12); computer identification and analysis of strategies used by students when solving problems of chemically-based scenarios online (13); or comparison of scores on achievement questions that had clicker, online quiz, or no antecedents (14). In each case, data were collected from instruments designed or modified to fit the research question. Two of the three (12, 14) used surveys as either the main or supplementary tools. All three tied the research question to the evaluation plan in the article's description of the research.

In two other studies (15, 16), surveys were used as the main tool of the evaluation to investigate the research question. In Bunce, Havanki and VandenPlas (15), the survey was constructed based on two theories that explain the process and factors that impact decisions to adopt change. In Dalgarno, Bishop, Adlong, and Bedgood (16), a survey was designed to collect self-reported data on the use of a virtual environment in a distance learning class to prepare for on-campus laboratory experiences. In each case, the tool of choice used in the evaluation plan was a survey created by the researchers.

The evaluation plan, in general, relies heavily on the instruments used to collect data. As can be seen in the studies cited here that span the topics of student learning, student behavior, or teachers' readiness to adopt change, surveys are often used as either the main or supplemental tools for data collection. Because surveys are often used in evaluation projects, it is important to review the possibilities for introducing error into this tool, which could result in invalid or inaccurate results. Several resources exist that can guide the researcher to the development or modification of good survey tools (17–23).

## **Qualitative Research and Tools**

The tool development in qualitative research is a more open-ended process than in quantitative research. Here, the researcher is actually the tool by which data is elicited. In other works, the researcher sets the environment that will allow the subject to respond with pertinent comments for analysis. This doesn't mean that time and effort are not needed to plan the environment for data collection. As described previously in the example of how clickers affect student learning, in order to obtain rich or in-depth data on how students use clickers in learning, it would be wise to provide both a clicker question the students have previously used in class and the assessment of learning, whether it is a question from a test or an explanation of the underlying concept of the clicker question, to guide the student's discussion of how clicker questions are used. This situation would also allow the researcher to judge the quality of learning by capturing the students' understanding in their own words. It is through the careful planning of the environment by the researcher that the students are able to express their understanding (learning) and describe how the clicker question process did or did not impact that learning. The researcher is the tool who allowed that data to be elicited.

## Data Analysis

In quantitative research even though analysis of data is typically done during the later stages of research, it should be planned at the start of the project. This is important to ensure that the data collected can be used to address the research question asked. Two important questions to consider when planning the research are: 1) What type of analysis (for example, comparison or prediction) best matches the research question? and 2) What should the end product of the analysis be to adequately address the question asked?

Qualitative research data analysis is a bottom-up approach where often a rubric is established by the initial in depth reading of subjects' comments. This rubric is then applied to the entire set of data for the purpose of generating more general insights into understanding the issues.

The most important thing is to match the type of analysis used to the research question asked.

### Matching Type of Analysis to the Research Question

#### *Quantitative Research*

We have already established that the wording of the research question determines the type of data analysis needed to address it. In this section, we look at that point more closely. For instance, comparisons between groups are indicated by phrases such as “better than”, “greater than”, “increase”, or “compared to” in the research question. As an example, a modification of our hypothetical clicker research question that has been modified could be “Does the use of clickers multiple times during the presentation of a topic significantly increase student achievement on a standardized exam as compared to students in the prior year taking the same exam under similar conditions but without the use of clickers in class?” Based on this modified question, a comparison between the two groups (with and without clickers in the two years) is proposed. These two groups represent the independent grouping variable and the scores on the standardized teacher-written exam are the dependent variable, or outcome. The take home message here would be whether one group scored significantly higher on the exam compared to the other.

Predictions, or relationships, between variables, as opposed to comparisons, are most often investigated using linear regression model-based statistics. Regression models, including correlations, are based on the generic line equation,  $y = bx + a$ . While the general linear model underlies many statistical tests, including those used for comparison (24), such a discussion is beyond the scope of this chapter. Regressions and correlations typically require continuous numeric data, but it is also possible to make predictions from categorical grouping variables using a process called dummy coding. Details regarding this technique can be found in statistics textbooks (24).

As an example of using regression to address a research question, our research question could be changed to “Can student scores on standardized teacher-written exams be predicted from their clicker use and academic aptitude as measured by SAT scores?” The use of the term “predicted” in the research question determines that a regression model is needed to address the question. Other terms such as “associated with” or “related to” in a research question describe correlational relationships rather than predictive relationships. A common misconception about the results of a regression or correlation is that the ability to predict one value from another implies that one variable *causes* the other. Rather, the results of a regression or correlation define only whether or not a relationship exists between the two variables not whether one variable causes the other.

### *Qualitative Research*

Analyzing the results of qualitative research typically involves transcribing interview data and organizing the data in electronic files; reading the transcripts, writing notes on effects seen; reviewing field notes taken during or shortly after the interviews; and reviewing the physical evidence from the interviews such as materials shown or used by the students during the interview. Next in the analysis is the iterative process of interpreting the data by developing categories or nodes to analyze the data. What is unique to this process is that the students’ words become the evidence for the categories. Explanations must be written describing the characteristics of the categories so that any other qualified researcher can reliably code the data according to the same categories. Validity and reliability for this type of analysis requires that multiple researchers will code the transcript data the same way. If not, then the effect seen by one researcher may not be strong enough to serve as a bona fide conclusion (4). The analysis process for qualitative research, which may require a smaller sample size than quantitative research, can also be much more labor and time intensive than quantitative analysis.

## **Relating the Results of Data Analysis to the Research Question**

### *Quantitative Research*

The end products of quantitative data analysis can be divided into descriptive and inferential statistics. As the name suggests, *descriptive statistics* present data by providing a description or summary. This summary can take the form of the mean, standard deviation, skewness, or frequency distribution for a variable. *Inferential statistics*, by contrast, are tests that allow the researcher to draw conclusions based on the data. Inferential statistics also allow the researcher to infer more generalizable results about the overall population from which the research sample is drawn and provide more control over possible errors in interpretation. There are several resources that describe the process of utilizing inferential statistics (24, 25).

Results of descriptive statistical analysis are often presented in data tables or displayed as bar graphs, line graphs, pie charts, or scatter plots. The ability to quickly, and often visually, summarize data is the main benefit to using descriptive statistics. Additionally, descriptive statistics can be used to check assumptions regarding the appropriateness of data for use in inferential statistical tests. This is especially true for the assumption that the data collected are normally distributed, i.e., take the shape of a bell-shaped curve. Examination of visual presentations of descriptive statistics can also help the researcher search for patterns in the data, which can be explored through further analysis. However, if the final data analysis only includes descriptive statistics, it might lead the researcher or reader to an interpretation of the significance of the results based on incorrect or unsupported generalizations of the data.

Inferential statistical tests provide a more objective analysis of the data that, if done correctly, controls for possible types of error such as Type I error (accepting a result as positive when it is not), denoted as  $\alpha$ , or a Type II error (accepting a result as negative when it is not), denoted as  $\beta$ . The power of a statistical test is related to the probability of Type II error through the relationship  $1-\beta$ . This means that if the chance of committing Type II error is 20%, the power of the test is 0.8. A power of 0.8 is interpreted as evidence that an effect that does exist will be detected 80% of the time. The power of a test is influenced by the size of the sample (24).

Inferential statistical tests include: *t*-tests, the multiple types of Analysis of Variance (ANOVA), correlation, and regression. Results of inferential statistical analysis are typically presented by reporting *p* values to show the probability of making a Type I error (false positive) along with values for the actual test statistic such as *t*, *F*, or *r* for *t*-test, ANOVA, and correlation, respectively. Most inferential statistics require the researcher to demonstrate that basic assumptions about the data have been met before the statistical test is run. Inferential statistics that require the assumption of a normal distribution of the dependent variable are known as parametric tests, and those that do not are known as nonparametric tests. More information about when to use nonparametric statistical tests can be found elsewhere (26).

As discussed, both descriptive and inferential statistics help interpret data, but in different ways. Descriptive statistics provide a building block for inferential statistics and complement the presentation of the inferential statistics by providing a visual summary of the data. Descriptive statistics are a useful tool but they typically serve as an intermediate in the interpretation of the data and conclusions produced by inferential statistics.

### *Qualitative Research*

Developing categories or codes of qualitative data is not the end to the analysis process in this type of research. Qualitative research is interpreted from more general themes that are supported by the categories found in the data. It is the higher order themes that help develop the model used to address the research question (4).

## Presentation of Results

### Quantitative Research

#### *Comparisons*

Presentation of the comparisons using descriptive statistics described previously in the modified quantitative question (“Does the use of clickers multiple times during the presentation of a topic significantly increase student achievement on a standardized teacher-written exam as compared to students in the prior year taking the same exam under similar conditions but without the use of clickers in class?”) would likely take the form of reporting the number of students in each group, the mean exam scores and standard deviations (SD) for each group, as shown in Table 1. Here the hypothetical data are also broken down by SAT score groups with high and low groups based on the mean SAT score of 600.

**Table 1. Descriptive Statistics for Exam Scores by Clicker Use and SAT Score Group**

<i>Clicker Use</i>	<i>SAT Score</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>
No	Low	15	74.00	9.09
	High	15	76.93	9.75
	Total without clickers	30	75.47	9.37
Yes	Low	15	78.33	6.61
	High	15	81.87	8.75
	Total with clickers	30	80.10	7.83
Total	Low SAT	30	76.17	8.11
	High SAT	30	79.40	9.44
	Grand Total	60	77.78	8.88

Visually these results could be presented in the form of bar graphs with the height of the graph indicating the mean exam score for each group of students. Figure 1 shows fictional data of 30 students without clickers having a mean exam score of 75.47 and 30 students with clickers having a mean exam score of 80.10. One problem with presenting these results using only descriptive statistics is that the reader has no way of knowing if the observed differences between groups are statistically significant. Additionally, the scale of the graph can be manipulated to either magnify or reduce the apparent differences between groups, as shown in Figure 1 graphs *a* and *b*. In Figure 1a, the mean standardized scores on the y axis range from 0.00 to 100.00 while in Figure 1b the range on the y axis is from 75.00 to 81.00, thus magnifying the difference in mean standardized exam scores between the two groups.

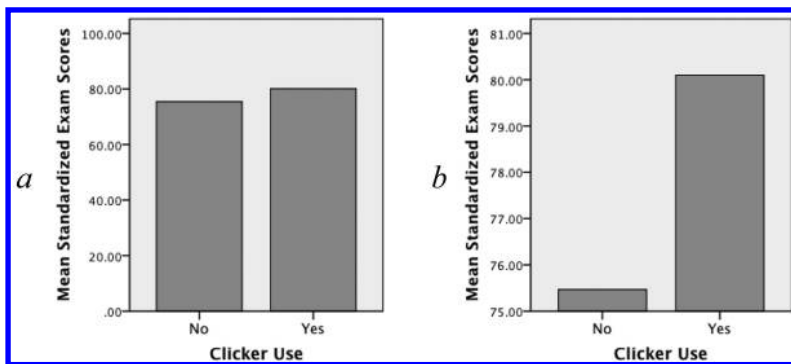


Figure 1. Bar graph of mean exam scores by clicker use with scale from 0-100 (a) and scale 75-81 (b).

Even though differences might exist between the exam scores based on clicker usage, these differences may be due to chance. Using only descriptive statistics puts the researcher in the position of having to argue that the difference in achievement scores either is or is not a meaningful result. Presentation of the results can lead the reader to make assumptions regarding the significance and meaningfulness of the differences. Further discussion of misleading ways to present data can be found in the reference by Wainer (27).

If the researcher includes SAT score data as a way to examine a possible intervening variable, this additional grouping can complicate the visual presentation. As seen in Figure 2, presenting all of this data requires the use of four columns representing the four possible classifications of students (No clicker—Low SAT, No clicker—High SAT, Clicker—Low SAT, Clicker—High SAT). Even with these distinctions in the data, no conclusions can be reliably drawn from the graph about whether differences displayed are statistically significant or due to chance.

Inferential statistics could be used for the original two-group comparison by conducting an independent *t*-test to look for a statistically significant difference in the mean exam scores between students who did and did not use clickers. Including the SAT variable requires the use of a factorial (two-way) ANOVA. The benefit of an ANOVA inferential approach is that the main effects of each variable (clicker or no clicker) are examined as well as effects due to interactions between the variables (SAT level and clicker use). More information on these tests and their underlying assumptions can be found elsewhere (24, 28).

Presentation of the inferential statistics' results in a table would include, but is not limited to, significance levels (*p* values) and values of the test statistics (*F* values), as shown in Table 2. The last two columns and first three rows in this table present the *F* statistic and associated *p* value for the main effect of clicker usage, SAT score group, and interaction between the clicker use and SAT score group, respectively. Further details on the other information presented in this table including the sum of squares (SS), degrees of freedom (*df*) and mean square (MS) values can be obtained from statistical references (24, 28).



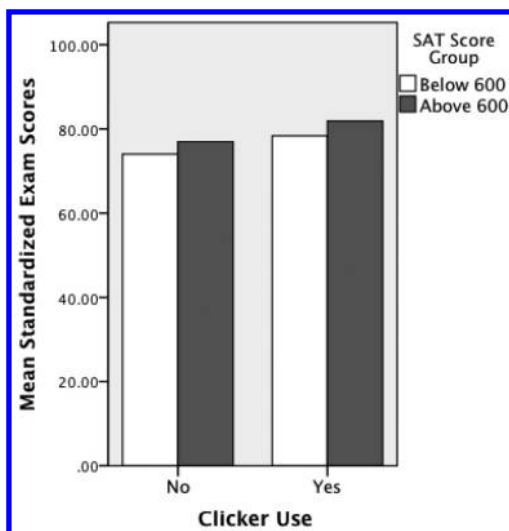


Figure 2. Bar graph of mean exam scores by clicker use with division by low (below 600) SAT score and high (above 600) SAT score.

**Table 2. ANOVA Summary Table for Exam Scores by Clicker Use and SAT Score Group**

Source of Variance	SS	df	MS	F <sup>a</sup>	p
<b>Clicker use</b>	322.02	1	322.02	4.32	<b>.04</b>
<b>SAT score group</b>	156.82	1	156.82	2.11	<b>.15</b>
<b>Interaction</b>	1.35	1	1.35	0.02	<b>.89</b>
Error	4170.00	56	74.46		
Total	4650.18	79			

<sup>a</sup> R<sup>2</sup> = .10, Adjusted R<sup>2</sup> = .06.

The analysis in Table 2 shows the statistically significant difference in exam scores between groups with and without clickers more clearly through the use of  $p$  as the level of significance ( $p$  less than 0.05 indicating less than a 5% chance of Type I error). The other two effects (SAT score and interaction of clicker use and SAT score) have  $p$  values greater than 0.05, indicating that no statistically significant differences were detected.

It is important to note that the convention of setting a cutoff for  $p$  of 0.05, or a 5% chance of committing Type I error, while commonly used in assessing the significance of statistical results, is not a hard and fast rule. Field (24) notes that 0.05 has “very little justification” (p. 78) in historic or modern statistical literature.

Instead, Field recommends using confidence intervals and effect sizes to judge statistical significance. An effect size is a way to report the magnitude of an effect in a standardized way. Another benefit of effect size is that it is not dependent on the sample size. Because large sample sizes are more likely to show highly significant  $p$  values, the effect size can be used to more objectively report if the significance of the  $p$  value should be considered important. There are many ways to calculate effect size. The reported effect size is dependent on the statistical test used. The most frequently reported effect sizes are Cohen's  $d$ ,  $r$ ,  $h^2$  (which is the same as  $r^2$ ), and  $w$  (sometimes reported as  $w^2$ ). Guidelines for effect sizes representing small, medium, and large effects can be found in statistical texts (24).

Table 2 thus presents inferential statistics indicating that there is a significant difference in achievement on a standardized teacher-written exam between students who did and did not use clickers ( $F_{(1, 56)} = 4.32, p < .05, = .23$ ), but no significant differences based on SAT scores ( $F_{(1, 56)} = 2.11, p > .05, = .13$ ) or the interaction of SAT scores and clicker use ( $F_{(1, 56)} = 0.02, p > .05, = .12$ ). The effect size for all of these comparisons is considered small and was obtained through a separate calculation, which can be found in statistics textbooks (24).

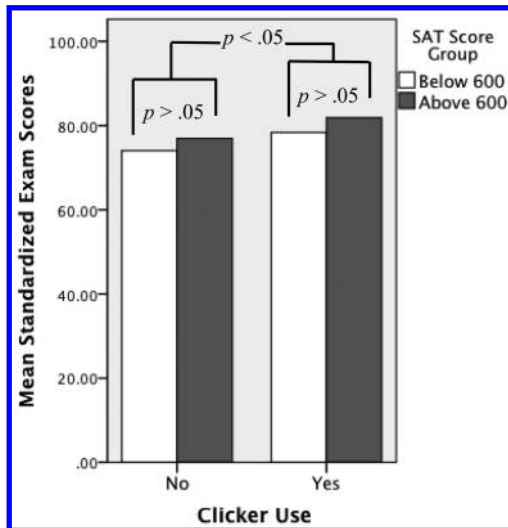


Figure 3. Bar graph of mean exam scores by clicker use and SAT score group with significant and non-significant group differences indicated.

Descriptive and inferential statistics can be used together for a more complete and easily interpretable result. Such a combination of descriptive and inferential statistics might use the same bar graphs as were used in the descriptive statistics but the differences in heights of the bar graphs can be marked as significant or non-significant based on the results of the inferential statistics, as seen in Figure 3. The significant  $p$  value indicates the significant difference between the combined with clicker and without clicker groups, regardless of SAT score group. This is the

main effect of clicker use shown in the first line of Table 2. The  $p$  values indicating no significant difference between low and high SAT score groups within the clicker use groups were determined from additional pairwise comparisons performed after the original overall statistical test. These additional comparisons are known as *post hoc* tests and are similar to  $t$ -tests. Information on *post hoc* tests and other ways to examine the results of factorial ANOVA analysis can be found in Field (24).

While adding this inferential information introduces additional visual complexity to the graph, it does provide a more defined interpretation of the differences between groups and allows for conclusions to be drawn more easily and accurately, regardless of the scale used.

### *Relationships*

Relationships such as regressions and correlations can be presented using descriptive statistics by creating a scatterplot to show whether or not two variables are associated with each other. This visual descriptive display does not necessarily indicate the strength of the relationship between the variables or if the relationship has statistical significance. Inferential statistics can be used to help interpret the scatter plot. In this case, either the parametric Pearson correlation coefficient ( $r$ ), or the nonparametric Spearman's rho, both inferential statistics using only one predictor (independent) and one outcome (dependent) variable, can be used (24). The benefit of using these inferential statistical tests in conjunction with the descriptive scatterplot is that the correlation coefficient, representing the slope of the line on the scatter plot, is standardized to a unitless value ranging between  $-1$  and  $+1$  and will have an associated  $p$  value indicating whether or not the relationship is statistically significant. With additional predictor variables, such as the SAT score variable in our example, more sophisticated analyses such as multiple regression or one of its variations can be performed. The use of multiple predictor variables usually results in a regression equation being presented instead of a visual scatterplot. More information on correlations and regressions can be found in most statistics textbooks (24).

### **Qualitative Research**

Presenting the data for qualitative research is dependent on which type of qualitative research (ethnography, narrative, phenomenology, grounded theory, or case study) is used. In phenomenology, results can be presented in the form of narration, tables, and figures. Tables could consist of frequencies of categories found in the data with one or more representative quotes included for each category. Figures might include a graphic demonstrating how the parts of the interpretation model fit together to explain the phenomenon under consideration. In grounded theory, a visual or graphic of the model or theory developed from the analysis of the data might be presented. The main presentation in each case is the accompanying narrative that explains to the reader how the conclusions were derived. By its nature, presentation of the results of qualitative research is more time and space intensive than reporting quantitative results.

## Summary

As we have seen in this chapter, the research question directly affects the research approach (qualitative, quantitative, or mixed methods); the evaluation plan developed, the tools selected, modified or constructed; the type of analysis of data performed; and the manner in which the results are presented. Thus the initial planning of a research question helps to develop a good match between research question and research approach; research approach and evaluation plan; evaluation plan and data analysis; data analysis and research question; data analysis and presentation; and research question and take home message. Careful and deliberate wording of the research question is key to the proper development of all other components in the research process. The research question is both the starting point and the end point of the planning process. The evaluation plan is what connects the research question to the results. The more highly integrated the research question and evaluation plan, the more effective the research.

## References

1. Duke University Academic Resource Center. *What is the difference between assessment and evaluation?* [http://duke.edu/arc/documents/The difference%20between%20assessment%20and%20evaluation.pdf](http://duke.edu/arc/documents/The%20difference%20between%20assessment%20and%20evaluation.pdf) (accessed Jun 25, 2014).
2. Bunce, D. M. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 976; American Chemical Society: Washington, DC, 2008; pp 35–46.
3. Bretz, S. L. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 976; American Chemical Society: Washington, DC, 2008; pp 79–99.
4. Creswell, J. W. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, 3rd ed.; SAGE: Los Angeles, 2013.
5. *User-Friendly Handbook for Mixed Method Evaluations*; Frechtling, J., Sharp, L., Eds.; National Science Foundation: Washington, DC, 1997.
6. Towns, M. H. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 976; American Chemical Society: Washington, DC, 2008; pp 135–148.
7. Talanquer, V. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 1166; American Chemical Society: Washington, DC, 2014; pp 83–95.
8. Abraham, M. R. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 976; American Chemical Society: Washington, DC, 2008; pp 47–66.
9. Bodner, G. M.; Orgill, M. *Theoretical Frameworks for Research in Chemistry/Science Education*; Prentice Hall: Upper Saddle River, NJ, 2007.
10. Yezierski, E. J. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 1166; American Chemical Society: Washington, DC, 2014; pp 11–29.

11. Cole, R. S.; Becker, N.; Stanford, C. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 1166; American Chemical Society: Washington, DC, 2014; pp 61–81.
12. Gron, L. U.; Bradley, S. B.; McKenzie, J. R.; Shinn, S. E.; Teague, M. W. *J. Chem. Educ.* **2013**, *90*, 694–699.
13. Case, E.; Stevens, R.; Cooper, M. *J. Coll. Sci. Teach.* **2007**, *6*, 42–47.
14. Bunce, D. M.; VandenPlas, J. R.; Havanki, K. L. *J. Chem. Educ.* **2006**, *83*, 488–493.
15. Bunce, D. M.; Havanki, K.; VandenPlas, J. R. In *Process Oriented Guided Inquiry Learning (POGIL)*; Moog, R. S., Spencer, J. N., Eds.; ACS Symposium Series 994; American Chemical Society: Washington, DC, 2008; pp 100–113.
16. Dalgarno, B.; Bishop, A. G.; Adlong, W.; Bedgood, D. R. *Comput. Educ.* **2009**, *53*, 853–865.
17. Fowler, F. J. *Improving Survey Questions: Design and Evaluation*; SAGE: Thousand Oaks, CA, 1995.
18. Scantlebury, K.; Boone, W. J. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 976; American Chemical Society: Washington, DC, 2008; pp 149–169.
19. Schwarz, N. *Soc. Cogn.* **2007**, *25*, 638–656..
20. Tavakol, M.; Dennick, R. *Int. J. Med. Educ.* **2011**, *2*, 53–55.
21. Tourangeau, R.; Rips, L. J.; Rasinski, K. *The Psychology of Survey Response*; Cambridge University Press: Cambridge, U.K., 2000.
22. Krosnick, J. A.; Presser, S. In *Handbook of Survey Research*; Marsden, P. V., Wright, J. D., Eds.; Emerald Group Publishing Limited: Bingley, U.K., 2010.
23. Bradburn, N.; Sudman, S.; Wansink, B. *Asking Questions: The Definitive Guide to Questionnaire Design- For Market Research, Political Polls, and Social and Health Questionnaires*; John Wiley & Sons: San Francisco, 2004.
24. Field, A. *Discovering Statistics Using IBM SPSS Statistics*, 4th ed.; SAGE: Los Angeles, 2013.
25. Sanger, M. J. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 976; American Chemical Society: Washington, DC, 2008; pp 101–133.
26. Lewis, S. E. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 1166; American Chemical Society: Washington, DC, 2014; pp 115–133.
27. Wainer, H. *Am. Stat.* **1984**, *38*, 137–147..
28. Pentecost, T. C. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series 1166; American Chemical Society: Washington, DC, 2014; pp 99–114.

## Chapter 3

# Making (and Using) Tests that Work: Cultivating Assessment Understanding To Support Teaching and Learning

April L. Zenisky\*

Center for Educational Assessment, University of Massachusetts, Amherst,  
119 Lathrop Street, South Hadley, Massachusetts 01075

\*E-mail: [azenisky@educ.umass.edu](mailto:azenisky@educ.umass.edu).

Tests are an integral part of evaluating the teaching and learning process. While education systems are at something of a crossroads in seeking a balance between instruction and assessment, it is critically important that instructors possess a working knowledge of tests and the test development processes in order to be able not only to create their instruments but to be able critically to evaluate, use, and explain the results of assessment and evaluation activities. This idea of a base competence with assessment principles forms the basis for the organization of this chapter, which draws on the 1990 *Standards for Teacher Competence in Educational Assessment of Students* (AFT, NCME, and NEA).

### Introduction

Today's multifaceted educational landscape is one in which accountability demands sometimes seem to be in direct competition with instructional priorities. In such a context, the very idea of tests and testing can evoke images of external mandates and irrelevance to the proceedings of actual classrooms. However, the present state of affairs in the science of test development - the field of psychometrics - is largely predicated on the alignment of curriculum, instruction, and assessment. Under this paradigm, in any given context, the curriculum defines the universe of the relevant knowledge, skills, and abilities. Instructors reference that universe in the development of lessons. When it comes to assessment, then,

any test to be constructed and/or used should be based on that curriculum and therefore assess the material taught. This principled approach provides a coherent and logical basis for appropriate test use, which is making inferences about what learners know and can do.

Given this background, the purpose of the present chapter is to provide a broad overview of test development practices and procedures, with special focus on test evaluation and test use. The chapter is structured around the *Standards for Teacher Competence in Educational Assessment of Students (I)*, guidelines that are now about 25 years old. They still serve a powerful purpose in framing the landscape of what educational assessment is (and, what it is not) in a way that is both accessible to and informative for practitioners. They also serve to provide a strategy for conceptualizing tests relative to test purpose and use. This chapter is organized around the seven principles that comprise these standards, and begins with an overview of tests and offers details on the typical steps in test development. These steps span a wide range of activities including test planning, development and assembly, administration, scoring, reporting, and use. Of special concern too is test evaluation, with guidance as to methods for determining test quality in the context of specific measurement needs. The final main section of the chapter builds on the ideas of the earlier sections to illustrate how test data can be represented and used in a variety of contexts, including results for individuals and for relevant groupings of learners.

## **An Overview: Teacher Competence in Educational Assessment of Students**

We begin with a nod to history: today's standards-based, accountability-driven educational system is largely viewed as a consequence of and reaction to the 1983 publication of *A Nation at Risk*, which is the report produced by former President Ronald Reagan's National Commission on Excellence in Education that called into question the preparation of America's youth to meet workforce needs in the global economy because of an educational system the report characterized as 'failing'. This document jump-started educational reform initiatives in the United States in the 1980s and 1990s, and the events that followed, including the National Educational Goals Panel meeting in 1989, emphasized the significant role of data - specifically educational test results - in evaluating student learning. It was against this backdrop of change that, in 1990, three groups - the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) - put forth the document entitled *Standards for Teacher Competence in Educational Assessment of Students*. The intent of these Standards was to provide guidance to both preservice and inservice teachers at all levels of education pertaining to the strengthening professional competency in the development and use of assessment in classrooms. These remain as relevant today as they were then, and indeed are almost prescient in highlighting the ways in which educators must be prepared to interact with assessment on a regular basis.

The AFT, NCME, and NEA *Standards* are organized around seven competencies, which these organizations viewed as the knowledge and skills critical to a teacher’s role as an educator. It is important to note that among these seven principles, some are more relevant to classroom assessment while others are connected more closely to ensuring teacher competence for participation in decisions related to assessment at various levels (school, district, state, and national levels, for example). The seven competences are listed below in Table 1.

The remainder of this chapter focuses on each of these standards in sequence, as a framework for understanding assessment options, test development, and appropriate test use.

It is recognized that the ideas presented above, in Table 1, were conceived of as being more directly relevant to K-12 educators. However, the principles espoused in these standards are broadly applicable across educational contexts, including the realm of higher education. The adoption of a principled approach to assessment (whatever form it takes) forms the basis for good evaluation and ultimately good decision-making. To reinforce this point, whenever possible throughout this chapter, examples and discussion will be couched in the higher education context.

**Table 1. The Standards for Teacher Competence in Educational Assessment of Students**

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3. The teacher should be skilled in administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.



## Standard 1: Choosing Tests

*Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.*

This standard is concerned with ensuring that teachers understand what assessment is and, perhaps most importantly, have the tools and skills to decide what assessment options are appropriate for the instructional decisions of interest in their own specific context. There are, as would perhaps be expected, numerous ways to think about the types of tests that are appropriate for use in educational settings. Stepping back for a moment, it is informative in this regard to consider the elements that together constitute what a test is. While there are many, many ways to differentiate various aspects of tests, here the focus is on defining tests by their purpose, by interpretation, and by format. These, among the various strategies, speak specifically to the types of data that instructors may need especially, given the educational decisions of interest. Some of the key considerations in each of these categories are given below.

### Assessment Types by Test Purpose

The idea of understanding tests on the basis of purpose is among the most fundamental ways of classifying assessments. It is rooted in the goal of understanding why an assessment is given and the kind of inferences about human performance that can reasonably be made on the basis of that test.

One clear and helpful perspective on assessment types relative to test purpose, interpretation, and use was put forth by Airasian and Madaus (2), who outlined four functional roles for assessments in educational settings. These are:

- *Placement assessment*: tests used to identify student knowledge at the outset of instruction
- *Formative assessment*: tests which are intended to monitor progress during instruction in a continuous feedback loop to both teachers and learners
- *Diagnostic assessment*: tests that effectively diagnose problems during instruction
- *Summative assessment*: tests used to assess achievement or improvement at the conclusion of instruction

As an approach to assessment development, these four types of tests are clearly differentiated from one another in that they span the range of assessment needs that are commonly present in educational settings, from the beginning of instruction to its conclusion.

### Assessment Types by Test Interpretation

A second critical distinction to be made between various tests is the notion of criterion-referenced tests (CRTs) and norm-referenced tests (NRTs). This distinction is predicated on the idea of test interpretation and how context is

given to a test score, because on its own, of course, a test score means nothing. On a norm-referenced test, the score an examinee receives is reported in terms of a relative position to the scores of other individuals in the relevant testing population. Norm-referenced tests typically provide both scale scores and percentile scores. A scale score is a transformation of a raw score onto a different scale for reporting, which is done to facilitate comparisons across different forms of a test (for example, year-to-year), while a percentile score is provided to help intended users understand that examinee's performance on that test instrument in the context of knowing where that score falls relative to other test-takers. These other test takers could be local (such as other sections of a course within an institution) or external (such as a national sample of students enrolled in a similar course across multiple institutions).

In reflecting on the interpretation of such norms, broadly this means that a learner who obtains a percentile score of 95 scored higher than 95% of test-takers in the reference group; a percentile score of 50 means the learner scored higher than 50% of people in the reference group (and lower than the other 50%). The choice of reference group, however, clearly impacts the types of inferences that can be made: typically external norm groups are larger and more stable, but could be less directly relevant as compared to local norms. The main criticism of the norm-referenced testing approach is that it does not support inferences about what a learner actually knows, as it is in essence designed to provide a rank ordering of performance independent of any external expectation of knowledge or skill.

The other type of test discussed here, criterion-referenced testing, is the approach to testing that fulfills that need for absolute interpretations, where examinee performance is independent of other examinees and is judged against mastery of content. Learner scores are not compared to other learners but rather are held against an absolute standard (the criterion) which can be defined relative to the knowledge or skill of interest that is being measured by the test. In criterion-referenced testing applications, which make up a sizable majority of tests used in many assessment contexts today, including higher education, what constitutes proficiency is determined during test development and in consultation with stakeholders such as educators. This typically occurs through a formal process of standard-setting, where groups of experts (often teachers, in educational settings) provide input to define acceptable performance at different performance levels (e.g., basic, proficient, and advanced; pass /fail). At the undergraduate level, especially for large courses, as part of course planning and syllabus development, there might well be some discussion of consensus about what constitutes an acceptable level of performance, with some degree of consensus reached.

### **Assessment Types by Item Format**

A third strategy for conceptualizing tests concerns the nature of the response provided by the examinee. This is most typically conceptualized as differentiating between selected-response items (most commonly, multiple-choice, but also including a range of other formats where the primary task is to choose from presented options) and constructed-response (item that involve

examinee-generated answers, including but not limited to written answers (short and extended text), oral questioning, performance assessment, and portfolios). It is important to note in this respect that the selected- and constructed-response labels are more commonly applied to item formats, and a test can, of course, draw on multiple item formats. This is a critical consideration in test development and evaluation, however, and the type of question asked impacts the nature of the information elicited from examinees, and as a result informs the nature of the interpretations that can be made about student knowledge.

The choice of item format is a critical element of test development and/or selection. Much has been made in recent years about a movement toward “performance-based assessment”, where test tasks are designed to elicit user-generated responses from test-takers (rather than a selection from presented options). There is nothing about either approach that is inherently better, however: it is the nature of the information sought that makes a test item format more or less appropriate. To know if a learner can write: ask them to write; to know if a learner can complete a spreadsheet, provide them with an opportunity to do so. Sometimes, it is necessary to ascertain if a learner has acquired factual knowledge, and there are question formats that can gather that information.

### **Other Ways of Differentiating Tests**

There are, of course, a few other ways to consider the nature of tests. One strategy is to differentiate *speed* tests from *power* tests, where the former typically involves presentation of a high number of items of relatively low difficulty (for example, simple math facts) to be completed within a specific time allotment for which quick recall is the measurement goal, while with a power test the intent for examinees is to show what they know without that time pressure. Accordingly, with power tests, there is likely to be wider variation in the difficulty level and no time limit (or, in practicality, a very generous one). The power approach aims to foster an environment where examinees can do their best without time pressure, for all intents and purposes.

The formality of a test is another distinction. Some tests are considered *standardized*, which Gallagher (3) defines as possessing the following characteristics:

- Developed by assessment specialists in collaboration with subject-matter experts.
- Field-tested and administered according to consistent and standardized procedures
- Scored and interpreted under uniform conditions

*Standardized*, in some ways, has evolved something of a pejorative meaning in the context of testing. This is likely due to increasingly high-stakes consequences associated with many tests and the widespread use (some would say overuse, per Downing (4),) of the multiple-choice format. However, the psychometric focus of standardization is ensuring consistency and integrity of the process across development and operational administration to promote

uniformity of interpretation. The test type held in contrast to standardized tests is a *teacher-made or classroom test*, which is developed in-house by an instructor or group of instructors to assess specific instructional objectives, typically with a greater degree of fidelity to what was taught.

## **Making Choices**

The preceding sections have provided a number of ways in which assessment options can be understood. There is a lot to consider, but all of the options presented are based on the idea that test users have options for tests, and these options are related to what users want to know about test-takers. Test purpose, test interpretation, item format, speed versus power, and formality level all contribute to support different types of interpretations and inferences about what students know.

## **Standard 2: Developing Tests**

*Teachers should be skilled in developing assessment methods appropriate for instructional decisions.*

The task of creating tests, in any context, is a significant one. Whether the test of interest is to be standardized for administration on a large scale or a local assessment for a classroom or institution, there are some broad principles of test development that can help guide the process and help ensure that the instruments developed provide users with the information desired, to support the proposed interpretations.

Downing (5) developed a listing of the “Twelve Steps for Effective Test Development”, which aims to systematize the process irrespective of specific testing context. That said, there are absolutely varying levels of formality and technical sophistication that are warranted depending on the testing context and purpose and the steps laid out here by Downing are defined so as to function as broad guidelines and not be prescriptive as to the expectations of the kinds of test development activities that are required for all instrument development efforts. These steps, accompanied by applications in classroom testing, are:

1. *Development of an overall plan*: This step helps to establish the most fundamental decisions about a test: here, the task is to define key elements of what is to be measured, test interpretations, format(s), and purpose, along with a timeline for development.
  - In the classroom, the plan is informed by material covered, the test purpose, and the kinds of information that instructors want to know about student learning.
2. *Content definition*: The domain to be assessed must be defined explicitly as the validity of inferences rests on the clarity of the content boundaries, from a validity perspective.

- Define the content (this week, this unit, this semester, this year) helps establish the boundaries for both instructors and learners.
3. *Test specifications*: Test specifications, also known as test blueprints, guide all of the test development activities relating to item development and form assembly by identifying test format, test length, the use of visual/audio stimuli as item components, and scoring rules.
    - This in essence is a plan for the test form. How many items, which item types, and what content is to be assessed is specified here.
  4. *Item development*: Here, item development is accomplished through well-established principles specific to item types, along with thorough training of item writers and a robust system for evaluating items (statistically and otherwise) before operational use.
    - Whatever item formats are to be used, the attention to detail in item development is critical. With selected response items, the stem and the distractors take on especial importance; with constructed-response items, the rubric requires considerable attention.
  5. *Test design and assembly*: The means by which test forms are constructed is likewise a critical step as it involves item selection for alignment to the operational blueprint.
    - This step is particularly relevant in classroom applications where an instructor has a bank of items to draw from, as the goal is to put together a set of items to assess the objectives specified.
  6. *Test production*: This step occurs irrespective of delivery mechanism (paper or computer), where a test is published and made available for use. A critical aspect of this step is also quality control.
    - For paper or computer-delivered tests, the items must be prepared in that medium, including instruction.
  7. *Test administration*: The types of issues that are relevant to this step include testing environment concerns as they relate to examinees with disabilities, proctoring, security, and timing.
    - Classroom assessments will vary in administration procedures, but the procedures in place for how tests are presented must be clearly established in advance of testing.

8. *Scoring of responses*: When scoring test responses, the key must be validated and quality controlled for, and item analysis carried out.
  - Scoring of student response is of course key as the outward-facing portion of testing, but for the instructor's own use, some review of performance in the form of item analysis can be tremendously informative for the instructor's own teaching practice, to understand how concepts were and were not understood by learners.
  
9. *Defining passing scores*: What constitutes acceptable performance must be determined with some logic and care.
  - There is a long tradition in classroom testing that scores of 70 constitutes a generally acceptable level of performance. This can be taken at face value, or instructors can choose to make the case for other levels of performance. The critical point here is that whatever passing score is implemented, that it be thoroughly considered and communicated.
  
10. *Reporting results*: The nature of the information that can be appropriately communicated to examinees depends on the choices made in prior steps (i.e., desired test interpretations, formats, and purpose), and this step also involves the timely release of accurate results.
  - In classrooms, there may be some specific learning management systems that are in use which may impact the nature of results reporting with learners. In any case, the data reported to learners should be at a relatively fine grain and actionable when possible.
  
11. *Item banking*: This step entails secure storage of test items for future use.
  - The principle of item banking was alluded to in Step 5, as over time a broad set of items can be built and drawn from as needed in specific classroom testing applications.
  
12. *Technical report*: Systematic and thorough documentation of all work completed related to test development is required in standardized testing. In other testing contexts, a less formal variation of a technical manual can be a useful reference or resource going forward.
  - Generally, the idea of formal technical documentation is unnecessary in classrooms. However, some compilation of decisions and their rationales can be helpful at a later date.

Again, while these steps encompass all of the formal work that goes into the development of a standardized test, the framework also provides a roadmap for the considerations that come up for classroom assessment. For example, “test production” reads as an oddly formal way to think about the specifics of test delivery in a classroom setting, but when assessing students instructors do have to make choices about how to get the items in front of the students, and whatever option is preferred or available to the instructor there are steps to be taken to make that happen (whether that is formatting an online delivery system or printing copies of paper tests).

## **Statistical Evaluation of Item and Test Quality**

The idea of evaluating items and tests from a statistical point of view is a key step in test development at all levels, including classroom tests. While the scale of large-scale standardized tests and classrooms tests is often quite different (and so the stability of the statistics is reduced, and less stringency is expected), item analysis can offer similar benefits to classroom assessments in terms of quality evaluations.

Gallagher (3) defines item analysis as “examination of the pattern or type of student response for each item of performance task in order to assess its effectiveness” (p. 326). The tasks of item analysis may vary in their formality according to the needs and intended use of an assessment, but when done in some form can provide key information that can help the test developer (whether a psychometrician or an instructor) revise a test to ensure that the items measure examinee knowledge as intended. By reviewing these data, users can ensure that items are individually and as a group operating at the intended level of difficulty, minimize the extent to which potentially problematic items are identified and cause disruptions during or after test administration, and provide developers, examinees, and users of test results alike a measure of confidence in the quality of the assessment.

There are perhaps three key analyses that are most helpful in the statistical review of test forms: item difficulty, item discrimination, and distractor analyses. These are discussed below.

### *Item Difficulty*

Item difficulty, as a basic statistic that speaks to the quality of test items, is an indicator of the proportion of examinees who successfully answered each item correctly. Other terms used for this index include *item easiness* or *p-value* (based on the idea of the statistic as a proportion). Computed by dividing the number of people who answered a specific item right by the total number who were administered the item, the resultant value ranges from 0.0 to 1.0 where lower values are items that are more difficult for the examinee group and higher values correspond to easier items.

Turning to the task of evaluating this statistic for each item in a test form, there is no absolute rule of thumb, because interpretation depends on an instructor's needs, their prerogative, and the content being assessed. If the content is relatively easy, then an instructor might well expect high item difficulty values for a given set of items. If the content of an item is particularly challenging or is relatively new, then perhaps the item difficulty would be lower. Computing the item difficulty statistic is clearly only a first step - it is necessary to review the data for anomalies, both with respect to possible gaps or misunderstandings in learner knowledge and also with an eye toward identifying potentially problematic items.

Note too that in terms of understanding what difficulty means, it is also important to consider that there may be some small impact of guessing on these statistics as well. Guessing occurs when a learner is not certain of the answer and chooses an option either based on elimination of one or more alternatives or at random. Interpretation of guessing is not quite so easy, however. When the test item of interest is a four-option multiple-choice question, the probability of simple random guessing is 0.25; for a five-option multiple choice item, it is 0.2. However, the observed proportions of learners choosing each option should be considered relative to the actual contents of those distractors. Some distractors are often better than others, and can be informative for teachers in the extent to which they highlight learners' proclivities to making common mistakes.

### *Item Discrimination*

In the context of item analysis, item discrimination is an important statistic that helps test developers identify items that differentiate between more capable and less-capable examinees. In essence, it characterizes the relationship between examinee performance on each item and their total score on the test, where a high item discrimination value indicates that high-achieving examinees do well on the item (as would likely be expected) and poorer examinees are generally not successful on the item. Item discrimination values range from -1.0 to 1.0, but in general items with low (e.g., lower than 0.2) or negative values must be carefully considered for inclusion on a test because they are statistically problematic. "Problematic" here means that high-achieving examinees are less likely to get the item correct than poorer examinees (which is likely due to a flaw in the item). There are several strategies available to compute item discrimination (6).

The kind of item discrimination values that should be expected in a given assessment depend on the nature of that assessment. Highly discriminating items are typically the more informative for spreading learners out along the score scale, and so that would be expected on a diagnostic assessment. In some mastery testing applications, lower discriminations are reasonable and expected.

In terms of reviewing items with low or no discrimination, it helps to look at the discrimination values and the item difficulty values together. Often, though not always, very easy or very hard items will have low or no discrimination. If all of the examinees are getting an item right, the item does not have any work to do to differentiate high and low-achieving learners; conversely, if no one answers an item right, there is also no differentiation of proficiency occurring.



## *Distractor Analyses*

As noted previously, distractor analyses can be a very important part of the work of test development, in the context of selected-response types of questions (specifically, multiple-choice). Distractor analyses seek to understand the performance of the incorrect response options. In test development, there is a basic assumption that the keyed correct answer is correct, and other response options for a given item are plausible but incorrect. At a straightforward level, a distractor analysis could be as simple as developing a frequency table and calculating the proportion of examinees who selected each option (which should add to 100% (including the percentage of omissions), when the keyed options are included).

When reviewing distractor statistics, there are a few things to consider. The rate at which certain distractors for a given item may be higher or lower than expected due to partial knowledge on the part of the examinee or a poorly constructed item that is confusing or keyed incorrectly.

## *Other Analyses of Interest*

One additional area of interest for evaluating test quality statistically involves reliability. Reliability matters in assessment because it characterizes the extent to which the instrument is consistent. This consistency matters in several respects, including consistency across tasks, across administrations, and across scorers.

Reliability is typically calculated through correlation. A few types of reliability metrics are described briefly below (Table 2).

The utility of each of these approaches to computing reliability is maximized when the count of students is quite large, and indeed larger than would typically be seen in most classroom situations. With formal standardized assessments, reliability coefficients are expected at or above 0.9, but in the area of classroom tests, reliabilities can be quite a bit lower (say, 0.4 or 0.5) without concern. These sorts of indices can be done, but should be taken in account alongside item analyses as described above, to provide a fuller perspective on test quality in the classroom given the data available.

The one other statistic of critical relevance to a discussion of reliability is the standard error of measurement (SEM). From the preceding review of reliability, it is clear that consistency is a desirable property of a test. The extent to which such consistency is present can be quantified using the SEM, because when a test has low reliability, there are likely to be large variations in examinee performance (expressed as high SEM values), and tests with high reliability provide results that are considerably more precise (associated with low SEM values).

Understanding SEM is largely task specific to a specific assessment. The main idea for understanding how to use SEM is that a test score, in and of itself, contains error. Imagine two students take a test and suppose one attains a score of a 36 and the other, 44. Ostensibly, these students are showing different levels of achievement. But, if the SEM on the test is 5, the true range for the first student's score is 31 to 41, and the range for the second student is 39 to 49. Now, their

skills do not appear to be quite so different, as they overlap a bit. Imagine further than the test has a cut-score of 40. The first student failed by obtaining a 36, but in considering the SEM, there's some possibility that the true skill of that individual is on the passing side of that cut-score. Similarly, the second student passed, but reflection on the SEM and the confidence band for that student shows that the "pass" could have been a "fail". The SEM helps to provide context for understanding student scores both in relation to one another and in terms of levels of knowledge and skill.

**Table 2. Reliability Indices**

<i>Reliability Type</i>	<i>Method and Procedure</i>
Stability	Test-retest reliability is evaluated by administering the same test to the same group before and after an interval. This approach estimates reliability of an instrument over time.
Equivalence	Equivalent forms reliability is evaluated by giving two different forms of a test to the same groups within a small time interval. This approach estimates the consistency of different forms of the same test.
Internal consistency	Split-half reliability is evaluated by administering a test once, dividing the test into two equivalent halves randomly (such as odd- and even-numbered items), and correlating the two halves. This approach establishes the reliability of two halves of a single test, which should be essentially equivalent to one another.
Consistency across raters	Interrater reliability is established by having two or more raters score a set of open-response items and correlating the judges' scores. This approach establishes the reliability of scorers for constructed-response items.

### **Standard 3: Giving and Understanding Tests**

*The teacher should be skilled in administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods.*

The expectations for teachers in the area of administration, scoring, and interpreting results speaks not only to the process of handing out tests and watching students complete items but also to the very purpose of assessment: to make appropriate inferences about what students know and can do, and to identify actions to be taken on the basis of results.

Historically, the psychometric community has regarded facilitating test interpretation as something of an afterthought to the test development process. It was generally viewed as reasonable to produce technically superb tests with low

standard errors of measurement, but when transmitting scores little or no context would be available to assist test users with using that information in a practical sense (7). Happily, however, the tide has turned with the advent of the era of accountability; with more data has come the expectation that scores have value and should be used. Much research and operational attention has gone into efforts to make test score reporting systems that are usable for a wide variety of intended audiences.

For instructors, there are a number of components of reporting that have direct bearing on the usability of reports. It is first important to note that reports of test performance can be created for individuals or for groups. Groups are typically composed on a hierarchical basis, beginning with a group such as class, then rolling up to other groupings such as multiple sections, a grade or year, a school, a district, a state, and a nation. Group reports can be conceptualized as list-style where individual results for all examinees in the list are provided in a sequence or in aggregate (where performance across individuals is summarized to make inferences about the group as a unit).

The next critical consideration in understanding reports is the choice of scale or metric used to express results. A test “score” is not a monolithic concept; rather, it encompasses a range of strategies for quantifying performance on an assessment. A few of the types of scores that may be provided on test score reports include:

- Raw score: This score summarizes the responses made by an individual (typically) as the total number of points earned by the student, typically expressed relative to the total points available.
- Percentage correct: This score is computed as the total number of correct answers relative to the total possible score available, and is expressed as a percent on a scale from 0% to 100%.
- Scale score: This score is a conversion of the raw score onto a scale that is common to all test forms for that assessment. It is done to facilitate interpretation of scores across multiple versions of the same test (such as from year to year or form to form).
- Percentile rank: This score type conveys information about an examinee’s relative position in a group in terms of performance. A percentile rank of 75 indicates that the individual scored higher than 75 percent of examinees who were in the reference group on which the percentiles were calculated.
- Grade equivalent: This score offers a strategy for understanding performance defined as the grade level at which a typical student obtains a specific raw score, computed using the performance of a norm group to establish what performance is typical at what point in the academic year. This score type is typically used in large-scale assessment with national norms in the K-12 assessment setting.
- Normal curve equivalent (NCE): This score is a normalized score where the score scale is created to have a mean of 50 and a standard deviation of 21.06 (selected to ensure that the NCE scale score range is 1 to 99).

Understanding test performance also requires that users have a working knowledge of how to use context to give meaning to scores. It is only through context that any of the scores listed above have meaning: for example, a grade equivalent score of 3.6 on a reading assessment only becomes meaningful if a test user knows what grade the test-taker is in and also has an understanding of what it means for a child to be reading at that level. If a fifth-grade student obtains that 3.6 grade equivalent score, that would perhaps be cause for concern; in contrast, a second grader who earns a 3.6 might be characterized as doing relatively well.

Some further sources of context for test scores that can help users to make sense of results include measures of central tendency, dispersion, reliability, and errors of measurement. Central tendency, of course, helps to add context by using indices such as the mean, median, and mode to characterize performance of a group. (Scores for an individual can also be held against those indices as an indicator of relative performance, of course.) Dispersion describes the spread of scores obtained and two examples of dispersion are standard deviation and standard error of measurement. The reliability of a test provides users with an idea of test quality as it relates to the consistency of a test to measure knowledge, skills, or abilities time and again, so that results are due to proficiency and not the product of spurious measurement error. As noted previously, the SEM of a test is an estimate of the amount of error associated with a test score. For a test with high reliability, the SEM is lowered; with less reliable tests, the SEM is higher. SEM is a statistic associated with test scores that can be used to compute a confidence interval around a test score, to provide a band within which a student's true score (8) is likely to occur.

One additional consideration of context in reporting is through the use of display strategies for communicating test results. Many agencies have relied on simple numerical listings, but increasingly are turning to a wider range of strategies to illustrate performance that integrate creative graphics to facilitate interpretations. For example, line graphs are often used to assist users in displaying mean performance for groups over time.

An interesting approach to displaying comparison scores is found in the online reporting system for the National Assessment of Educational Progress (NAEP, online at <http://nces.ed.gov/nationsreportcard/>). NAEP has long been a leader in reporting efforts (9), and the National Center for Educational Statistics (NCES) has a number of displays in use that illustrate particularly innovative approaches to communicating performance between states. When a user interacts with the online reporting tools, that person can click to make any jurisdiction assessed by NAEP the reference group, and the map then automatically populates with color to illustrate performance of all other jurisdictions to clearly show those with statistically higher or lower average performance, or those that are statistically equivalent to the reference group in performance. These displays are often changeable by users, where (for example) National Public schools can be set as the reference group in some applications, and the performance of all other states are coded relative to that (or vice versa).

## **Standard 4: Using Tests**

*Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school.*

As an area of competence in assessment, it is critically important that institutions foster an environment where teachers have the tools and skills to use assessments effectively for specific needs. This of course connects to other standards, especially the first where the nature of the test to be used is clearly delineated and the testing goals explicitly stated.

There are a wide variety of ways that tests can be used in educational settings. When focusing on individuals, test results can be helpful for a) instructional planning (including differentiating between students and helping to identify areas in need of remediation), b) individualizing instruction, c) identifying the needs of exceptional students, d) monitoring progress over time, e) informing families about student achievement, and f) providing information to help students make decisions about future education and career options. Rolling up results for groups, such data provides insight for not only instructional planning but also program evaluation at different levels of aggregation.

The recommendations from the Standards in this area suggest that skills relevant here include the ability to use test data in an actionable way, to make plans for appropriate next steps and to avoid misconceptions and misuses. By reflecting on individual scores, instructors can identify particular needs for individuals, but aggregate test score data at the total test level and in content subareas of interest can help instructors identify broad areas of strength and weakness among groups of students.

## **Standard 5: Grading Learners**

*Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.*

At times, the idea of assessment seems quite formal and quite distanced from the reality of grading as an ongoing and necessary part of evaluation of performance. However, in practice, grading is an integral activity in classrooms that speaks to an instructor's proficiency in understanding what constitutes learning and monitoring achievement over time. This standard addresses expectations about not only being able to competently evaluate student work, however; it also lays out ideas relating to processes and procedures. It is necessary for teachers to have in place structures that clearly explain how the various assessment mechanisms used in classrooms (such as, but not limited to quizzes, tests, project, activities, and other assignments) are together used to characterize student performance. Above and beyond that, grading practices should be fair and rational and be defensible.

In developing an approach to grading, there are several factors to consider. There may be established policies at institutions that impact the extent to which individual instructors can develop their own approaches. Also, personal philosophies about grades can affect how instructors look at the grading process. This emerges through reflection on what grade symbols are used, what aspects

of performance are included in grades, and how elements of work could be combined to represent overall performance.

## **Standard 6: Communicating Results**

*Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.*

In the process of developing a working level of expertise with respect to these various competencies, one goal articulated here in this standard is for instructors to become familiar with the language of assessment and with the task of talking about tests with stakeholders such as students and students' families. Whether referencing results of formal/standardized assessments or informal/classroom practices, teachers should be prepared to discuss interpretations with these stakeholders and communicate how test data may impact actions taken with respect to a student's educational experience.

Communicating results is an opportunity for instructors to connect with students, and while the number of test takers impacts the extent to which communication can be in person, there are ways in which test results can be communicated back to learners that foster action and improvement where appropriate. Some examples of report data that can be communicated include score breakdown by subarea, with points earned versus points available shown. It can also be helpful to illustrate how learners did on items of different formats, and where available, provide correct answers to questions answered wrong.

## **Standard 7: Promoting Fairness**

*Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.*

The final standard of interest to be discussed here concerns fairness. Fairness enters the assessment process both in the development of assessments (in terms of methods) and in test use.

### **Fairness in Test Development**

On the side of development, it is necessary to consider how the activities undertaken to build, administer, and evaluate a test promote equity across test takers. The principles of universal test design, which seeks to apply the idea of "the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (10) to the realm of testing, in the following ways (11):

- *Inclusive assessment population:* A test instrument must be conceived of as being accessible for the entire population of examinees (being mindful of opportunity considerations for sub-populations such as individuals with disabilities, limited English proficient students, and other groupings based on social or racial/ethnic group membership).

- *Precisely defined constructs*: A test should measure what it intends to measure, and minimize the extent to which it measures other information (construct-irrelevant variance).
- *Accessible, non-biased items*: As part of item development and review, test developers must institute clear procedures for ensuring the quality, clarity, and a lack of ambiguity of items, as well as have in place both quantitative and qualitative mechanisms for evaluating items with respect to potential sources of bias.
- *Amenable to accommodations*: Objectively, accommodations refer to any changes in content, format, or administration to facilitate completion of a test form by individuals unable to take a test under the original standard conditions (12).
- *Simple, clear, and intuitive instructions and procedures*: The nature of the task and the method for providing a response should be expressed in a way that is simple, clear, concise, and understandable to test takers. This likewise applies to directions for test administration.
- *Maximum readability and comprehensibility*: The formal idea of readability is the likelihood that text is comprehensible by a particular group of individuals, and can be calculated using various text features. In the context of fairness, as relevant here, readability involves ensuring that the test materials are presented at a level and in a format that is accessible and understandable for the examinee population of interest.
- *Maximum legibility*: The contents of a test form should be able to be read with ease by examinees, with an eye toward not only the font but also text size and the design and layout of the page (including any item elements such as tables, passages, and graphics).

Each of these elements offers a way for test development procedures to consider the full range of test-takers from the outset of test development, and ensure that the testing instrument and its administration procedures are appropriate for all learners. It is important to note, however, that the points listed above are indeed conceptualized in the high-stakes, standardized testing realm. The direct applicability of some of these tenets may be lower in classroom settings, but they are indicative of the issues that can arise and what the guidance from the psychometric field is for handling such issues. Balancing these ideals with practical considerations poses its own set of challenges, but accessibility can be prioritized in many ways, both big and small.

### **Fairness in Test Use**

On the side of use, the focus on fairness is on the appropriateness of interpretations and the extent to which decisions made on the basis of test scores are supported by evidence. At a basic level, from the outset of test development, a test is developed to accomplish a specific purpose. The methods and decisions made for a test that is summative in nature are quite different from those made for a test that is intended to provide diagnostic results, which is different yet again from a test that is to be used for placement. This invokes the critical topic of

validity, which concerns “the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by the proposed uses of tests” ((12), p. 9).

There are of course multiple ways to think about validity evidence, including in the context of test content, response process, internal structure, relationships to other variables, and consequences. Validity evidence comes in many shapes and sizes, and should be thought of as a property that occurs as a matter of degree, not as a categorical yes-or-no. For instructors, the main point to be made about validity is that it is important that they develop an awareness of what a specific test instrument that they are using can and cannot be used for.

## Final Thoughts

The reality of today’s educational system, from early education through graduate and professional preparation, is that assessment is integral. For instructors, establishing a base understanding of the principles and practices of assessment from the perspective adopted in this chapter (of being an informed and aware developer and user of test data) is frankly necessary. The competency-based approach advocated for by the AFT, NCME, and NEA in the development and publication of these *Standards* offers an important framework for understanding the dimensions of assessment as they impact classrooms. The primary goal for educators is to view assessment as a vital and connected part of their work.

To this end, the key underlying theme of these standards - even nearly twenty-five years after their original publication - remains knowledge and understanding in communication. Educators are quite often the first line of communication to help examinees themselves (and other stakeholders, including families) put test results in their proper context and to help them identify the appropriate next steps (whatever those might be, depending on test context and intended uses of scores). Whereas tests and testing can undoubtedly be a complicated topic for both practice and discussion, current expectations of test use likewise necessitate preparation and training in assessment as a matter of their responsibility to learners alongside curriculum development and instructional methods.

## References

1. American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), National Education Association (NEA). *Standards for Teacher Competence in the Educational Assessment of Students*; American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), National Education Association (NEA): Washington, DC, 1990.
2. Airasian, P. W.; Madaus, G. F. *Meas. Eval. Guid.* **1972**, *4*, 221–233.
3. Gallagher, J. D. *Classroom Assessment for Teachers*; Merrill: Upper Saddle River, NJ, 1993.
4. Downing, S. M. In *Handbook of Test Development*; Downing, S. M.; Haladyna, T. M., Eds.; Erlbaum: Mahwah, NJ, 2006; pp 287–302.



5. Downing, S. M. In *Handbook of Test Development*; Downing, S. M.; Haladyna, T. M., Eds.; Erlbaum: Mahwah, NJ, 2006; pp 3–26.
6. Haladyna, T. M. *Developing and Validating Multiple-Choice Exam Items*, 2nd ed.; Erlbaum: Mahwah, NJ, 1999.
7. Zenisky, A. L.; Hambleton, R. K. *Educ. Meas.: Issues Pract.* **2012**, *31*, 21–26.
8. The concept of true score is key to interpreting test scores. A test provides a test score which must be understood as an indicator of proficiency based only on that which can be observed. A test score cannot be accepted as an absolute gauge of true, innate knowledge because an assessment requires a demonstration of performance and a test is an instrument that provides a means for evaluation of that demonstration. Error enters the measurement process in a range of ways (through the person or through the instrument) and the job of psychometrics is to minimize that to the extent possible through quality test development and administration procedures.
9. Zenisky, A. L.; Hambleton, R. K.; Sireci, S. G. *Appl. Meas. Educ.* **2009**, *22*, 359–375.
10. Center for Universal Design. *What is Universal Design?* Center for Universal Design, North Carolina State University: Raleigh, NC, 1997.
11. Thompson, S. J.; Johnstone, C. J.; Thurlow, M. L. *Universal Design Applied to Large Scale Assessments*; University of Minnesota, National Center on Educational Outcomes: Minneapolis, MN, 2002.
12. American Educational Research Association, American Psychological Association, & National Council of Measurement in Education *Standards for educational and psychological testing*; AERA: Washington, DC, 1999.

## Chapter 4

# General Statistical Techniques for Detecting Differential Item Functioning Based on Gender Subgroups: A Comparison of the Mantel-Haenszel Procedure, IRT, and Logistic Regression

Lisa K. Kendhammer<sup>1</sup> and Kristen L. Murphy<sup>\*,2</sup>

<sup>1</sup>Department of Chemistry, University of Georgia,  
140 Cedar Street, Athens, Georgia 30602

<sup>2</sup>Department of Chemistry and Biochemistry,  
University of Wisconsin-Milwaukee, 3210 N. Cramer Street,  
Milwaukee, Wisconsin 53201

\*E-mail: kmurphy@uwm.edu.

Many decisions are made based on assessment results from course grades to entrance into programs. The data from assessments, therefore, must be fair and valid. The validity and fairness of the data should also be considered by subgroups for the test overall and by individual test items. Differential Item Functioning (DIF) occurs when subgroups that are matched on equal abilities perform statistically different on an assessment item. Detection of DIF items can be used to ensure the data from examinations are as fair and valid as possible. Not only is detection of items that exhibit DIF important for making sure the data from the assessments are fair, but also for educators and researchers to understand better the reasons why these items are exhibiting DIF. While various methods are used to determine DIF, three of the more common ones include the Mantel-Haenszel procedure, the Item Response Theory model, and the Logistic Regression model. An overview of each method, as well as a comparison of the results, will be presented.

## Introduction

There are many motivations to study Differential Item Functioning (DIF). DIF occurs when subgroups that are matched on equal ability perform statistically different on an assessment item. First, it is known there is the movement from many national organizations for equality in education between subgroups of students. Examples of subgroups can include gender, race, ethnicity, socio economic status, etc. However, why is the focus on individual items? Suppose there were three items on a 40-item assessment that exhibited DIF that favored one subgroup of students. This would mean that subgroup of students could potentially score three points higher on the assessment than the equally ability level students in the other subgroup would. It would not be considered fair towards the unfavored subgroup of students and could potentially be detrimental. This leads to a second motivation that certain decisions are often based on the scores that students receive on these assessments. If everyone strives for education and career options to have greater equality, it is important to make sure the assessments that are guiding education and career options produce data that are as fair as possible.

Although the terms gender and sex are often used interchangeably they do have different meanings. To keep in line with current literature the term gender will refer to the actual biological differences of sex and not the social constructed identity (1, 2). Gender equality is a phrase that many people encounter daily. These two words may bring different thoughts and different emotions to mind depending on context in which they are thought of or spoken. In education, one of the terms most often heard relating to gender equality is gender parity. Gender parity refers to both male and female students having equal opportunity to enroll in school (3). When looking at the number of female students enrolled in an institution compared to the number of male students enrolled that is often referred to as the Gender Parity Index (GPI) (4). GPI is a statistic that in most countries is easily attainable and is often one of the first places to look when investigating gender equality in education. In the United Nations 2013 report on the Millennium Development Goals there are 8 goals that are hoped to be met by 2015, and one of them specifically focus on gender equality. Goal 3 is to “promote gender equality and empower women.” The goal focuses on education, wage-earning jobs, and women with governmental positions. For education there is near equality of genders in primary education, but not as much with secondary or tertiary education. However, this is region dependent and in some areas there are more female students enrolled in higher education than male students (4). Another agency that is interested in gender equality is the United Nations Education, Scientific and Cultural Organization (UNESCO) which has the UNESCO Priority Gender Equality Action Plan for 2014-2021 (5). When looking at the GPI, it gives information mainly about the students who attend the school but not how they are treated. To understand truly if schools are achieving gender equality one must look much deeper. This could involve not only the learning content, curriculum, and textbook material but also the attitudes of the students and teachers. While some of these measures may be more difficult to study, a readily available measure is how the students perform on different assessments

(3). There have been many studies in the last 50 years that investigated gender differences on all types of assessments. These studies have investigated gender differences on different subjects, content areas, test structure, item structure, and item order, as well as others. It has been recognized that there are generally no differences based on gender subgroups on overall assessments testing multiple skills and abilities (6). However, an overall gender equality may be masking inequalities when considering specific content areas or format types of items and more investigation into differential performance by gender subgroups, by components of assessments, or by test questions is warranted.

## **Gender Differences**

In 1974 Maccoby and Jacklin did a well-known review looking at many suppositions of gender differences including that verbal abilities tend to favor female students and quantitative and spatial abilities tend to favor male students. The age-old notion that males perform better in math and science was also investigated. The study confirmed the previous conclusions that questions on math and science examinations that involved visual-spatial factors tended to be favored by male students whereas questions that involved verbal skills tend to be favored by female students (7). A study conducted by Cleary had similar results, stating the female students across all age groups performed better in verbal tasks, whereas, male students across all age groups tended to perform better on science tests (8). More than twenty years after the Maccoby and Jacklin study, Nancy Cole and Educational Testing Service (ETS) conducted a study that looked at gender differences in education. It was found that the previously identified gap between male and female students in math and science had closed. In fact, when looking at entire subjects there was no difference in overall performance by gender subgroup. When considering aspects of assessment items or skills, however, there were some differences between genders that were found. First, verbal skills still seemed to favor female students as was reported in the Maccoby and Jacklin study. Second, when looking at open-response questions there was sometimes no favor, while other times the direction that was favored was split. Questions that required a written response tended to favor female students. However, open response questions that required responses with the construction of a figure tended to favor male students (6). These results suggest that some items involving visual-spatial skills may show an advantage for male students even though earlier results showed that visual-spatial skills had a very small difference favoring male students. This may result from the combination of the item's format being open response and the addition of the visual-spatial component. The result that open response questions with a written component favored female students strengthens the notion that female students do better with verbal skills. Another interesting finding in the Cole study was that gender differences tended to increase with the students' age (6). This means that even if the differences don't occur at a younger age, once students reach high school, college, or graduate level these differences could increase, making them more of a concern. This conclusion has been reached by other researchers as well (8-10).

Aside from these broad, large studies, there have also been many others studying both gender differences in content and format of the items. There are many layers when looking at the content of items. One could look broadly at the subject in general such as science, English, or humanities. While this may be interesting information it doesn't go into enough detail on which specific areas these differences occur. Also, many studies show that although there may be gender differences within subjects, there are more variables to consider than content areas. The results from different studies about the gap between genders in science and mathematics are inconclusive overall in whether this gap is real. In addition, some studies have shown that certain areas of science also have some gender differences. One difference is that female students tend to do better on health-type questions (9, 10) whereas male students out-perform at chemistry and physics, especially in the areas of electricity and mechanics (6, 9, 11). One of the areas in science where female students have been shown to excel over male students is in science-inquiry-type questions (11).

Besides gender differences based on the content of the assessment, another area of consideration with regards to gender differences is the format of the items. Like the study by Maccoby and Jacklin, many studies have shown that male students tend to do better on items that involve visual or spatial skills. Hyde and Linn conducted a meta-analysis where they studied gender differences on three aspects of spatial ability which included spatial visualization, mental rotation, and spatial perception. Of the three, two of them were found to have a gender difference that favored male students, with spatial visualization showing no difference (10). This phenomenon was found in other studies as well, including a later study by Linn which found that while the gap is narrowing for spatial visualization, a difference in performance was still found for spatial reasoning. Another study found there was an advantage on most visual skills that favored male students (12, 13).

Although there is some alignment between these studies and the determination of favor with regards to general assignments of content and format, there are also some conflicting findings between these studies. This is possibly due to gender differences that occur more on specific content areas and formats. This suggests that instead of looking at tests that cover broad aspects of a domain or multiple content areas, instead gender differences should be studied on individual items within one content area. In order to conduct this analysis meaningfully, one must compare the performance of a group of male students and female students who are matched at equivalent intellectual abilities and examine for a performance difference on a single assessment item. When that item shows a statistical difference, that is known as differential item functioning (DIF) (14). DIF can be used not only to examine gender differences on an item, but can be used for any subgroup such as race, ethnicity, socio economic status, religion, etc. Many large testing associations are using DIF to vet questions on assessments based on gender and other subgroups to ensure test equality (15, 16). Along with large-scale assessment publishers conducting DIF analysis, other researchers are examining DIF on assessments as well. Examples include Hambleton and Rogers who investigated DIF based on race on the New Mexico High School Proficiency Exam (17); Schmidt and Dorans who investigated DIF based on race on the SATs

(18); and a gender DIF study conducted on the National Education Longitudinal study of 1988 (19), along with others (20, 21).

## **Statistical Methods for Detecting Items that Exhibit Differential Item Functioning**

Considering the importance of examining for DIF, especially on assessments that have an effect on a person's academic future or career, it is also important to consider different statistical methods that can be used to study DIF. A graphical representation of the performance of an item, similar in characteristic to an item characteristic curve that we will call an item plot, can aid in understanding these differences.

An item plot is constructed from the probability of the item being answered correctly based on either the students' standardized score on the assessment or on a latent trait such as the students' ability level. Groups of students can be then separated into their subgroupings and plotted separately. When comparing item performance between subgroups a single graph can be generated with both plots where the differences can then be examined (22). When there is no DIF, the two plots will be the same (see Figure 1a). When the two plots are not the same, this indicates the presence of DIF. Uniform DIF is present if the two plots look similar but one is consistently higher than the other. This is indicating that the probability of one subgroup outperforming the other is consistently greater across all ability levels (see Figure 1b). Nonuniform DIF is present if the two plots cross. For example, in Figure 1c, for the students with lower abilities there is a higher probability of male students outperforming female students. However, for students with high abilities the performance switches and now there is a higher probability of female students outperforming male students (see Figure 1c).

Item plots are useful to visually inspect if an item could possibly exhibit DIF and help determine if further investigation is needed. There are a number of different methods that can be used to detect items that exhibit DIF such as Standardization (23), the Mantel-Haenszel procedure (24), Logistic Regression (25), SIBTEST (26), and Item Response Theory (IRT) methods (17). While each of these methods have been used, the Mantel-Haenszel procedure, Logistics Regression and IRT are the more common methods and will be the three that are discussed below (17, 23, 25, 27). All of these methods will identify items that exhibit DIF, but they use different theories to do so and vary in their levels of sensitivity of detection.

### **The Mantel-Haenszel Procedure**

The Mantel-Haenszel procedure is one of the most commonly used methods to detect DIF. It is readily available in many statistical packages and fairly easy to use (23). The Mantel-Haenszel procedure was introduced as a method to determine DIF by Holland and Thayer (24). The Mantel-Haenszel procedure is a method which uses a  $2 \times 2$  contingency table to determine the probability of one subgroup answering the item correctly versus the other subgroup. There is

an additional part of the contingency table which is sometimes represented as an  $m$  (or  $K$ )  $2 \times 2$  contingency table, meaning that this table is represented for some matching criterion. Most often this is represented by the score they received on the assessment. For each item on a 100 point assessment, a  $2 \times 2$  contingency table would be constructed for all the participants who earned 100 points, then for all participants who earned 99 points and so on. An example of this contingency table is shown in Table 1. The Mantel-Haenszel common odds ratio is shown in Equation 1, where  $p$  is the proportion of participants who got the item correct, and  $q$  is equal to  $1-p$ . Both representations from Table 1 and Equation 1 are from the article DIF Detection and Description in the book Differential Item Functioning (14).

$$\hat{\alpha}_{MH} = \frac{\sum_i p_{r_i} q_{f_i} N_{r_i} \frac{N_{f_i}}{N_i}}{\sum_i p_{f_i} q_{r_i} N_{r_i} \frac{N_{f_i}}{N_i}} = \frac{\sum_i a_i d_i}{\sum_i b_i c_i} \quad \text{Equation 1}$$

The Mantel-Haenszel Chi-Squared statistic determines if an item is favored by one subgroup (at a certain test score or interval) by testing if the proportion of the correct responses from one subgroup is the same as another (22). If not then that item exhibits DIF. This statistic is calculated by Equation 2, where  $m$  is equal to the score level of the studied item.

$$MH - \chi^2 = \frac{\left[ \left| \sum_m a_m - \sum_m E(a_m) \right| - 0.5 \right]^2}{\sum_m Var(a_m)} \quad \text{Equation 2}$$

$$\text{where } E(a_m) = E(a_m | \alpha = 1) = \frac{N_{rm} N_{1m}}{N_m} \quad \text{Equation 3}$$

$$Var(a_m) = Var(a_m | \alpha = 1) = \frac{[N_{rm} N_{1m} N_{fm} N_{0m}]}{[N_m^2 (N_m - 1)]}$$

The -0.5 in equation 2 is used as a continuity correction. According to Holland and Thayer (pg 134) the continuity correction is used to “improve the accuracy of the chi-squared percentage points as approximations to the observed significance levels (24).” While the Mantel-Haenszel procedure is commonly used to detect uniform DIF, other methods are more useful in detecting items with nonuniform DIF (17, 23, 25).

### Item Response Theory

Another common method for detecting DIF is item response theory. Item response theory (IRT) is a model that predicts how the students should respond to the item based on their latent trait ability (22). Instead of using the pure raw scores like the Mantel-Haenszel procedure, IRT converts the raw scores into

log-odds therefore, transforming the non-linear data into linear data. There are three common models for dichotomous data; the one parameter model, the two parameter model and the three parameter model. The standard logistic function for item response theory is shown in equation 4.

$$P_i(\theta_s) = \frac{e^x}{1 + e^x} \quad \text{Equation 4}$$

For this equation,  $P$  is the proportion of subjects ( $s$ ) with an ability level of  $\theta$ , who answered the item ( $i$ ) correctly and  $x$  is a representative symbol that will change depending on the parameter model. An item response model for dichotomous data has three main parameters of interest: the item discrimination, the item difficulty, and a pseudo-guessing parameter.

The one parameter model accounts for the item difficulty only as a result of the latent trait ability (28). The difficulty of the item (or item location) is represented as  $b$ , and is a measure of how hard the item was for that group of individuals (22). For this case the  $x$  in equation 4 would be given by equation 5, where  $D$  and  $a$  are constants.

$$x = Da(\theta_s - b_i) \quad \text{Equation 5}$$

The second model is much like the first except this time the item's discrimination is accounted for as well which is shown in equation 6. The discrimination of an item or the slope (represented as  $a$ ) is a way to determine how well the item distinguishes between students with a high ability level and those with a lower ability level (22).

$$x = Da_i(\theta_s - b_i) \quad \text{Equation 6}$$

The third model accounts for the item's difficulty, discrimination and the pseudo-guessing parameters as seen in equation 7. The pseudo-guessing parameter (represented as  $c$ ) accounts for students being able to guess correctly on multiple-choice or true-false type questions. With other types of questions the probability of answering the question correctly approaches zero at lower latent trait ( $\theta$ ) values (29). However, if the students are able to guess the probability would instead approach a higher probability in accordance with the number of available options (i.e. 25% for a multiple-choice item with four responses).

For three parameter fit the equation represented by  $x$  is the same as the two parameter fit, except there is also the addition of a pseudo-guessing factor.

$$P_i(\theta_s) = c_i + \frac{(1 - c_i)e^{Da_i(\theta_s - b_i)}}{1 + e^{Da_i(\theta_s - b_i)}} \quad \text{Equation 7}$$

However for the detection of DIF, the guessing parameter often produces a large standard error and can negatively impact the power of detection. Therefore when investigating for DIF, it is common to compare only the  $a$  and  $b$  parameters for the participants (30).



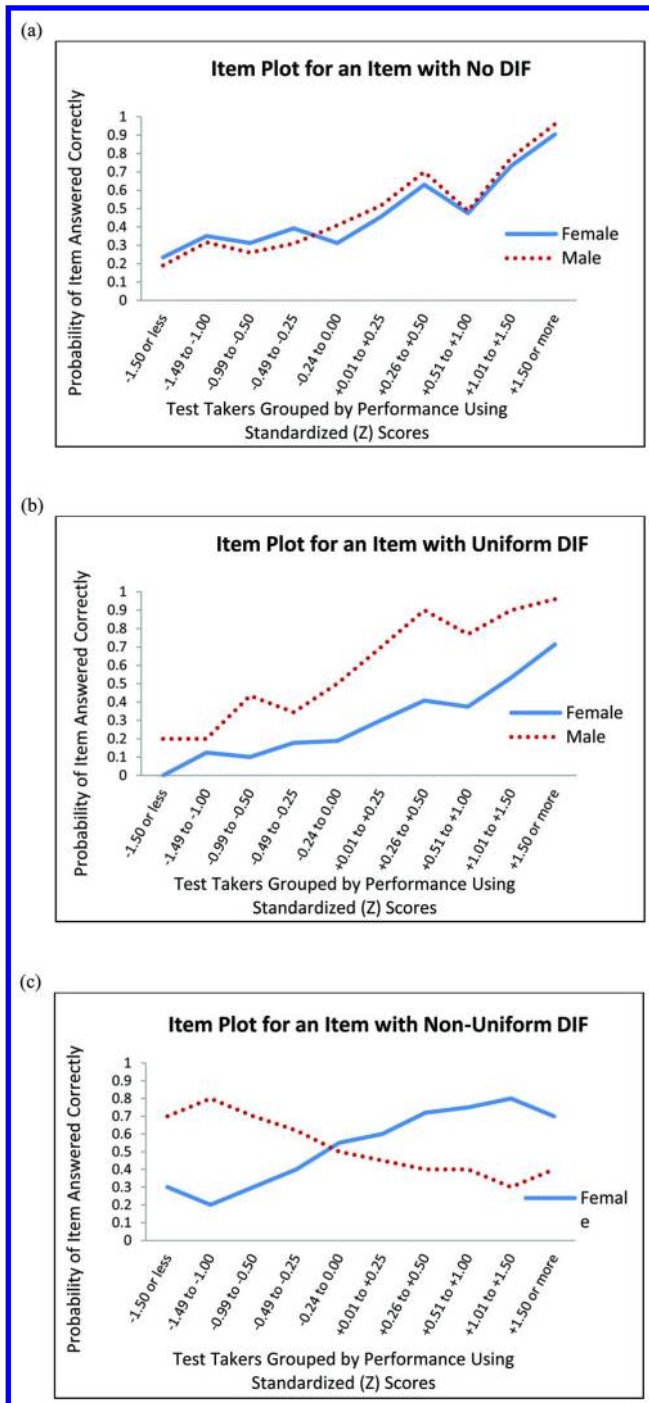


Figure 1. Item plots exhibiting (a) item with no DIF (b) item with uniform DIF and (c) item with non-uniform DIF.

**Table 1. 2 × 2 Contingency Table**

	<i>Performance on item</i>		<i>Performance on item i</i>
Tested Group	1	0	
Reference (r)	$a_i$	$b_i$	$N_{ri} = a_i + b_i$
Focal (f)	$c_i$	$d_i$	$N_{fi} = c_i + d_i$
	$N_{1i} = a_i + c_i$	$N_{0i} = b_i + d_i$	$N_i = a_i + b_i + c_i + d_i$

When using IRT for DIF analysis, one of the three models is used to create an item characteristic curve (ICC) for each subgroup. Each ICC is placed on the same scale by assuming the ability levels as determined between the subgroups are equivalent. The ability or proficiency level of students is often determined by the score on the test under analysis (or an internal measure of proficiency). Any differences in the parameters between the two different groups can then be analyzed to determine if DIF is present (31). Some of the limitations of using IRT methods to determine DIF are that a large sample set is needed and the researcher needs to understand the theory behind the method and parameter estimation in detail (23). While there is not a definite cut off for sample size, it has been suggested that for dichotomous data around 200 participants are need for a two parameter model, though others suggest at least 500 participants (32).

### Logistic Regression

Lastly, another procedure that is commonly used to detect DIF is logistic regression. Logistic regression is another mathematical model that works by predicting if the students will get the question right based on an observed variable (usually the score they receive on the assessment) (25). The logistic regression model is shown in equation 8, where  $\theta_s$  is the observed ability of the subject (s) (21, 25).

$$Y = b_0 + b_1 (\theta_s) + b_2 (\text{gender}) + b_3 (\theta_s \times \text{gender}) \quad \text{Equation 8}$$

In the above formula  $b_0$  is the intercept,  $b_1$  is the effect to which the score on the assessment has,  $b_2$  is the effect to which subgroup membership has, and  $b_3$  is the interactive effect between the score on the assessment and the subgroup membership. If  $b_2$  and  $b_3$  are equal to zero then that would represent no DIF because essentially there is no difference between the genders. Uniform DIF is detected when  $b_2$  is added to the prediction and non-uniform DIF is detected when  $b_3$  is added to the prediction. If  $b_2$  is not equal zero but  $b_3$  is equal to zero than there is uniform DIF. This is saying that for both genders the slope on the item plot is approximately the same but the intercept is different. Lastly if  $b_3$  is not equal to zero than that represents non-uniform DIF. This means that the slopes for both genders on the item plot are different and are independent of the results of  $b_2$  (25). Logistic Regression has the added advantage over other methods as it can more easily predict items with nonuniform DIF as well as items with uniform DIF (33).

## Comparison of Statistical Method to Determine DIF

As each method has its advantages and limitations, many researchers will use multiple methods to detect DIF. Sometimes they use them as a comparison and other times it's complementary to make sure all DIF items are detected. Some examples include using the Mantel-Haenszel procedure and IRT (17), the Mantel-Haenszel procedure and SIBTEST (26), and the Mantel-Haenszel procedure and Logistic Regression (25). As the authors of this chapter have studied identification of items exhibiting persistent differential item functioning and determining possible causes of DIF in these items, they wanted to do a comparison of common methods used for detection of DIF on one of the more common assessments used in their field. The following example is a study comparing three different methods to detect items that exhibit DIF: the Mantel-Haenszel procedure, Logistic Regression, and IRT.

### Methods

This research was conducted at a large, public, doctoral university in the Midwest. The 900 participants in the study were general chemistry I students from three semesters (the fall of 2009, and the spring and fall of 2010) and separated by gender,  $n(\text{male students}) = 401$ ;  $n(\text{female students}) = 499$ . There were two different instructors with all other components of the courses the same. The gender assignments were self-reported and if someone did not report their gender their data was not used for the study. DIF analyses using the Mantel-Haenszel procedure, Logistic Regression, and IRT (2 parameter model) were performed on the ACS DivCHED EI First Term General Chemistry Paired Questions 2005 final examination that contained 40 multiple choice questions. All students in the three semesters of testing took the exam under the same conditions (as a final exam and with the same testing requirements) and for the same stakes (same contribution towards their final grade). Additionally, all students in the three semesters were prepared the same way, using the same teaching methodology and covering the same curriculum. Each of these items was analyzed by content area and format by the authors and another expert in the field. Once both the content area and format were assigned for each item, the three met to discuss their assignments for each item until a unanimous assignment was reached. The content areas included many different areas that are taught in a college general chemistry I course. There were 75 content areas which ranged from classification of matter to phase diagrams. The formats of the items were classified as visual-spatial (VS), specific chemical knowledge (SCK), reasoning (R), and computation (C). These format types were adapted from another DIF study that included factor analysis of item formats (19). The items could be classified with multiple formats depending on how it was constructed.

## Results

Of the 40 items on the assessment there were 13 items in total that exhibited DIF: 12 items that were identified based on the Mantel-Haenszel procedure, 10 items were identified using Logistic Regression, and four items were identified based on IRT as shown in Figure 2. The breakdown of these items by uniform or non-uniform DIF is given in Table 2. Examples of uniform and non-uniform DIF item plots are shown in Figures 1b and 1c. Of the 12 items that were identified using the Mantel-Haenszel procedure, 11 of those items exhibited uniform DIF and the other item was identified as having nonuniform DIF. For the 10 items that were identified as exhibiting DIF using Logistic Regression, eight of those exhibited uniform DIF and the other two were identified as exhibiting nonuniform DIF. All of the items that were identified as exhibiting DIF using IRT were identified as uniform DIF. For the four items that were identified using IRT, these were also identified by both Logistic Regression and the Mantel-Haenszel procedure.

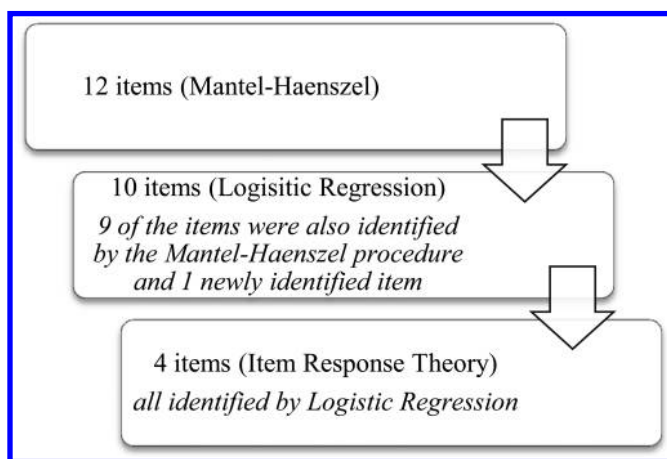


Figure 2. Number of DIF items identified by three different DIF detection methods.

**Table 2. Uniform and Nonuniform DIF Items by DIF Detection Methods**

<b>Type of DIF</b>	<b>Number of items by method for DIF detection</b>		
	<i>Mantel-Haenszel</i>	<i>Logistic Regression</i>	<i>Item Response Theory</i>
Uniform	11	8	4
Non-Uniform	1	2 <sup>a</sup>	0

<sup>a</sup> 1 of the non-uniform items identified by Logistic Regression was also identified by the Mantel-Haenszel procedure.

The items were then considered by the direction of favor and classification of the items that exhibited uniform DIF. Focusing first on the four items that exhibited DIF detected by all methods, all were found to favor male students. The first item was classified with a content area of general stoichiometry and formats of VS, R, and C. The second item was classified with a content area of limiting reagents and formats of VS, and R. The third item was classified with a content area of general properties of aqueous solutions and formats of VS, and SCK. The last item was classified with a content area of kinetic molecular theory and a format of R. This is shown in Table 3.

Besides the four items that were identified by all three methods there were five other items that were identified by both the Mantel-Haenszel procedure and Logistic Regression. Four of these items exhibited uniform DIF and one item exhibited non-uniform DIF. The items with non-uniform DIF have no direction of favor, because by definition the favor flips depending on the ability levels of the participants. For the items that exhibited uniform DIF one of them favored female students and the other three favored male students. The item that exhibited non-uniform DIF was classified with a content area of chemical reactions and formats of VS, and R. Of the four items the exhibited uniform DIF the first item was classified with a content area of ionic radius, a format of SCK and was the item that favored female students. The second item was classified with a content area of classification of matter and a format of SCK. The third item was classified with a content area of limiting reagents and formats of VS, and R. The last item was classified with a content area of kinetic molecular theory and formats of SCK, and R. This is shown in Table 4.

**Table 3. Uniform DIF Items by Direction of Favor and Classification Detected by Item Response Theory, Logistic Regression and Mantel-Haenszel Methods**

	<i>Favor</i>	<i>Content Area</i>	<i>Format<sup>b</sup></i>
Item 1 <sup>a</sup>	male	stoichiometry	VS, R, C
Item 2	male	limiting reagent	VS, R
Item 3	male	aqueous solutions	VS, SCK
Item 4	male	kinetic molecular theory	R

<sup>a</sup> Item numbers are arbitrarily assigned and do not correspond to test item numbers <sup>b</sup> C = computation; R = reasoning; SCK = specific chemical knowledge; VS = visual-spatial or reference component.

There was one item that was only identified as exhibiting DIF by Logistic Regression and not by the Mantel-Haenszel procedure or IRT. This item exhibited non-uniform DIF and was classified with a content area of electronegativity and formats of SCK, and R. Because this item exhibited non-uniform DIF there is no overall consistent favor associated with this item.

**Table 4. DIF Items by Direction of Favor and Classification Detected by Logistic Regression and Mantel-Haenszel Methods Only**

	<i>Type</i>	<i>Favor</i>	<i>Content Area</i>	<i>Format<sup>b</sup></i>
Item 5 <sup>a</sup>	non-uniform	-- <sup>c</sup>	chemical reactions	VS, R
Item 6	uniform	female	ionic radius	SCK
Item 7	uniform	male	classification of matter	SCK
Item 8	uniform	male	limiting reagents	VS, R
Item 9	uniform	male	kinetic molecular theory	SCK, R

<sup>a</sup> Item numbers are arbitrarily assigned and do not correspond to test item numbers <sup>b</sup> C = computation; R = reasoning; SCK = specific chemical knowledge; VS = visual-spatial or reference component <sup>c</sup> - no favor or direction as this was a non-uniform item.

There were three additional items that were only identified as exhibiting DIF using the Mantel-Haenszel procedure. All three of these items exhibited uniform DIF with the first two favoring female students and the last one favoring male students. The first item was classified with a content area of general stoichiometry and formats of VS, and C. The second item was classified with a content area of molecular shape and a format of R. The last item was classified with a content area of kinetic molecular theory and a format of SCK. This is shown in Table 5.

**Table 5. DIF Items by Direction of Favor and Classification Detected by Mantel-Haenszel Methods Only**

	<i>Favor</i>	<i>Content Area</i>	<i>Format<sup>b</sup></i>
Item 10 <sup>a</sup>	female	stoichiometry	VR, C
Item 11	female	molecular shape	R
Item 12	male	kinetic molecular theory	SCK

<sup>a</sup> Item numbers are arbitrarily assigned and do not correspond to test item numbers <sup>b</sup> C = computation; R = reasoning; SCK = specific chemical knowledge; VS = visual-spatial or reference component.

Considering the items collectively that exhibited uniform DIF, eight items favored male students and three items favored female students. These 11 items were classified into eight different content areas with only one content area overlapping into two items with each favoring a different gender. For the items that exhibited DIF that favored female students there were three different content areas. For the items that exhibited DIF that favored male students, two of the content areas had multiple items: there were two items with the content area of limiting reagents and three items with a content area of kinetic molecular theory. The format of visual-spatial or reference component was more common for items favoring male students with four items containing this format as opposed to only one item that favored female students. No other discernable trend based on format

was identified. For the two items that exhibited non-uniform DIF they were in two different content areas (chemical reactions and electronegativity) and had three different formats (VS, R, SCK), with both items having the format of R. This is summarized in Table 6.

**Table 6. Uniform and Nonuniform DIF Items by Direction of Favor and Classification**

<i>Favor</i>	<i>Number of items</i>	<i>Content Area<sup>a</sup></i>	<i>Format<sup>b</sup></i>
female	1	stoichiometry	VR, C
	1	ionic radius	SCK
	1	molecular shape	R
male	1	classification of matter	SCK
	1	stoichiometry	VS, R, C
	2	limiting reagent	VS, R
	1	aqueous solutions	VS, SCK
	3	kinetic molecular theory	R or SCK
-- <sup>c</sup>	1	chemical reactions	VS, R
	1	electronegativity	SCK, R

<sup>a</sup> only the content area of stoichiometry overlapped between the direction of favor <sup>b</sup> C = computation; R = reasoning; SCK = specific chemical knowledge; VS = visual-spatial or reference component <sup>c</sup> no favor or direction as these were non-uniform items.

DIF detection and analysis can be an important validity check to use when making decisions associated with test performance. While the study presented here was illustrative of the different methods available for detecting DIF, it is also important to note that the sample used was actual student data and not theoretical data. This warrants an important consideration or limitation in interpreting DIF results. When using DIF results to make judgments about the validity of data produced from the assessment items, the statistical analysis only provides statistical information. Considering this another way, there must always be a consideration of the statistical error of the DIF detection and the possibility that the results are due to chance rather than an actual performance differential. Therefore one must be careful and proceed with caution when making statements about the items that exhibited DIF. In previous studies, however, it has been shown that when comparing the Mantel-Haenszel procedure and IRT, both under-detected the amount of DIF items (17). With larger sample sizes (1,000 participants total) another study found that for uniform DIF both the Mantel-Haenszel procedure and Logistic Regression were able to detect 100% of the items that exhibited DIF (25). Given the low probability of over-detecting the amount of DIF items, it seems reasonable that some conclusions can be drawn from the above data presented.

First the data suggests that the Mantel-Haenszel procedure was the most sensitive at being able to detect items that exhibited uniform DIF. There were 11 items that exhibited uniform DIF detected by the Mantel-Haenszel procedure, whereas eight of those eleven items were detected by Logistic Regression and four of those eight by IRT. Coupling these results with the fact that the Mantel-Haenszel procedure is fairly easy to use and readily available in many software packages this makes it a promising method to use for detecting uniform DIF items.

For items that exhibited nonuniform DIF the data suggests that Logistic Regression was the most powerful detecting two items, and IRT was the least powerful not detecting any items. However, contrary to previous studies the Mantel-Haenszel procedure did detect one item that exhibited nonuniform DIF (17, 25).

When considering the content and the format of the items that exhibited possible DIF, a number of conclusions can be drawn. Out of the 11 items that exhibited uniform DIF, eight items exhibited DIF favoring male students and three items exhibited uniform DIF favoring female students. There was only one content area for the items the exhibited uniform DIF that favored both genders and that was general stoichiometry. All other items had different content areas suggesting that one of the possible causes of DIF could be content area (16, 20). There were also two different content areas that had multiple items that exhibited uniform DIF. The content area of limiting reagents had two items and the area of kinetic molecular theory had three items that exhibited uniform DIF all favoring the same subgroup that reinforced this conclusion.

When considering the items that exhibited nonuniform DIF as well as the format of the items, the data are inconclusive. With only two items that exhibited nonuniform DIF, there are not enough data to make conclusive statements. No pattern emerged when considering the format of the items either by uniform versus nonuniform DIF or by gender. For the items that exhibited DIF favoring female students one of each format was used overall on the three items. For the items that exhibited DIF favoring male students the format of computation was used once, the formats of visual-spatial and specific chemical knowledge was used four different times each, and the format of reasoning was used five times. For the items that exhibited nonuniform DIF the formats of visual-spatial and specific chemical knowledge was used once, and the format of reasoning was used twice overall in the two items.

## Conclusion

The results of this study suggest that for this sample size the Mantel-Haenszel procedure was the most sensitive in detecting uniform DIF and Logistic Regression was the most sensitive in detecting nonuniform DIF. However, it may be helpful to do a more in-depth analysis. Using a two-stage iterative process to determine DIF could possibly lead to a more definite knowledge of which items were “real” DIF and not either over- or under-detected by the different methods (34). One must always consider the sample size as well. For this study, it is possible that



a larger data set would increase the sensitivity of the IRT analysis because larger sample sets can improve the accuracy of estimation if the 2- or 3-parameter model is used (23). Overall the results from this study suggest that if wanting to detect both uniform and non-uniform DIF, multiple methods should be used.

The use of assessments in many settings will continue as will using the outcome on the assessments to make judgments and decisions. Because these judgments and decisions can have high-stakes consequences for the participants, it is the responsibility of the practitioner who constructs and administers the assessment that the data from the test and related decisions are fair and produce valid results. Concurrent with making sure the data of the assessments produce valid and fair results, it is also a requirement that those results are fair for different subgroups of testtakers. The use of DIF analysis can provide practitioners with information to make informed decisions on the use of assessment outcomes. Only through the critical use of analyses of assessments can the related judgments and decisions work towards fairer tests that could lead to greater equality among participants.

## Acknowledgments

We would like to thank Anja Blecking for her help serving as an expert rater in categorizing the items by both content area and format. We would also like to thank the many students who participated in this study. This material is based upon work supported in part by the University of Wisconsin-Milwaukee (Research Growth Initiative Grant under Grant No. 101X259).

## References

1. Crawford, M.; Marecek, J. *Psychol. Women Q.* **1989**, *13*, 477–491.
2. Weisstein, N. *Kinder, Kirche, Kuche as Scientific Law: Psychology Constructs the Female*; New England Free Press: 1968.
3. Subrahmanian, R. *Int. J. Educ. Develop.* **2005**, *25*, 395–407.
4. *The Millennium Development Goals Report*; United Nations: New York, 2013; pp 1–68.
5. *UNESCO Priority Gender Equality Action Plan 2014-2021*; United Nations Educational, S. a. C. O., Ed. Paris, France, 2014; pp 1–67.
6. Cole, N. S. *The ETS Gender Study: How Females and Males Perform in Educational Settings*; Educational Testing Service: Princeton, NJ, 1997; pp 1–36.
7. Maccoby, E. E.; Jacklin, C. N. *The Psychology of Sex Differences*; Stanford University Press: Stanford, CA, 1974.
8. Cleary, T. A. Gender Differences in Aptitude and Achievement Test Scores. In *Sex Equity in Educational Opportunity, Achievement, and Testing*; Pfeleiderer, J., Ed. Educational Testing Service: Princeton, NJ, 1992; pp 51–90.
9. Sjøberg, S. *Scand. J. Educ. Res.* **1988**, *32* (1), 50–60.

10. Linn, M. C.; Petersen, A. C. A Meta-analysis of Gender Differences in Spatial Ability: Implications for Mathematics and Science Achievement. In *The Psychology of Gender: Advances through Meta-analysis*; Hyde, J. S., Linn, M. C., Eds.; The Johns Hopkins University Press: Baltimore, MD, 1986; pp 67–101.
11. Friedler, Y.; Tamir, P. *Res. Sci. Technol. Educ.* **1990**, 8 (1), 21–34.
12. Linn, M. C. In *Gender Differences in Educational Achievement*; ETS Invitational Conference, New York, NY; Pfliegerer, J., Ed.; Educational Testing Services: New York, 1991; pp 11–50.
13. Halpern, D. F.; LaMay, M. L. *Educ. Psychol. Rev.* **2000**, 12 (2), 229–246.
14. Dorans, N. J.; Holland, P. W. DIF Detection and Description: Mantel-Haenszel and Standardization. In *Differential Item Functioning*; Holland, P. W., Wainer, H., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1993; pp 35–66.
15. Holme, T. *J. Chem. Educ.* **2003**, 80 (6), 594.
16. Kendhammer, L.; Holme, T.; Murphy, K. *J. Chem. Educ.* **2013**, 90 (7).
17. Hambleton, R. K.; Rogers, H. J. *Appl. Meas. Educ.* **1989**, 2, 313–334.
18. Schmitt, A. P.; Dorans, N. J. *J. Educ. Meas.* **1990**, 27 (1), 67–81.
19. Hamilton, L. S.; Snow, R. E. *Exploring Differential Item Functioning on Science Achievement Tests*; 483; National Center for Research on Evaluation, Standards, and Student Testing: Los Angeles, CA, 1998; pp 1–43.
20. Zenisky, A. L.; Hambleton, R. K.; Robin, F. *DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices*; University of Massachusetts Amherst: Amherst, MA, 2003; pp 1–22.
21. Zumbo, B. D. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*; Directorate of Human Resources Research and Evaluation, Department of National Defense: Ottawa, ON, 1999.
22. Crocker, L.; Algina, J. *Introduction to Classical & Modern Test Theory*; Holt, Rinehart and Winston: New York, 1986.
23. Clauser, B. E.; Mazor, K. M. *Educ. Meas: Issues Pract.* **1998**, 31–44.
24. Holland, P. W.; Thayer, D. T. Differential Item Functioning and the Mantel-Haenszel Procedure. In *Test Validity*; Wainer, H., Braun, H. I., Eds.; Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ, 1988; pp 129–145.
25. Swaminathan, H.; Rogers, H. J. *J. Educ. Meas.* **1990**, 27 (4), 361–370.
26. Shealy, R.; Stout, W. *Psychometrika* **1993**, 58 (2), 159–194.
27. Mazor, K. M.; Kanjee, A.; Clauser, B. E. *J. Educ. Meas.* **1995**, 32 (2), 131–144.
28. Embretson, S. E.; Reise, S. P. *Item Response Theory for Psychologists*; Lawrence Erlbaum Associates, Inc.: Mahwah, New Jersey, 2000.
29. van der Linden, W. J.; Hambleton, R. K. Item Response Theory: Brief History, Common Models, and Extensions. In *Handbook of Modern Item Response Theory*; van der Linden, W. J., Hambleton, R. K., Eds.; Springer: New York, 1997; pp 1–31.
30. Hambleton, R. K.; Swaminathan, H.; Rogers, H. J. *Fundamentals of Item Response Theory*; Sage Publications: Newbury Park, CA, 1991; p 174.

31. Thissen, D.; Steinberg, L.; Wainer, H. Detection of Differential Item Functioning Using the Parameters of Item Response Models. In *Differential Item Functioning*; Holland, P. W., Wainer, H., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1993.
32. Thorpe, G. L.; Favia, A. *Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications Psychology Faculty Scholarship [Online]*, 2012; Paper 20.
33. Karami, H. *Int. J. Educ. Psychol. Assess.* **2012**, *11* (2), 59–76.
34. Zenisky, A. L.; Hambleton, R. K. *Educ. Psychol. Meas.* **2003**, *63* (1), 51–64.

## Chapter 5

# Organic Chemistry Practice Exam: Helping Students Gain Metacognitive Skills To Excel on the Full-Year ACS Exam

Susan M. Schelble,<sup>\*,1</sup> Milton J. Wieder,<sup>1</sup> David L. Dillon,<sup>2</sup>  
and Ethan Tsai<sup>2</sup>

<sup>1</sup>Department of Chemistry, Metropolitan State University of Denver,  
P.O. Box 173362, CB 52, Denver, Colorado 80217

<sup>2</sup>Department of Chemistry, Colorado State University-Pueblo,  
2200 Bonforte Boulevard, Pueblo, Colorado 81001

\*E-mail: [sschelbl@msudenver.edu](mailto:sschelbl@msudenver.edu).

Students who must take a high-stakes final examination that covers a full year of the organic chemistry curriculum often wonder how best to prepare for such an assessment. Many institutions use the products from the American Chemical Society (ACS) Examinations Institute as one metric for student learning. In order to help students prepare for this kind of final examination, a committee of authors from several institutions created a 50-question Organic Practice Exam. About 1700 students from dozens of institutions used this examination prior to taking the ACS organic chemistry final examination. This chapter describes the outcomes and impact the Organic Practice Exam had on student ACS organic chemistry final examination performance. It will also describe student metacognitive information from practice examination questions and a comparison to expert item analysis.

## Introduction

In 2010, a committee of organic chemists worked = to create an Organic Practice Exam for a full-year course. The construction of the exam paralleled that of the General Chemistry ACS Practice Exam (1) and of an Organic ACS Full-Year Exam currently being used by the Exams Institute (2). The goals of providing organic students with a practice exam were similar to those cited for the General Chemistry ACS Practice Exam (1). Like all students preparing for a high stakes ACS final exam, organic students are actively seeking study materials, such as the Study Guide (3), this practice exam, and other materials they may find from a variety of sources. The latter often have multiple-choice questions, but frequently have not undergone the rigors of constructing questions by a committee and evaluating those individual items and the entire exam for reliability, validity, discrimination, difficulty, and cognitive load. The Organic Practice Exam for a full-year course was written to address all of the aforementioned factors.

## Construction of the Practice Exam

The process for writing the Organic Practice Exam began by forming a committee of item writers. This is similar to the process that the Exams Institute uses for constructing an ACS Exam. Our work for the practice document was somewhat smaller in scope, however. We used nine authors from five different institutions of various sizes and missions in Colorado. This differs from the process used by an official exam preparation, where typical committees have 15-20 members from across the United States.

Like an official exam committee, this group set out to prepare exam items that represent the scope of a full-year course. They used the same general topics as those used to prepare the OR04 Organic Chemistry two-semester exam (2).

The Organic Practice Exam committee used the list of topics to construct 120 multiple-choice questions. The committee looked for a variety of topic coverage and difficulty, and pared the number of questions to 50 (from the original 120). Questions were also constructed to have a variety of projected measures of levels of cognition. The committee sought to include both lower and higher order cognitive skill metrics (4, 5), a percentage of *recall*, *algorithmic* and *conceptual* problems that mirrors current ACS secure exams (6), and sustainability literacy (7).

This exam construction approach is similar to that of an ACS secure exam. Secure exams are developed by committee from two versions (A and B) of an exam, where each version has 50-70 questions that are selected from a total pool of 200 or more items. Both A and B versions have similar topics represented. These versions are subsequently beta-tested and the final questions are chosen based on the statistical data derived from the beta-tests.

The practice exam committee only developed one 50-item version of the exam. The 50 questions were divided into 10 *content* categories as an abridged grouping of topics from which the committee designed the exam, originally. These are summarized in Table 1.

**Table 1. Groupings of 50 Exam Items into One of Ten Topics**

<i>Content Group</i>	<i>Content Areas</i>	<i>Item #</i>
1	Nomenclature Stereochemistry	1, 2, 8, 9, 10,
2	S <sub>N</sub> 1/S <sub>N</sub> 2 E1/E2	6, 14, 17, 18, 19, 23, 36,
3	Addition to alkenes alkynes	7, 21, 22, 25, 37, 40,
4	Stability Acidity Mechanisms	3, 4,5, 11, 12, 13, 14, 15, 24,
5	Aromatic Substitution Aromaticity	27, 28, 50
6	Reduction Oxidation	26, 29, 31,
7	Spectroscopy	20, 33, 38, 45, 49
8	Carbonyl additions and substitutions Enolates	32, 34, 35, 41, 42, 43
9	Synthesis	30, 39, 48,
10	Radicals Pericyclic	44, 46, 47,

Like the secure exam process, the practice exam included items that would be tested with smaller pools of student volunteers before being selected for the final version of the exam. Although this was done on a smaller scale than the process for a secure exam, the questions that were selected for the final version of the practice exam did undergo testing for difficulty of item, discrimination indices, and topic coverage.

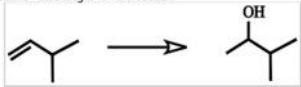
Like the General Chemistry Practice Exam (1), but unlike the typical ACS secure exams, each item for the Organic Practice Exam had a mental effort assessment. The same theories (8) used by the General Chemistry Practice Exam about self-assessment abilities of students were used with the Organic Practice Exam. The latter exam, however, was designed to be unsecured. After taking the exam students could check their answers, and use additional information including the key to help organize their study plans for the final. They were encouraged to examine areas where they excelled and/or needed work using the topics in Table 1.

When taking the exam, each content question was followed by a mental effort rating question. The content questions were answered on a bubble sheet from numbers 1 to 50. The mental effort questions that followed each content

item ranged from numbers **51** to **100**. The students submitted answers on a form for computer grading with 100 items on one sheet. Thus, in column 1 on the scantron, they could answer the content question, and in column 2 the mental effort question. This system also allowed for computing the content and mental effort data separately, or as unit pairs, and ensured that the mental effort was evaluated immediately after each content question. An example is exam question number **25**. This question is assigned to *content* category 3 (Table 1): “Additions to alkenes and alkynes.” After completing the question, each student rated their *mental effort* required to answer the question. This was done using a Likert scale, where **1** corresponded to *low* mental effort and **5** to *high* mental effort. The students were directed to rate their perceived *effort*, not their perceived *difficulty* of the question. It is conceivable, therefore, that on occasion a difficult question might have a *low mental effort* because the student would choose to guess. On the other hand, a modestly difficult question might require a great deal of *mental effort*, which the student chooses to use because they feel it will lead to a correct answer.

An example of how a paired set of questions appeared to students is illustrated in Figure 1. Question number **75** is the *mental effort* rating for content question number **25**.

**25.** Which reagents could be used to make this alcohol as the only possible alcohol product from the reaction with 3-methyl-1-butene?



(A) 1. Br<sub>2</sub>            2. H<sub>2</sub>O  
 (B) 1. BH<sub>3</sub>            2. NaOH, H<sub>2</sub>O<sub>2</sub>  
 (C) 1. H<sub>2</sub>SO<sub>4</sub>        2. H<sub>2</sub>O  
 (D) 1. Hg(OAc)<sub>2</sub>      2. Na BH<sub>4</sub>

**75.** How much mental effort did you expend on question # 25?

(A) very little  
 (B) little  
 (C) moderate amounts  
 (D) large amounts  
 (E) very large amounts

Figure 1. Example of test item followed by respective mental effort rating (1).

The initial version of the Organic Practice Exam was administered in the spring of 2010 to 35 students. Several questions were removed and replaced because they were too difficult and/or did not discriminate between well-prepared students and those who were guessing.

An example of an item that was removed is shown in Figure 2.

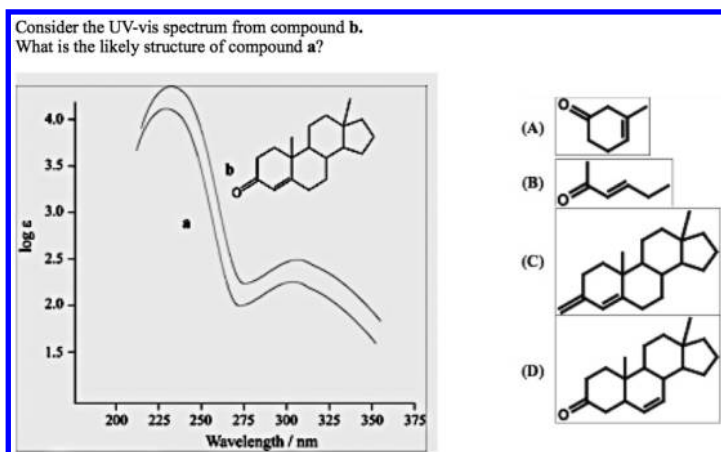


Figure 2. This question was answered correctly by less than 10% of students. Statistics also indicated that students performing well on other parts of the exam performed poorly on this question, so it was a poor discriminator.

The final version of the organic practice exam, titled *May 2011*, containing 50 items, was reset and used starting in 2011 and concluding in 2014.

## How the Practice Exam Was Used

The final version of the Organic Practice Exam as used by a dozen different undergraduate institutions, and several thousand students. Instructors used the exam with numerous approaches, but always with a goal of helping their organic students achieve a deeper mastery of the subject. The exam was written specifically to be unsecured.

One typical approach was to administer the Organic Practice Exam about 3-4 weeks before the completion of the second semester organic chemistry course. Students recorded their answers on scantrons and on paper. Shortly after completing the exam, students were given a key and a copy of the questions. Students were also given a version of the organic content topics from Table 1 and asked to do a self-evaluation about their strengths and weaknesses regarding the exam material. This is illustrated in Figure 3, where student computed his/her relative performance in the 10 content areas, and ranked the order for study planning.



	Content Areas	Item #	% Correct	Order by study need: 1 = lowest% correct; 10 = highest % correct
1	Nomenclature Stereochemistry	1, 2, 8, 9, 10,	80	8
2	S <sub>N</sub> 1/S <sub>N</sub> 2 E1/E2	6, 14, 17, 18, 19, 23, 36,	4	2
3	Addition to alkenes alkynes	7, 21, 22, 25, 37, 40,	33	3
4	Stability Acidity Mechanisms	3, 4, 5, 11, 12, 13, 14, 15, 24,	60	6
5	Aromatic Substitution Aromaticity	27, 28, 50	46	7
6	Reduction Oxidation	26, 29, 31,	100	9
7	Spectroscopy	20, 33, 38, 45, 49	40	4
8	Carbonyl additions and substitutions Enolates	30, 32, 34, 35, 41, 42, 43	57	5
9	Synthesis	39, 48,	0	1
10	Radicals Pericyclic	44, 46, 47,	46	7

Figure 3. Student completed self-assessment, from which he/she could design study plan.

Students completed the self-assessment of skills and were encouraged to incorporate this information into the plans they would develop for studying for the final. Many used this information to determine the sections of the ACS Organic Study Guide (3) that needed the most effort.

Whereas students had questions and answers after taking the practice exam, they had a new opportunity to work with faculty and peer-leaders both in and outside of the classroom. Faculty and student leaders tried to use the practice exam to help students identify the basic skills that need to be mastered to answer each question on the exam correctly. They also worked with students to help them learn how to integrate the basic items that will lead to an informed solution to the question.

Variations on exam usage were reported. Some faculty administered the exam in parts and distributed the questions and corresponding answers throughout the several weeks before the scheduled final exam. This allowed faculty to help students review material in smaller portions while still being able to give feedback based on the questions and the subtask skills needed for each.

One instructor used the practice exam in a large classroom with clickers and collected both individual and group data. These data are not included in this current report.

All users of the Organic Practice Exam encouraged students to self assess their basic skills and increase their ability to solve problems that had higher order cognitive loads by learning how to integrate basic skills.

Whereas approximately 1700 students have used this exam, some smaller subsets of students provided information about exam items, mental effort, and correspondence of performance on ACS exam used for a course final. The subsequent evaluation was provided by students who signed the corresponding consent form from the IRB of this study.

## How the Practice Exam Was Evaluated

The Organic Practice Exam was evaluated in several steps during its construction. After the initial beta-testing, the final 50-question exam was evaluated for difficulty, discrimination, overall test reliability, cognitive complexity, student mental effort, match to ACS Anchoring Concept Content Maps (9, 10), impact on final exam performance, types of artifacts used on the practice exam (6), and comparison of these various parameters. Exam answers on DataLink-Scantrons were evaluated using the Apperson GradeMaster 600 Test Scanner (11). Software that is included with this grading program was used to determine difficulty, discrimination, and overall test reliability.

### Difficulty

Each item was measured for difficulty in the manner typically used by ACS secure exams. Difficulty is directly related to the number of correct responses, or indirectly to the number of incorrect responses. The item on the exam with the highest difficulty had 90% correct (10% incorrect); the item with the lowest difficulty had 20.7% correct (79.3% incorrect). The ideal question range was between 20% incorrect and 80% incorrect. Questions outside this range often were poor discriminators. We tried to avoid having too many questions that were either too easy or too difficult.

### Discrimination

This parameter was measured with a slightly different software application than is typically used for ACS secure exams. The item discrimination was calculated using the point biserial rating (12). This is a correlation statistic that estimates the degree of relationship between two dichotomous scales, and is abbreviated  $r_{pbi}$ . This rating ranges from  $-1.00$  to  $+1.00$ . Any positive rating would indicate a positive correlation. For our exam questions, this is desirable. An exam question with a high positive  $r_{pbi}$  indicates that a student who was performing at a high level on the entire exam tended to get this question correct; a student who was performing at a low level on the entire exam tended to get this question incorrect.

A negative  $r_{pbi}$  is undesirable because it indicates that weak students are getting the question correct more frequently than strong performers on the rest of the exam. The question in Figure 4 that was discarded had a  $-0.20$   $r_{pbi}$ . Technically, the  $r_{pbi}$  is calculated based on work from Linacre (13). We set the lower limit for an acceptable question at  $+0.15$ . This is similar to the discrimination limit used by ACS exams. We did not set an upper limit, as high discrimination is very desirable.

The biserial correlation coefficients for the Organic Practice Exam ranged from 0.15 (weakest correlation) to 0.58 (strongest correlation).

## Test Reliability

The overall *reliability* of the exam was also calculated with Apperson software. The KR20 calculation is for the Kuder Richardson (14) Coefficient of Reliability for Binary Data. This statistical measurement is used to examine the reliability of exam items to determine if items within the entire exam obtain the same results over a population of testing subjects.

The KR20 for the Organic Practice Exam was 0.849. A KR20 of 0.9 or more indicates a homogenous set of data (15), so this is quite a good correlation of exam reliability. For 50 test items this is considered to be a strong estimate of the *reliability* of this multiple-choice exam (16).

## Cognitive Complexity

This is a numerical value that is determined by “experts” in the chemical education community and, in this case, the organic education community. *Cognitive complexity* assignments for organic chemistry exam questions (17) differ slightly from *Cognitive complexity* assignments for general chemistry items (18).

The same cognitive complexity instrument was applied here as in the work published by Raker, Trate, Holme, and Murphy (17). For this practice exam, each item was analyzed with the same organic rubric by 5 different experts. This included identification of the following for each exam question:

1. Number of **subtasks** and rating each as *easy*, *medium*, or *hard*.
2. An **amplification** factor (integrating subtasks) rating as *easy*, *medium*, or *hard*.
3. The role of **distractors** was rated as *selection*, *elimination*, or *evaluation*.

For this *complexity rating*, each expert tried to predict the thought process that a typical student might undertake while trying to answer a question. Subtasks would include all student processes such as definitions, recall of mechanisms, stereochemical rules, stabilities, three-dimensional structure, reaction outcomes, rearrangements, solvent effects, etc. The more subtasks and respective difficulties, ranked as *easy* (**E**), *medium* (**M**), or *hard* (**H**), the higher the projected cognitive load the student will experience when answering a question.

The *amplification* score results from predicting the extent of activity a student must apply in order to integrate all of the subtasks, these are ranked as *easy* (**E**), *medium* (**M**), or *hard* (**H**).

The *distractor role* is a measure of the process a student needs to consider in order to arrive at the correct answer. If the student can determine the correct answer before looking at the responses, this *role* would be classified as *selection*. This would be the lowest *cognitive* rating. If the correct answer can be ascertained by *eliminating* one or two choices, this would be rated as a medium *cognitive load*. Finally, if each possible answer must be *evaluated* before answering a question,

this would be rated as the *highest cognitive load*. An example of a question needing *evaluation* would be one that required the ranking of 3 or 4 ions (radicals) for relative stability. In this case, each possible answer must be *evaluated* separately.

An example of the rubric applied to question 25 (Figure 1), by one expert rater is shown in Figure 4.

	Subtasks	Difficulty	Amplification	Distractor
25	Reaction			
	Additions of water to alkenes	E M H		
	Reagents that can provide water addition	E M H		
	Regio selection of addition	E M H	E M H SEL	ELIM EVAL
	Predication of rearrangements	E M H		

Figure 4. Each subtask was evaluated; the amplification (in terms of working memory and cognitive load) was determined to be hard; the role of the distractor was selective.

After all experts applied qualitative ratings to each question on the exam, numerical scores were assigned to each. These quantitative evaluations were made using the same instrument previously described (18). For question number 25, one medium **subtask** has a score of 2; three hard subtasks have a score of 6; one hard **amplification** has a score of 3; and a **distractor** of *selection* has a score of 0. This rating adds up to 11.

The experts rated all 50 questions for the Organic Practice Exam. On average these ranged between 3.750 and 11.70 for *cognitive complexity*.

## Student Mental Effort

Students rated each item using a Likert Scale. These items were averaged (241 students completed ratings for each question) and were used to look for correlations with *cognitive complexity* and *difficulty*.

## ACS Anchoring Concept Content Map (ACCM)

Ten major anchoring concepts were developed for all undergraduate chemistry (9, 10). The 50 questions used on the Organic Practice Exam were each assigned to one of ten categories of *ACCM*. These differ from the 10 content categories described in Table 1.

The *ACCM* categorization was done after the practice organic exam had been prepared, and thus shows some gaps in equitable coverage of the *Big Ideas* from the Content Map. Eleven questions (of 50) on the Practice Organic Exam fall into *ACCM* category 5. This involves reactions and is understandable. Only three questions each fall into *ACCM* categories 4 and 10. These probably should be considered when writing a future practice exam. Table 2 summarizes these assignments for the practice organic exam.

**Table 2. Number of Questions Assigned to Each ACCM Category**

<i>Numbers of 10 “Big Ideas from Content Map</i>	<i>Anchoring Concepts from ACS Exams Institute<sup>10</sup></i>	<i>Number of Questions in each category</i>
1	Matter consists of atoms that have internal structures that dictate their chemical and physical behavior.	4
2	Atoms interact via electrostatic forces to form chemical bonds.	4
3	Chemical compounds have geometric structures that influence their chemical and physical behaviors.	4
4	Intermolecular forces, electrostatic forces between molecules, dictate the physical behavior of matter.	3
5	Matter changes, forming products that have new chemical and physical properties.	11
6	Energy is the key currency of chemical reactions in molecular scale systems as well as macroscopic systems.	5
7	Chemical changes have a time scale over which they occur.	7
8	All chemical changes are, in principle, reversible and chemical processes often reach a state of dynamic equilibrium.	4
9	Chemistry constructs meaning interchangeably at the particulate and macroscopic levels.	5
10	Chemistry is generally advanced via empirical observation.	3

### **Types of Artifacts on Practice Exam**

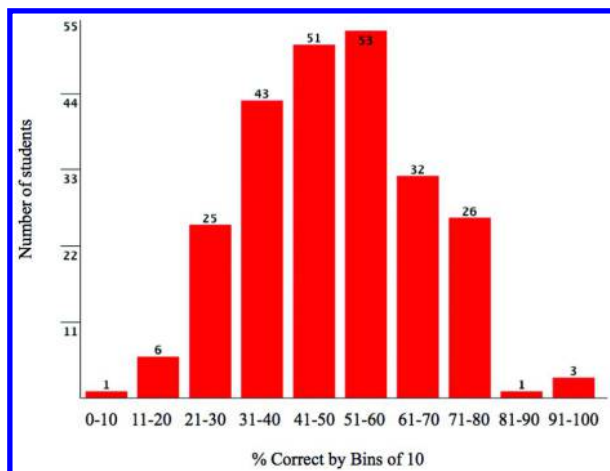
The 50 questions on the Organic Practice Exam were divided into three categories, just as was done with the historical investigation of ACS organic exams from 1949-2012. These areas are: *recall*, *algorithmic*, and *conceptual questions*. Table 3 is a summary of this assessment. The Organic Practice Exam compares (6) very well to the 2004 Full-Year Organic secure exam. All students in this study (269) took the 2004 ACS final exam. Some (162) also took the practice organic exam; others (107) did not.

**Table 3. Number and Percent of Question Types for the 50-Question Practice Organic Exam**

<i>Artifact Type</i>	<i>Number of Questions</i>	<i>Percent of Questions</i>
Recall	3	6%
Algorithmic	7	14%
Conceptual	40	80%

## Results

The results are multifaceted and include information for the instructors about how much students were able to learn from their teaching. Students were also able to assess what they learned metacognitively, examine the depth of their learning, and determine how to increase their mastery before taking a final. Since the creation and use of the practice organic examination in 2010 and 2011, respectively, partial results have been reported at several junctures (19–21). The results reported here summarize large compiled data from users of the Organic Practice Exam, and compares and contrasts these to reported results about using other practice exams in preparation for ACS secure finals (1, 17).



*Figure 5. Performance of 241 students on organic practice exam (50 questions) by % correct distributed across 10 bins. Low score: 0.0%; High score: 96%; Mean: 50.1% (SD) Median 50.0%; Standard Deviation: 15.3.*

## Instructor Information

Instructors were given performance results for individual classes taking the organic practice exam. Thus, each faculty member could use their specific data base to assess how benchmarks for their course were being met. They could also use the data to help students assess their skills. Usually this occurred when 70-80% of the second semester organic lecture course was completed.

This report compiles the results from all 241 students who completed the exam and provided mental effort ratings. This larger pool of data from multiple institutions and instructors examined noted trends that can inform instructors and students. The overall performance on the practice exam is summarized in a histogram in Figure 5. The number of students (*y*-axis) in each of 10 bins, by percent correct (*x*-axis) shows a typical distribution for the 241 exam takers. The highest score was 48 (of 50) correct; the lowest was 0 correct. The average was 50% correct.

Each item was rated for *difficulty* based on the number of *incorrect* responses. The more *incorrect* (or least number of *correct*) responses, the lower the *difficulty value*. The 50 questions were divided into 10 bins of 10% each as illustrated in Table 4. There were two questions that were rated *very easy* (bins 1 and 2) and none that were rated *very hard* (bins 9 and 10). In future iterations of the exam, these questions would be eliminated, because they give less informative data than more challenging questions.

**Table 4. Number of Questions of Varying Difficulty Going from *Easy* to *Hard* Top to Bottom**

<i>Bin #</i>	<i>% Incorrect</i>	<i>Number of Questions</i>	<i>Percent of Questions</i>
1	0-10	1	2
2	11-20	1	2
3	21-30	5	10
4	31-40	5	10
5	41-50	8	16
6	51-60	16	32
7	61-70	11	22
8	71-80	3	6
9	81-90	0	0
10	91-100	0	0

In addition to *difficulty*, as determined by the % correct for all 241 students taking the 50-question practice exam, each question was rated by faculty experts for *cognitive complexity*. The average complexity ranged from 3.75 to 11.70. The mean *cognitive complexity* was 7.37 (standard deviation 1.70); the median was 7.25. It was predicted that the questions with the highest *cognitive complexity* would have the lowest average percent correct and vice versa. In other words, a negative correlation would be observed when plotting *difficulty* as a function of *expert cognitive complexity* ratings. A plot of this data confirms this prediction, as shown in Figure 6.

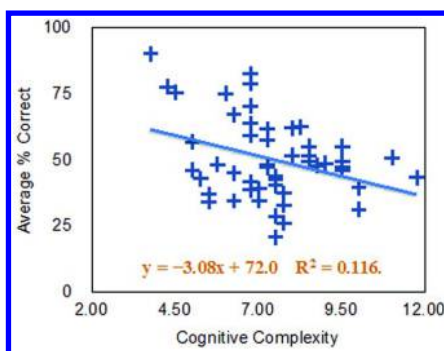


Figure 6. A linear correlation of student performance averages, as measured by difficulty, with respect to expert cognitive complexity. As predicted, an inverse relationship exists.

This equation is as predicted and trends in a similar fashion to the ACS Practice Exam reported earlier (18). The latter had a similar slope ( $-4.53x$ , when adjusted for percent), but a better correlation ( $R^2 = 0.195$ ). The weaker correlation on the data reported here versus a similar use of this instrument probably is a reflection of a weaker inter-rater reliability (between 0.70 and 0.75) of the experts in this study versus 0.83 in the earlier study (18). This study also had a smaller expert pool (5 vs 8). The expert raters for the Practice Organic Exam did not undergo a common training session. Still the current study shows a trend for complex questions to have a lower success for students, and less complex questions have higher average percent correct, just as the previously reported data. Thus application of the instrument developed for reliability and validity (18) to this current set of questions by a new group of *experts* shows very similar trends as illustrated in Figure 6, and provides information about cognitive complexity to faculty using practice exams. Some specific examples illustrate this.

An example of a question that had high average student success (75.3% correct) and a low rating for *cognitive complexity* (mean of 4.00) is question number 13 (Figure 7). This question is one that was rated (Table 3) as an *algorithmic* question (6); is in *ACCM* (10) category 4; and is in content category 4 (Table 1). This item correlated as predicted.



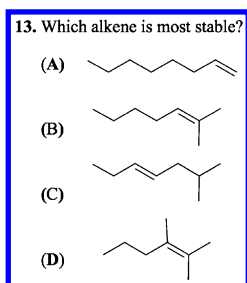


Figure 7. Students were able to apply algorithmic rules about alkene stability. 75% chose the key, "D" (9.5% chose A, 3.7% chose B; 10.0% chose C).

Faculty raters noted that there were only one or two *easy* subtasks (1-2 points) needed to answer this question. They rated the *amplification* as *easy*, 1 point, and the *distractor* role as *evaluation*, 2 points. Other levels of difficulty also had expected correlations.

An example of a question that had low average student success (31.0% correct) and a high rating for *cognitive complexity* (mean of 10.0) is question number 48 (Figure 8). This question is one that was rated (Table 3) as a *conceptual* question (6); is in *ACCM* (10) category 5; and is in *content* category 4 (Table 3).

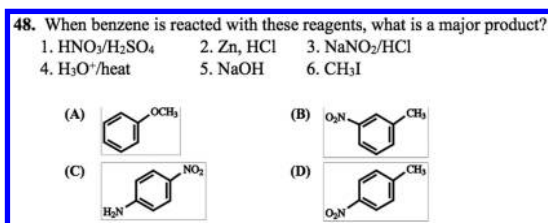


Figure 8. Students needed to complete many subtasks, namely multi-step reaction series to solve this question. 30.7% chose the key, "A" (34.0% chose B, 10% chose C; 24.1% chose D).

This item had great correlation between performance (low) on a *cognitively complex* (third highest) question. The stem of this question is typical for organic synthesis curriculum, and well-prepared students often perform very well on these types of questions, while poorly-prepared students often guess or choose an answer that seems most familiar.

An example of an item that did not correlate as well (percent correct vs. *cognitive complexity*) is number 25 (Figure 1) Faculty experts gave this question a modest mean *cognitive complexity* of 7.50. The student performance on this question was the third poorest (of 50) where only 28.6% chose the key, "D". The incorrect distractors were chosen as follows: 14.9% A, 24.9% B, 31.1% C. This

question outcome is particularly informative to instructors. It appears that experts believe this type of “addition to alkenes” question has a much higher mastery level for students than the evidence supports. The *cognitive complexity* was probably under-rated. More importantly, this outcome would encourage instructors to look at strategies for teaching this topic, as well assessing student mastery, especially where complex *amplification* is applied to difficult subtasks.

The Organic Practice Exam showed a strong reliability overall, with a KR20 rating of 0.849. Individual questions on the exam were also rated by discrimination using point biserial,  $r_{pbi}$ . Questions that had negative  $r_{pbi}$  were rejected outright. Questions where  $r_{pbi} < 0.15$  were also not included. Most of the questions used on the exam had moderate to strong discrimination (0.25–0.64) as summarized in Table 5. Six questions had weaker discrimination, but were retained because they measured important content for the course.

**Table 5. Percent of Questions of Varying *Discrimination* Going from *Weak* to *Strong* Top to Bottom**

$r_{pbi}$	<i>Number of Questions</i>	<i>Percent of Questions</i>
0.15-0.24	6	12
0.25-0.34	17	34
0.35-0.44	17	34
0.45-0.54	7	14
0.55-0.64	3	6

Question number **48** (Figure 8) had not only strong *cognitive complexity* correlation to student performance, but also had one of the highest measures of *discrimination*, 0.56. This question would be considered to have a low *difficulty* (30.7% correct); high *cognitively complex* (mean of **10.0**), and useful in discriminating between students who performed well on the rest of the test (well-prepared) and those who did not (had greater tendency to guess).

Easy questions can also have high measures of *discrimination*. A question was written to test understanding of acid/base mechanism steps. This was a *conceptual* question (6); is in *ACCM* (10) category 7; is in content category 4 (Table 1); and also has one of the highest measures of *discrimination*, 0.56. This question, however, had an *easy* rating and was answered correctly by 78.4% of students. Experts rated this question as moderately easy, with *cognitive complexity* mean of **6.75**. Still, it was useful at discriminating between students who performed well on the rest of the test and those who did not. From this question’s results, instructors can conclude that they are teaching this topic at a high level to the students who are well prepared.

The least *discriminating* question ( $r_{pbi} = 0.15$ ) on the exam was number 9 (Figure 9).

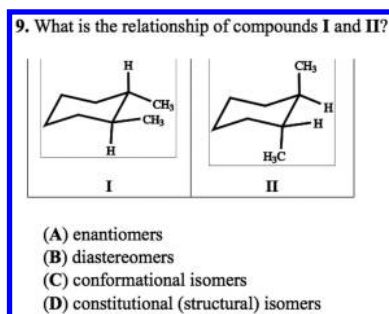


Figure 9. Students needed to complete many subtasks to solve this question. 21.2% chose the key, “A” (24.5% chose B, 39.0% chose C; 14.5% chose D; one student left blank).

This question tested as the item with the lowest *difficulty* on the practice exam (79.3% were incorrect). This question was one that was rated (Table 3) as a *conceptual* question (6); and is in *ACCM* (10) category 3; and is in content category 1 (Table 1). Faculty experts rated this question for *cognitive complexity* at 7.500 (average). As instructors, this outcome is puzzling and generates some discussion about the mismatch of the data. The most common answer was distractor C. When making this selection, the student is probably reaching for “low-hanging fruit”. It appears that strong and weak students alike (low  $r_{pbi}$ ) looked at the equatorial versus axial representation and assigned this as a conformational isomer without considering more depth of stereochemistry. By doing so, the students reduced this problem from a complex *conceptual* item to mere *recall* (6). The exam writers thought this was an important question to pose to students. However, the challenge here remains testing the ability to recognize that these structures are enantiomers with better discrimination. The committee of exam writers is looking at how this question could be rewritten.

Question number 9 introduces another important component of the results, the student-reported mental effort ratings. Students also rated this question as needing low *mental effort*, yet they performed very poorly.

### *Student Metacognitive Information*

Students were able to assess their abilities in organic chemistry after completing about 70-80% of the year-long course. They were able to examine how they performed versus how much *mental effort* they estimated was needed to answer each question. A comparison of student *mental effort* ( $n = 241$ ) to performance is predicted to have a negative correlation. This is indeed the case, as shown in Figure 10.

This equation is as predicted and trends in a similar fashion to the ACS General Chemistry Practice Exam reported earlier (18). The latter had a similar slope ( $-22.3x$ , when adjusted for percent), but a better correlation ( $R^2 = 0.534$ ). The weaker correlation found here is difficult to rationalize. This current study involved a large pool of students from a dozen institutions. Perhaps it is a reflection of inconsistent instruction about mental effort ratings.

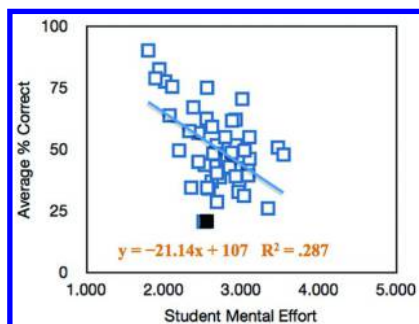


Figure 10. A linear correlation of student performance averages, as measured by percent correct, with respect to student-rated mental effort. As predicted, an inverse relationship exists. The filled-in data point represents question number 9.

Question number 9 again is an outlier, represented by the filled-in square data point in Figure 10. For this question, students gave a *mental effort* rating that was a mismatch for the *difficulty* rating, much like the faculty *cognitive complexity* rating. Students' ratings for mental effort on question number 9 under-estimated the complexity of the question, as did the faculty rating for cognitive complexity. In this case, the student and faculty ratings matched. Both sets of ratings for evaluating question number 9 indicated that the item should have been easier for the students than it was.

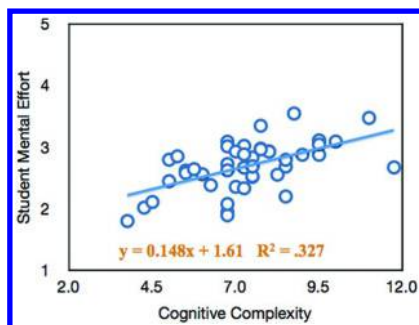


Figure 11. A positive linear correlation of student-rated mental effort, as measured by Likert scale averages, with respect to cognitive complexity exists.

For most of the questions on the Organic Practice Exam, there was a positive correlation between these two measures, as shown in Figure 11. Student mental effort ratings ranged from 1.79 to 3.54 (average was 2.69; standard deviation was 0.39). Faculty *cognitive complexity* ratings ranged from 3.75 to 11.75 (average was 7.37; standard deviation was 1.69).

This equation is as predicted and trends in a similar fashion to the ACS General Chemistry Practice Exam reported earlier (18). The latter had a similar slope (+0.199x) and a slightly better correlation ( $R^2 = 0.348$ ). These results confirm the general utility of the instrument developed to assess organic exams.

One of the trends noted in current assessment and organic curriculum (6) is the increase in questions that incorporate spectroscopy, whereas those that include qualitative analysis have diminished. Question number 37 is an example of the former. This is shown in Figure 12.

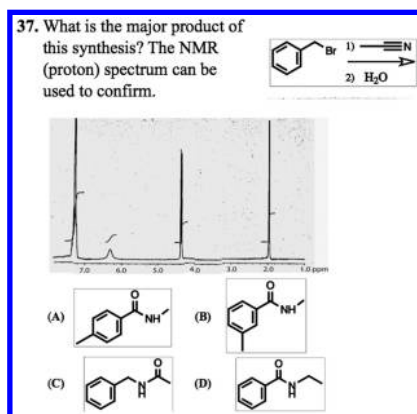


Figure 12. This was a difficult question (49.4% incorrect). Faculty rated it as the most cognitively complex question on the exam. Students rated it as one of the highest in mental effort. 50.6% chose the key, "C" (13.7% chose A, 5.0% chose B; 30.3% chose D; one student left blank).

This question had great statistics. Faculty gave this question the highest *cognitive complexity* rating of 11.75; students gave it one of the highest mental effort ratings, of 3.47. Yet many students found it worth the effort of solving the question with 50.6% correctly answering the question. The point biserial on this question showed high discrimination (0.37). This type of question is quite desirable as it likely reflects *higher order cognitive skills*, HOCS, (4).

**Table 6. Summary of Average on the 2004 OR04 ACS (2) Exam and Hour Exams**

<i>Group</i>	<i>Mean/Median/Std Dev: # Correct (of 70)</i>	<i>Mean/Median/Std Dev Percentile</i>	<i>Mean/Median/Std Dev Hour Exam %</i>	<i>Hr. Exam %/ ACS Percentile/Correlation (R<sup>2</sup>)</i>
All in Study	35/33/10	40/34/25	75/76/13	
Practice Exam	35/33/10	40/34/25	77/76/12	
No Practice	35/34/10	40/37/24	73/75/15	
ACS EI Norms (22)	39.2/38.5/12.2	50/48		
All Practice (162)				77/40/0.46
Low 1/3 (55)				66/23/0.30
Mid 1/3 (54)				76/36/0.08
Top 1/3 (53)				91/63/0.35
No Practice (107)				73/40/0.25
Low 1/3 (36)				56/30/0.03
Mid 1/3 (36)				75/37/0.01
Top 1/3 (35)				89/57/0.38

Also included in this table are the average percent hour exam grades of this same group of students and the norms for the OR04 (22). Data summarizing the hour exam average (column 1) for totals of each group and by approximate thirds based on average hour exam percentages. The corresponding 2004 OR04 percentile scores are listed in column 2 of this table. The correlation between the two is listed in column 3.

## Practice Exam Impact on Preparation for the Final ACS Exam

We expected the students who used the organic practice exam as part of their preparation for the final exam to surpass the performance of those students who did not. In order to investigate this, we looked at the performance of students taking the 2004 Organic ACS full-year exam. For this pool, 162 students had used the practice exam as part of their preparation for the final. The control group was 107 students who did not take and submit the practice exam. This study looked at student performances on the course hour exam (written by instructor) averages compared with their ACS exam scores. These are summarized in Table 6.

Several trends are evident from Table 6. All students participating in the study scored lower than the posted norms from the ACS Examination Institute (22). All students whether they participated in the practice exam as part of their preparation for the final, or did not, averaged 35 correct questions. This is 50% of the total questions on OR04, 70. Of interest, the average percent for all students taking the practice exam (241 from six institutions) was also 50%.

From Table 6, it appears that there is no significant difference in the final exam performance of students who took the Organic Practice Exam and those who did not. Students who did not take the practice exam had a slightly lower hour exam average before taking the final (73 versus 77), but performed equally well on the final.

To display the outcomes in a more visual way, the percentile results for the 2004 OR04 exam were divided into ten bins. These are summarized as a bar graph in Figure 13.

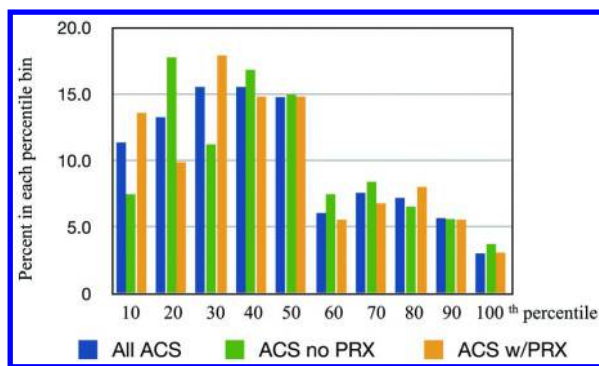


Figure 13. This summarizes 2004 ACS OR04 Percentile Performance.

Most students taking the ACS final exam were in the bins between the 20<sup>th</sup> and 60<sup>th</sup> percentiles. As a group, there was no measurable difference between the groups taking the practice exam as part of their study plan compared to those who used other methods to prepare. Sometimes, the group that took the practice exam appeared to do better on the final than those that did not. One case is the group in the 8<sup>th</sup> bin (71<sup>st</sup>-80<sup>th</sup> percentiles). More often, the opposite was the case as in the 6<sup>th</sup> and 7<sup>th</sup> bins (51<sup>st</sup>-70<sup>th</sup> percentiles). A similar pattern occurred for the low

performing groups in 1<sup>st</sup> and 3<sup>rd</sup> bins, where the percent of students performing at these low levels was higher for those taking the practice exam than for those not taking it. From a different perspective, the performances on the hour examination averages were parsed into thirds for both groups (practice exam group of 162; control group of 107)

Overall, the group taking the Organic Practice Exam seems to have a higher correlation with their respective performances on the ACS final (0.46 versus 0.25). Both groups showed a higher correlation between the final and their hour exam scores in the top third than in the lower or middle thirds. This data is illustrated in Figure 14 as a bar graph.

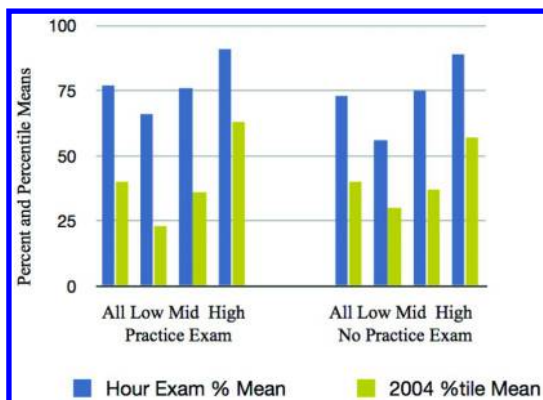


Figure 14. This summarizes 2004 ACS Percentile Performance (yellow/gray) with respect to average hour exams (blue/black) before the final. On the left are the students (all and by thirds) that took the practice exam. On the right are the students (all and by thirds that did not).

The final exam (ACS OR04) performance of the students taking the practice exam showed no measurable difference from those not taking the practice exam. This is also true of the middle third of each group, where the scores were identical for the hour exam average and the ACS final percentile averages. The students in the top third of the group taking the practice exam appear to have made some gains over their counterparts who did not take the practice exam, but these were not significant differences. The bottom third who took the practice exam demonstrated the least benefit from the process. This is not unlike previously reported results for the General Chemistry Practice Exam (1). We might posit the same hypothesis for these results. Perhaps the lower performing students did not benefit as much from the practice exam because were not able to use the results to develop a plan of study.

The overall lack of difference in final exam performance between students taking the practice exam and those not taking it seemed very surprising, at first. However, given the fact that the practice exam was used by the faculty and student study groups to help the entire class (not just those who took the exam) prepare



for the final, perhaps this should have been anticipated. After it was taken, all students attending study sessions or course lectures could ask questions about how to improve their toolbox of individual subtask skills and how to integrate those subtasks (*amplification*) and how to minimize *cognitive load* when taking a multiple-choice final exam. Other factors might also be involved with the noted outcomes. Perhaps a practice exam intervention came too late in the course to change the skill set needed to master the final exam.

Whereas there is no objective evidence that the practice exam improved the final exam outcome for the test group any more than the control group, students seek out this and other (3) practice opportunities. Students who provided data from the practice exam, as well as those who learned from post-practice exam activities benefitted equally. Faculty might consider methods of using practice exam interventions earlier in order to improve effectiveness.

## Summary and Future Uses of Practice Exam

Students and faculty can learn useful information by using a practice exam to help prepare students taking an ACS final exam. Instruments can be used to compare student perceptions as well as faculty predictions of exam question rigor. These have been shown to correlate as predicted.

Faculty can use practice exam data to ascertain areas where students need remedial help understanding certain content in organic chemistry.

At least one question on the Practice Organic Exam showed that using spectroscopy could probe higher order cognition skills in students. The spectroscopy question was had a low *difficulty* value (49.4% incorrect); faculty experts gave it the highest *complexity rating* (11.75); students gave it one of the highest *mental effort* ratings (3.47). For this question, the *increased mental effort* was rewarded for 50.6% of the students with a successful answer. That being said, when a question involved solving a problem using evidence that was not specifically encountered in the class or lab courses, students did not statistically demonstrate *higher order cognitive skills, HOCS*, (4). One question on the Practice Organic Exam asked about expected outcomes for organic qualitative analysis. Only 26.1% answered this correctly. This question also had borderline *discrimination* ( $r_{pbi} = 0.21$ ); faculty rated it *cognitively complex* (7.75); students rated it as needing high *mental effort* (3.34). And yet, with the effort put forth by students the results were not as favorable as with the spectroscopy question. Because the latter is taught specifically in many courses, this might indicate comprehension of the topic, but not necessarily *HOCS*. The real measure of *HOCS* might be solving problems when looking at evidence from a problem that was never directly taught.

Another example of a question that illustrates *HOCS* competency is the one in Figure 4. This was rejected as testing poorly because of low difficulty and negative discrimination. Faculty believed that this question was *cognitively complex*, but students who had developed *HOCS* should have been able to see similarities in patterns of UV-vis spectra (even if they had not been directly exposed to this topic). Faculty also believed that students with *HOCS* would recognize the importance

and projected analytical similarities for all conjugated carbonyls. Again, a small pool of evidence told us this was not the case. These are examples of highly *conceptual* questions that challenge students for mastery and faculty for teaching this mastery. This is not a unique challenge to organic chemistry, as learning command of conceptual questions is noted for introductory levels of chemistry (23, 24).

As faculty we sometimes hear the philosophy that only instructor-prepared assessments with opportunities for partial credit can measure *HOCS*. This seems unlikely. Students with a true mastery of these skills should excel in the multiple-choice format as well. Perhaps we should conclude that *HOCS* is indeed a difficult topic for mastery of students, but one that is needed for preparing our future scientists. The students taking organic chemistry as part of a plan to prepare for the Medical College Admission Test (MCAT) are told "...the MCAT is a standardized, multiple-choice examination designed to assess the examinee's problem solving, critical thinking, and knowledge of science concepts..." (25). Clearly there is a belief that multiple-choice questions can do this. Chemical educators have long promoted the value of teaching the *thinking* skills needed to master all sorts of assessments, not just the *recall* or *algorithmic* (26). The preparation of the practice organic exam demonstrated the challenges of preparing questions that accurately measure *HOCS*.

Students using the practice exam were exposed to methods for assessing their own gaps in knowledge and how to construct a deeper mastery of the subject. This current study did make an interesting observation about student mental effort on a practice exam. When looking at the students who took the practice exam, we noted that those students, who entered the final exam from the lower third of the average for instructor-written hour exams, reported a higher mental effort than the middle or the top third. These trends were noted from the data on the bar graph, but were not determined to be significantly different. However, the indicated trend is not surprising (Figure 15), as the top third is the group that is more "prepared" for the final, and thus, they seem to experience lower mental effort when answering questions.

The take-home message here might be to emphasize the importance of being prepared.

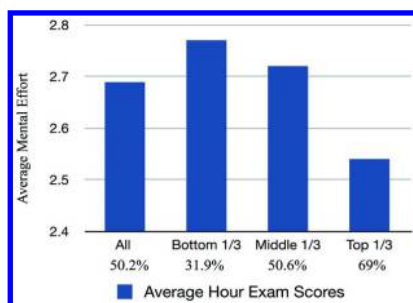


Figure 15. Students with higher average hour exam scores, rated questions to have lower mental stress when completing the practice exam.

The data acquired from the practice organic exam indicate that it is possible to prepare multiple-choice items as part of a quality department assessment program (27–29). By analyzing the exam described in this chapter we were able to note areas that needed to be augmented, and some that needed to be diminished. Whereas content seemed to be equally distributed, as shown in Table 1, the *ACCM* categories need a better distribution of questions (Table 3). Faculty preparing the exam intended to have one or more questions that addressed the importance of *global sustainability*, (7) but were unable to write an item of this nature that tested well. Faculty constructed a majority of items that they rated as *conceptual* (6). However, as illustrated by question number 25, students often approach these problems by trying to reduce them to *recall* if they can. In this case, *recalling* that axial versus equatorial defines conformational isomer led students to make an incorrect choice. It might be interesting to consider how many other questions students approached in a similar manner. Finally, faculty might consider how to help students develop skills to master these types of questions.

The authors of this practice examination discovered that it is not only challenging to teach students to solve *cognitively complex* problems, but also to write multiple-choice questions that discriminate student success. The importance of learning from high quality laboratory curriculum (5, 30) is also something the authors want to investigate with practice exam data.

The next iteration of this practice exam will include a total of nine NMR questions. These will be parsed into three levels (31).

1. Basic Level from Lecture
2. Advanced Level from Using Instrument
3. Research Depth from Organic II Project Based Lab Course

The nine questions will be added to the practice exam as part of the assessment for an NSF grant (32), which provided students with regular access to a high field NMR as part of their undergraduate organic laboratory courses. When used with exam items that have already been statistically measured as valid and reliable, the nine NMR questions will replace some of the items (especially in *ACCM* group 5) and will be used to probe how the use of the instrument impacts depth of understanding of NMR as part of organic chemistry.

## Acknowledgments

The authors would like to thank all colleagues who helped construct exam question items: Melvin Druelinger, Douglas Dyckes, Vanessa Fishback, Thomas Bindel and Donald McElwee. They also thank Karen Knaus and Margaret Asirvatham for advice on statistical data collection and analysis, and Kristen Murphy for setting the exam. Finally, the authors would like to thank the students who contributed results for this study and possible future looks at practice examination data: University of Colorado-Denver, Colorado State University-Pueblo, Converse College, Jamestown College, Goshen College, and Metropolitan State University-Denver.

## References

1. Knaus, K. J.; Murphy, K. L.; Holme, T. A. *J. Chem. Educ.* **2009**, *86*, 827–832.
2. American Chemical Society Division of Chemical Education's Examinations Institute. *OR04 Organic Chemistry 2004-Exam for two-semester Organic Chemistry*; Ames, IA, 2004.
3. American Chemical Society Division of Chemical Education's Examinations Institute. *Preparing for Your ACS Examination in Organic Chemistry: The Official Guide*; Ames, IA, 2011.
4. Zoller, U.; Fastow, M.; Lubezky, A. *J. Chem. Educ.* **1999**, *76*, 112–113.
5. Zoller, U.; Pushkin, D. *Chem. Educ. Res. Pract.* **2007**, *8*, 153–171.
6. Raker, J. R.; Holme, T. A. *J. Chem. Educ.* **2013**, *90*, 1437–1442.
7. Zoller, U. *J. Chem. Educ.* **2012**, *89*, 297–300.
8. Paas, F. G. W. C.; Van Merriënboer, J. J. G. *Hum. Factors* **1993**, *35*, 737–743.
9. Murphy, K.; Holme, T.; Zenisky, A.; Caruthers, H.; Knaus, K. *J. Chem. Educ.* **2012**, *89*, 715–720.
10. Raker, J. R.; Holme, T. A.; Murphy, K. *J. Chem. Educ.* **2013**, *90*, 1443–1451.
11. Apperson Products, 851 SW 34<sup>th</sup> Street, Bldg. B, Renton, WA; <http://www.appersonedu.com/> (accessed September 3, 2014).
12. Brown, J. D. *Understanding Research in Second Language Learning: A Teacher's Guide to Statistics and Research Design*; Cambridge University Press: Cambridge, MA, 1988.
13. Linacre, J. *Rasch Meas. Trans.* **2008**, *22*, 1154–1157.
14. Kuder, G. F.; Richardson, M. W. *Psychometrika* **1937**, *2*, 151–160.
15. Guilford, J. P. *Fundamental Statistics in Psychology and Education*; McGraw-Hill: New York, 1965.
16. Bodner, G. M. *J. Chem. Educ.* **1980**, *57*, 188–190.
17. Raker, J. R.; Trate, J. M.; Holme, T. A.; Murphy, K. *J. Chem. Educ.* **2013**, *90*, 1290–1295.
18. Knaus, K. J.; Murphy, K. L.; Blecking, A.; Holme, T. A. *J. Chem. Educ.* **2011**, *88*, 554–560.
19. Schelble, S. M.; Murphy, K. L.; Knaus, K. J. *Assessing Conceptual Concepts of Organic Chemistry*. Presented at the 21<sup>st</sup> Biennial Conference on Chemical Education, Denton, TX, August 3, 2010.
20. Schelble, S. M.; Murphy, K. L.; Knaus, K. J. *Outcomes from the Use of a Collaboratively Developed Organic Chemistry Practice Exam*. Presented at the 241<sup>st</sup> National Meeting of the American Chemical Society, Anaheim, CA, March 31, 2011.
21. Schelble, S. M.; Murphy, K. M. *Study of Student Growth Through the use of a practice ACS Examination*. Presented at the 244<sup>th</sup> National Meeting of the American Chemical Society, Philadelphia, PA, August 20, 2012.
22. American Chemical Society Division of Chemical Education, *Composite Norms, Organic Chemistry Form 2004*; <http://chemexams.chem.iastate.edu/national-norms/or04.html> (accessed September 3, 2014).
23. Nurrenbern, S. C.; Pickering, M. *J. Chem. Educ.* **1987**, *64*, 508–510.

24. Nakhieh, M. D.; Lowrey, K. A.; Mitchell, R. C. *J. Chem. Educ.* **1996**, *73*, 758–762.
25. *Association of American Medical Colleges and related marks*; <https://www.aamc.org/students/applying/mcat/about/> (accessed September 3, 2014.)
26. Moore, J. W. *J. Chem. Educ.* **2008**, *85*, 763.
27. Towns, M. H. *J. Chem. Educ.* **2010**, *87*, 91–96.
28. Bergendahl, C. *J. Chem. Educ.* **2005**, *82*, 645–651.
29. Bretz, S. L. *J. Chem. Educ.* **2010**, *89*, 689–691.
30. Duis, J. M.; Schafer, L. L.; Nussbaum, S.; Stewart, J. J. *J. Chem. Educ.* **2013**, *90*, 1144–1150.
31. Schelble, S. M.; Murphy, K. M. *Measuring Learning Benchmarks when Students have Access to High-Field NMR*. Presented at the 247<sup>th</sup> National Meeting of the American Chemical Society, Dallas, TX, March 16, 2014.
32. National Science Foundation TUES (DUE #1245666), August 2013–August 2015.

## Chapter 6

# Feedback in Testing, the Missing Link

Jamie L. Schneider,<sup>\*1</sup> Sara M. Hein,<sup>2</sup> and Kristen L. Murphy<sup>3</sup>

<sup>1</sup>Chemistry Department, University of Wisconsin River Falls,  
410 S. 3rd Street, River Falls, Wisconsin 54022

<sup>2</sup>Chemistry Department, Winona State University, 175 W. Mark Street,  
Winona, Minnesota 55987

<sup>3</sup>Department of Chemistry and Biochemistry, University of  
Wisconsin-Milwaukee, 3210 N. Cramer, Milwaukee, Wisconsin 53211

\*E-mail: [jamie.schneider@uwrf.edu](mailto:jamie.schneider@uwrf.edu).

There are many strategies to assess student learning but a very common approach is still individual tests, often in multiple-choice formats. Many college science courses assess student learning with several unit tests and a cumulative final exam. Much research has been published on reforming the content and design of summative tests. Cognitive psychology literature suggests that feedback is essential to enhancing long-term memory and student performance on repeat assessments. Although there is ample psychology literature on feedback affects, especially with rote memory applications, the research on Science, Technology, Engineering, and Mathematics (STEM) testing with feedback is limited. This chapter provides a summary of key literature on testing feedback and its effects on undergraduate student performance. Where possible, references from STEM literature will be cited including the design of our chemistry testing feedback research study; however, much of the literature comes from the area of cognitive psychology.

### Introduction

Gibbs and Simpson (*1*) describe the functions of feedback: “to correct errors, develop understanding through explanations, generate more learning by suggesting further specific study tasks, promote the development of generic skills by focusing on evidence of the use of skills rather than on the content,

promote meta-cognition by encouraging students' reflection and awareness of learning processes involved in the assignment, and encourage students to continue studying" (pp 19-20). By definition, summative assessments are used to provide evidence for evaluating or grading students. Does this definition preclude that summative assessments cannot serve as learning opportunities for students? Some argue that summative assessments are distinctly separate (2, 3) suggesting that students tend to pay less attention to feedback when evaluations count significantly toward the students' final grade. Yet, many practitioners frequently attempt to provide some type of feedback to students typically ranging from test scores to posted answer keys to class discussions with hopes of students utilizing the feedback in some way. What feedback best practices should be employed in these individual testing experiences, especially for unit exams leading to a cumulative final exam, potentially to promote student learning? There is a plethora of literature on testing affect and testing feedback affects in the cognitive psychology literature to promote student learning. The difficulty with this literature is that there are many conflicting reports depending on experimental and classroom conditions. This chapter will provide a summary and some key lead references to help illustrate the importance of testing and testing feedback on student learning at the collegiate level. Recommendations will be made for areas of further study, specifically as it relates to testing in STEM.

## **Testing Design**

There are several inherent principles (or assessment tasks) to consider when writing a test or exam. Vitale, Romance, and Dolan (and references cited within) describe four that are commonly recognized (4). First, tests provide a sample of student behavior—the answer a student gives to the question directly. These behaviors could range from a written answer to a true/false choice. Second, one can infer or conclude about other observable student behaviors of interest using the testing data. This is probably the most used reason for giving a test. How well a student has learned the material presented is generally the inference instructors make based on a how a student performs on the test. Test validity is the third testing principle. This indicates the extent to which the test is credible for making the types of inferences desired. The last principle is reliability which focuses on the amount of consistency of a measurement outcome. This is especially important if assessment is to be carried out across various samples. Careful attention to these principles should allow the instructor to make informed instructional decisions about test design.

A common testing format in undergraduate science courses includes the use of multiple-choice items (5). These items are regularly machine scored allowing for quick generation of statistical reports of student performance. Other advantages include allowance for the inclusion of more questions in a short period of time, assessment of a wider range of topics, higher reliability, and possible diagnostic information by careful study of incorrect answer choice frequencies (6). Limitations include difficulty associated with constructing valid

items, challenges in writing plausible incorrect responses or distractors, scores influenced by students' reading ability, inability to measure all instructional goals, and scores influenced by guessing (6).

Can multiple-choice tests be used to assess students' knowledge and reasoning proficiencies? The answer to this question is debatable; it depends in large part on how the exam is written (7–10). In short answer and problem-solving questions, the thought process of students can usually be followed in constructed response exams even if a student arrives at an incorrect answer. This is not the case in most multiple-choice exam formats. With multiple-choice exams, it is difficult to assess a student's ability to apply what they know with a multiple-choice exam unless it is written purposefully for that outcome.

Along with testing and item formats, instructors need to think about the frequency of testing and the overall cumulative nature of repeat tests. The frequency of testing will be discussed in the next section on Testing Affect. Distributed practice literature (11, 12) suggests that tests and practice tests that require repeated review of earlier material enhance student performance on cumulative final exams. In courses that consist of a collection of topics, this distributed practice needs to be purposefully integrated into the test design. However, courses that have a hierarchical structure naturally build this distributed practice because later topics build upon knowledge of prior topics. Many introductory chemistry courses have a hierarchical structure allowing for a naturally distributed practice on unit exams.

## Testing Affect

As mentioned in the introduction, much effort from practitioners is placed on test design, but what happens after the test? Is a student's course content understanding affected by the act of taking the test? Test affect is the effect on student learning from just taking a test. It can also be described as the long-term effect whereby long-term retention is improved by successful retrieval of information through taking a test. Although much literature focuses on the cognitive changes from taking a test, there is also research on the affective changes from taking a test. The affective changes as a result of test taking are beyond the scope of this article; however, Crooks provides a nice literature summary which highlights changes in students' intrinsic and continuing motivation, expectations for learning and study behaviors, anxiety, self-efficacy, attributions for success and failure, and motivational aspects of competitive, individualistic, and cooperative learning structure (13). As illustrated by this lengthy list, classroom evaluation (testing) has a tremendous impact on students. Crooks (13) states that evaluation is "one of the most potent forces influencing education. Accordingly, it deserves careful planning and considerable investment of time from educators" (p 467).

Bangert-Drowns, Kulik, and Kulik published a thorough meta-analysis on the effects frequent classroom testing in pre-college and college mathematics, science, and social studies courses has on students' performance on end of instruction cumulative exams (14). They used data from 35 studies that met the following criteria: took place in real classrooms, used identical instruction



between comparison groups except for frequency of testing, utilized conventional classroom tests and a common summative end-of-course exam, and contained no serious methodological flaws. Twenty-nine of 35 studies found positive effects as a result of frequent testing. Thirteen of these positive effect studies were statistically significant; whereas, only one of the negative reports was statistically significant. In the 35 studies, frequent testing raised achievement scores by an average of 0.23 standard deviations. However, the achievement increase (or decrease) was highly varied between studies ranging from 0.96 to -0.80 standard deviations. They also carried out regression analysis taking changes in test frequency between control and experimental groups into account. They found that adding more tests had a smaller effect on gains in performance on end-of-term exam with each additional test. The first semester test showed the largest gain on the end-of-term exam. For the studies utilized in this meta-analysis, there was no mention of test feedback other than to say that “ordinary classroom tests are often used without feedback” and that “students are usually aware that their test performance is a one-time event that contributes to the student’s academic record” (p 97) (14).

Classroom research investigating the value of practice exams has shown few significant gains (15–17). Like with the studies mentioned above, many of these studies focused more on testing and less on feedback optimization. A more recent example can be found in chemistry utilizing a practice exam. Knaus, Murphy and Holme (18) investigated change in Z-score comparisons between students who took a multiple-choice practice exam one week prior to a final exam and students who did not. Practice exam total score and category scores were provided one day after testing. Students who took the practice exam showed a decreased performance on the final exam compared to their performance on three unit exams given prior to the practice exam. However, upon closer inspection, students who performed below average on the practice exam showed positive changes in Z scores and students with above-average scores showed negative changes. The authors (18) suggest that perhaps “students who do well believe that they need less study and subsequently have lower performance” (p 831). The authors note the importance of coaching students on how to use practice exam information. Practice tests may also be an important area to think more closely about optimizing testing feedback.

In 2010, Butler published a study that looked carefully at the effects of repeated testing or repeated studying with careful control of the feedback (19). Participants were asked to read four passages on a computer that contained facts presented in a single sentence and concepts derived from multiple sentences. After each passage, participants were then asked to reread the passage (restudy) or take a cued-recall format test (or short answer test) on the passage. After students entered their test answers which were typically 1-3 sentences long, they were shown the question with the correct answer. One week after the first session, participants were asked to take a final cued-recall test which contained factual and conceptual questions that were either verbatim, rephrased, or inference questions compared to tested questions given previously. Butler found that participants consistently scored higher on the final test for repeated testing conditions compared to restudy conditions. This pattern was consistent for experiments involving factual and

conceptual items as well as verbatim, rephrased, or inference questions. Butler (19) concluded that “relative to repeated studying of passages, repeated testing led to better performance on new inferential questions that required the application of previously learned information to produce a new response” (p 1124).

## Testing Feedback

In the early studies on test affect, there was considerable variability in the effects of repeat testing on student performance suggesting that there were additional factors to consider. One important factor to control for is testing feedback, as noted in the 2010 Butler study. Providing optimal testing feedback may not be a priority for many practitioners. Critics suggest that summative tests are too focused on what an external person (the instructor) can gain from the results rather than on gains in learning for the user (the student) (20). When utilized appropriately, testing feedback has the potential to satisfy this critique by providing the student opportunities for learning (21, 22). Wiggins (20) defines feedback as “information that provides the performer with direct, usable insights into current performance, based on tangible differences between current performance and hoped-for performance” (p 182). Many of the arguments surrounding the one-sided nature of tests has fueled discussions in higher education to promote more formative assessment practices including distributed practice events like quizzing and practice exams. In many ways, the sheer design of formative assessments is to provide feedback to both instructors and students. Formative assessment often involves a combination of corrective and timely feedback designs, as well as social constructs to permit interactions with other students (23). Several authors have written about general classroom assessment practices (1, 24). Many more suggestions from chemistry courses are contained within this book. Although there are many instances of positive outcomes from providing feedback, there are also a few noted situations that have been shown to have no effect or even negative effects on student learning which include: tasks that students are capable and willing to produce their own feedback, very easy tasks where feedback is unnecessary, and instructional feedback provided prior to completion of the tasks (25). Interestingly enough, the use of pretests diminishes the negative effect of these types of feedback. Therefore the use of pretests combined with feedback does pose an interdependent combination that leads to positive effects (25).

In addition to feedback provided through formative assessment techniques, testing feedback on unit exams has the potential to provide students learning opportunities, particularly for courses that have a hierarchical structure (26). As mentioned earlier, multiple-choice testing is a popular testing format in STEM courses. Multiple-choice exams present a bit of a conundrum for the effects on student learning. The act of taking the test can improve retention as noted by the testing effect; however, students reading or endorsing the lure items (or distractors) may result in the acquisition of incorrect knowledge (27, 28). This affect has been shown to increase for less prepared students (28). Utilizing appropriate feedback can reduce the negative effects and enhance the

positive effects on student learning (25, 29). Butler and Roediger (29) also found that feedback minimized differences in these negative effects for students with differing test preparation. The rest of this chapter will provide feedback literature on studies utilizing mainly multiple-choice testing formats. Illustrated in the examples provided, there are several variations in the type and timing of the feedback.

## **Types of Multiple-Choice Testing Feedback**

Several types of corrective feedback have been studied including: indicating right/wrong (verification), providing the correct answer, allowing answer-until correct, and providing explanations. It should be noted that all of these types of feedback provide some type of item by item corrective feedback to the student. Providing no feedback or even posting total score correct, referred to as non-corrective feedback, consistently has lower effects on student learning compared to corrective feedback (29). It is also generally accepted that providing the correct response is more effective than simply indicating whether the response is correct or incorrect (29–31). This correct/incorrect feedback provision is referred to as verification feedback. Hancock, Stock and Kulhavy found that participants spent less time processing correct/incorrect feedback compared with feedback that provided the correct answer in a study that utilized multiple-choice items from College Board Achievement tests that included history, social studies, world cultures, biology, chemistry, and physics (32). Lhyle & Kulhavy (33) suggested that feedback designed to “lead students to process, study, or apprehend the feedback more closely should increase the amount of correction that takes place” (p 320) based on work from a study involving a multiple-choice test on passages related to the structure and function of the human eye.

Answer-until-correct feedback format was originally developed by Pressey in 1926 (34). In this format, participants answered multiple-choice questions by selecting from the choices until the correct answer was revealed. More recently, a commercially available product called the Immediate Feedback Assessment Technique (IF-AT®); [www.epsteineducation.com](http://www.epsteineducation.com), accessed July 2014) was developed to offer an answer-until-correct format in the context of large classroom settings without technology requirements. This technique will be discussed further in the next section because this technique also relates to feedback timing.

Is there a limit to the benefits of providing more detailed corrective feedback? The use of elaborative feedback is more complex than simply noting the correct answer choice. It can contain an explanation for the correct answer choice or a representation of the original study materials. Butler, Godbole, and March noted that educators and researchers generally have an assumption that elaborative feedback is superior to correct answer feedback (35). However, Bangert-Drowns, Kulik, Kulik, and Morgan (25) state that they did not find a relationship between the amount of information and feedback effects in the conclusion of a meta-analysis on effects of feedback in test-like events. They proposed two possible explanations for this finding. First, the studies used in their meta-analysis emphasized fact-based retrieval which “makes students most interested in the correct answers, and

thus they do not mindfully attend to more detailed explanations” (p 234) (25). Second, they suggested that the content assessed was too simple or specific and may not require elaborate feedback. They also suggested that elaborate feedback “may be more important in the building of conceptual frameworks, drawing of inferences, or applying of rules in complex situations” (p 234) (25).

Butler, Godbole, and March expanded on earlier studies to examine transference of knowledge via testing with new inference questions utilizing a computer testing system (35). While the protocols resembled earlier work, the application to transference questions addressed the earlier suggestions provided by Bangert-Drowns, Kulik, Kulik, and Morgan (25). In this study, students read 10 passages that contained two critical concepts per passage. After reading each passage, they completed a short-answer test on the critical concepts that consisted of definition questions. Immediately after answering each question, subjects were either given no feedback, correct answer feedback (a re-representation of the question and statement of the correct answer), or explanation feedback (a representation of the question, a statement of the correct answer, and two additional statements taken from the passage to explain the concept). Students then returned two days later to take a short answer test, half of which contained new inference questions. Significantly more inference questions were answered correctly when explanation feedback was provided as compared to no feedback or correct-answer feedback. Although this study did not use multiple-choice formats, it is important to note this discrepant result when questions required transfer of knowledge.

## **Timing of Multiple-Choice Testing Feedback**

Although researchers agree that corrective feedback is superior to non-corrective feedback, the optimal timing of feedback continues to be debated (36). Theories supporting the effectiveness of immediate feedback stem from original behaviorist theories of reinforcement (37). The major premise of this research is that delays in feedback reduce the efficacy of the feedback. This is particularly evident for students with intellectual challenges. The greater the delay in feedback, the less opportunity there is for learning (38). Proponents of delayed feedback have argued that the delay allows students to forget errant responses as tendencies are strengthened by taking tests (21). This is because incorrect responses must be forgotten or they will interfere with the acquisition of correct responses. By delaying feedback, students will be less apt to repeat errors in subsequent exams. Because multiple-choice exams offer other options or distractors for answers, the incorrect answers may interfere, causing students to focus on incorrect concepts. This phenomenon is termed the Delayed Retention Effect (DRE) (22). Delayed feedback also allows for spaced presentation of thinking—that is, students thinking about the test questions on more than one occasion. Spaced presentations of information are associated with better learning compared to massed presentations. This is currently thought to be the most important positive aspect of delayed feedback (26).

The timing of feedback can vary from immediately after each item, immediately after the test, or delayed after the test. One confusing issue in the literature is how these different timings can be described and utilized by different researchers. Some classroom researchers describe immediate feedback for feedback given in the next class session after a test (39). More commonly however, experimental psychologists describe this type of feedback as delayed (40). The expression immediately after the test might imply immediate feedback; however, some researchers describe this as delayed feedback because it was given after the student had moved on to other thought processes. Still other delayed spacing may range from a few days to a few weeks (36). In addition, the duration between the initial test and feedback relative to the final test can be varied. Researchers also use the terms massed presentation versus spaced presentation to indicate feedback that is delivered during and/or immediately after the testing event versus feedback that is delivered at a later time separate from the testing event (26). The last consideration from the feedback literature focuses on the outcomes the researchers are interested in studying. Some studies focus on retention of correct responses; whereas, other studies focus on correction of incorrect responses (36). The positive accumulation of both effects would result in better performance with repeat testing.

Memory enhancement from retrieval of information, perhaps during a testing event, supports incorrect information just the same way that correct information is enhanced. Students often believe that if they are able to recall the information, then it must be correct (41). Immediate and delayed feedback provides information that is retrieved by students at a later time, and it also helps to run interference with assumptions on correctness. Students can use the new knowledge of correct and incorrect answers if subsequent questions on the material are offered after the initial test (42).

In 1988, Kulik and Kulik published a meta-analysis on what would now be considered historical studies on the timing of feedback (36). From the meta-analysis of 53 studies on feedback timing, they found that classroom studies that utilized quizzing consistently showed positive effects for immediate over delayed feedback with average effect size of 0.28. However, laboratory experiment results were much less clear with some conditions favoring delayed and some favored immediate feedback. One such example that consistently favored delayed feedback was experimental designs aimed at studying acquisition of test content. In this design, “test item stems are used as the stimulus material and the correct answer is the response to be learned” (p 80) (36). The immediate feedback condition was almost always found to have lower affects compared to the delayed feedback condition with an average affect size of -0.36. In list-learning experiments, results were highly variable, but overall they favored immediate feedback conditions with an average affect size of 0.34. Kulik and Kulik (36) concluded “that delayed feedback appears to help learning only in special experimental situations” (p 94).

Given the positive endorsements of immediate feedback, Michael Epstein developed the Immediate Feedback Assessment Technique (IF-AT®) form to easily provide answer-until-correct immediate feedback in a classroom setting without the use of technology (38). The IF-AT® multiple-choice answer sheets

have a thin opaque film covering the answer options. Students are instructed to scratch off the film to reveal a blank box (incorrect) or a box with a star (correct) and then to continue to scratch-off another option and so on until the star is revealed. The idea is that getting students to actively generate the correct response after they made an error may engender deeper processing and further retrieval processes compared to simply reading the correct answer (43, 44). Anderson contradicted this idea by suggesting that selection of multiple incorrect answers may cause competing memories and may actually interfere with the ability to remember the correct response (45).

The earlier IF-AT® research involved introductory psychology classroom studies that utilized multiple-choice items that primarily assessed student's knowledge of definitions of basic psychology concepts (46–49). In a more extensive study, 611 undergraduates took five multiple-choice course exams and a final exam (50). The final exam contained 50 new items and 10 items randomly selected from each course exam. A subsample of these students (467 out of 611) took long term retention tests at 3, 6, 9, and 12 months after the final exam consisting of 25 new items and 5 items randomly selected from each course exam. Students were partitioned into one of four groups: end-of-test feedback group (answers were recorded on bubble forms and students reviewed answer sheet, test items, and the correct solutions for 30 minutes immediately after the test was complete), delayed feedback group (answers were recorded on bubble forms and student reviewed an answer sheet, test items, and the correct solutions for 30 minutes 24 hours after the test was complete), immediate feedback group (answers were recorded on an IF-AT® form and students were allowed to continue to scratch off the waxy coating until the correct answer was revealed), and control group (answers recorded on bubble form and students were given their machine-scored bubble forms back the next class period). All participants used a bubble form to record final exam responses. Because no significant differences were observed for the end-of-test feedback group and the delayed feedback group, both of these groups were pooled into one delayed feedback group in the results.

Final exam “scores were higher on the repeated test items for (a) both feedback groups than for controls, and (b) for the immediate feedback group than for the delayed feedback group” (p 397) (50). The scores for the novel test items showed very little differences between the three groups. They compared final exam data on earlier to later course exams (weeks 3 to 11) to measure the potential to forget correct responses and to correct initially incorrect responses during a semester. These were measured as conditional probabilities or fractions. For instance, correct answer conditional probabilities are determined by finding which items were correctly answered on the initial test and still answered correctly on the repeat test. The number of items consistently answered correctly on both initial and repeat tests is divided by the number of items answered correctly on the initial test. A high value represents consistency in a student's ability to answer items correctly on both the initial and repeat tests. Correction conditional probabilities focus on the items that were incorrect on the initial test but now are correct on the repeat test. Here a high number represents a high proportion of corrects from initial to repeat test. The immediate feedback group had higher conditional probabilities for maintaining the correct answer and for correcting an

initially incorrect response when compared to the delayed and control groups. The post-semester data showed that the attrition rates were quite constant across the three treatment groups over the 3 to 12 month period. However, the conditional probabilities of maintaining the correct answer or correctly answering an initially incorrect response were consistently higher for the immediate feedback group compared to the delayed and control groups. In their concluding remarks, they state that the “immediate feedback transforms the multiple-choice examination into a learning opportunity and the student into an active learner” (p 406) (50).

More recently, two additional studies were published utilizing IF-AT® forms for testing in a pharmacokinetics course and a higher-level athletic training course. The pharmacokinetics study included questions requiring assimilation of multiple elements for successful problem solving (51). Students were given three exams including a final exam. Exam questions were arranged in order of increasing cognitive activity according to Bloom’s taxonomy. The results of the IF-AT® multiple-choice exams were compared to mixed-format exams (50% open response questions, 50% mix of multiple-choice, true-false and short-answer) given in a prior year when immediate feedback was not provided. No significant differences in final exam test scores were observed. This study directly contrasts some of the earlier work on testing with IF-AT® forms; however, the authors note that there may have been some variability in item difficulty between the two sets of exams used in this study. It is also possible that immediate feedback on problem-based questions was not sufficient to affect learning. In the athletic training study, students took a traditional multiple-choice exam or an IF-AT® multiple-choice exam for exam 1 and reversed the response option for exam 2. The next class period after each exam, the instructor reviewed the examination with all students. One week after each exam, the students took a follow-up exam that contained the same materials as the initial test but with reordered items and answers. They found that both groups had statistically significantly higher scores on the follow up exam, with no difference based on method. One issue with this methodology was that both groups actually received delayed feedback. Both of these studies expand the research into new content areas and both contradict some of the earlier IF-AT® work.

In a 2007 study, Butler, Karpicke, and Roediger challenged some of the classroom IF-AT® work stating that although the design was convenient for classroom studies it did not effectively isolate the type and timing effects (52). In the IF-AT® studies, students in the IF-AT® group immediately actively engaged in generating an answer; whereas, the delayed feedback group passively read the answers. Butler, Karpicke, and Roediger utilized computer based laboratory experiments where participants read passages from GRE, TOEFL, and SAT study guides and after each passage answered fact-based multiple-choice questions. Over the course of the 12 passages, each participant experienced two types of feedback conditions (standard and answer-until-correct (AUC)) and two control conditions (no test and test with no feedback). The standard feedback consisted of an indication of accuracy (correct/incorrect), a re-presentation of the question, the response selected, and the correct answer. The participants experienced the feedback condition immediately after each item or after a 10- minute delay while carrying out a distractor task following the test. Both groups were exposed to

the standard and AUC feedback in the same fashion—requiring both groups to actively engage with the feedback. The next day, participants returned to take a final cued-recall test that contained exactly the same item wording as the multiple-choice items but the answers were either words or short phrases. The feedback conditions outperformed the control conditions. The delayed and immediate feedback groups had similar proportion of correct answers on the final cued test regardless of whether they received standard or AUC feedback. The delayed feedback group outperformed the immediate feedback group (0.73 vs 0.68 proportion correct;  $p = 0.61$ ); however, this result was not significant.

To investigate long-term retention further, they carried out a second experiment in which the final cued-recall test occurred one week after the initial test and the delayed feedback was given one day after the initial test. Once again, the delayed and immediate feedback groups experienced the same active engagement with the feedback. The type of feedback was not a factor in influencing the proportion of correct responses. However, the timing of the feedback was a factor. The delayed feedback generated statistically higher proportion of correct responses (0.60 vs 0.70;  $p = 0.93$ ). As was mentioned in the Kulik and Kulik meta-analysis, the laboratory experiments more often favored delayed feedback over immediate feedback. Butler, Karpicke, and Roediger (52) suggest that a “viable explanation for the superiority of immediate feedback in some studies is that students sometimes may not fully process feedback after a delay unless required to do so (as in laboratory studies)” (p 280).

## **Student Response to Feedback**

In higher education, students have remarked that the process involving feedback delivery (or lack thereof) is an unsatisfactory experience and that a true educational experience must include a method for closing the loop on assessments (53). Students also generally favor multiple-choice tests over open-ended types of assessments such as essays and short-answers (54). Therefore, the implementation of feedback in multiple-choice exams presents an interesting array of factors that must be considered. Besides the aforementioned testing effect, it has been shown that the engagement of the retrieval process, on previously tested items, leads to better performance on a final exam. Kang, McDermott, and Roediger were able to link the testing effect with both format and corrective feedback and found that student performance on subsequent tests was affected by both (55). In this study, an initial test was given in either multiple-choice or open response form in which the students received no feedback. Under these conditions, students who took the multiple-choice test performed better on the final compared to the students who took an open-response test. In their second experiment, corrective feedback was given under the same conditions. In this case, students who took open-response tests outperformed students who took multiple-choice tests. Interestingly, the amount of effort involved in retrieving information on open-response formats played a significant role in the increase on the final tests as compared with multiple-choice testing. Although one might assume that students would prefer to do less work on an exam, the amount of work involved to achieve the correct



answer seems to have a direct impact on student performance. The transference of knowledge in multiple-choice tests, through feedback acquisition, may have its limits in terms of student engagement. One way to increase engagement may be to scaffold questions to provide the necessary framework that renders feedback retrieval necessary.

What determines if a student wants to look at feedback or when they do not want to? This question is difficult to answer as much of the literature presents the effects of learning when feedback is used but little of it presents information about students' desire to use it or the best method by which to provide it. One might assume that students who select incorrect answers might be more likely to look at feedback than students who select correct answers. However, in a study that controlled for access to feedback in a computer-based study, it was found that there was no significant difference in the amount of time students spent looking at the feedback for correct versus incorrect answers (56). These results suggest that when given the option, students will look at feedback regardless of correctness. Studies have also shown that students will perform better on tests by simply knowing that feedback will be provided (57, 58). However, the timing and acceptance of the feedback is based on students' initial understanding about their individual intelligence (59). Students who believe that their intelligence is an ever-changing entity anticipate provided feedback will help them on their exam; whereas, students who believe that their intelligence is a stagnant condition do not welcome the feedback and view it as a negative representation of their abilities. These findings connecting views of intelligence with feedback would suggest that feedback is helpful to some students and hurtful to others. Additionally, students bring preconceived notions about their abilities to each test. These predisposed confidence levels have been investigated and present the challenge of differentiating between the types of questions that lead to differing confidence levels (21, 29, 60).

A number of studies involved the investigation of students' responses to feedback specifically in STEM courses. Mullet, Butler, Verdin, von Borries, and Marsh conducted a paired study between immediate and delayed feedback methods in an engineering course to determine which method was preferred for learning and which method students perceived to be better (61). This feedback was provided for homework assignments. The study indicated that what students perceived to be better was not in congruence with what was found for increased learning. Delayed feedback led to better performance on the final exam even though students reported that they preferred immediate feedback. It is important to note however, that though the feedback provided to the students was immediately available for the immediate feedback condition, students typically did not access the feedback until an average of 4-5 days later; whereas delayed access elapsed an average of 14 days. These conditions have been shown to allow for spaced studying, which can lead to larger learning gains (62). In a study that included human biology students, Fyfe, Meyer, Fyfe, Ziman, Sanders, and Hill found that demographic factors (age, gender, study experience, status, language spoken) or anticipated grade also greatly contribute to student perception about the usefulness of feedback (63). The authors found that students prefer immediate feedback, but the preference for the construct of the feedback, such as whether it is

corrective or personalized to the type of answer provided, depends greatly on the background of the students. This combination of influences indicates that careful and deliberate construction of multiple-choice exams should be followed with constructive and useful feedback if facilitation of additional learning is desired.

One last consideration regarding feedback is the possibility of increased test anxiety if feedback is given during the testing process, as is the case with use of the IF-AT® form. DiBattista and Gosse studied the relationship between students' reactions to using the IF-AT® form for their first major test in an introductory research design and statistics course and their levels of test anxiety and trait anxiety (64). Similar to some previous studies, they found that test anxiety and test performance were inversely related. However, students' preference for the IF-AT® form over traditional bubble forms was not related to test anxiety, test performance, or other demographic variables. They also found that immediate feedback actually reduced the test-related anxiety for the majority of students. They proposed that this reduction could be from the fact that anxiety is reduced when students respond correctly to an initial attempt (something that will likely happen more often than not during a test) and from a "gaming" affect produced by scratching off the coating on the form. Lastly, there was no indication that students with higher levels of test anxiety and poorer performance found the method of feedback to be anxiety provoking.

## **Recent STEM Testing Studies on Feedback**

Throughout this discussion on studies examining feedback, one particularly important theme emerged: the importance of designing studies where the inferences about feedback are clear. It is important to extricate the type of feedback from the timing of feedback and what contributes to subsequent student performance. Finally, because these studies can have real implications for student performance on summative exams that contribute to a student's course grade, they must also be designed so that potential harm is minimized. Considering the results based on the use of practice exams in chemistry (18), there is potential that students who take a practice exam may have reduced performance on subsequent exams. Therefore, investigators must also consider student perceptions of feedback on practice exams and effects on student studying or additional learning in preparation for final exams.

As one example of experimental design to separate the type of feedback from the delay in feedback, a study was developed that examined feedback timing using a repeat testing model to study the effect on students' additional learning in preparation for summative exams in chemistry. All components of this study were conducted in the final two to three weeks of the first semester of a two semester sequence of general chemistry. Testing took place during a laboratory period or at an additional time scheduled outside of class time. In some cases, more than one testing time option was given to students to accommodate schedules. A general schematic of the phases of the study and the timing of the feedback is shown in Figure 1.

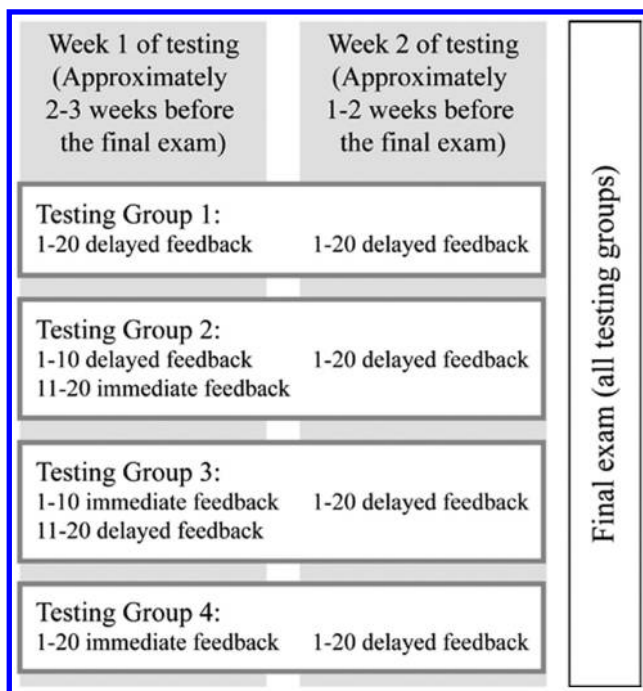


Figure 1. Schematic of testing phases of the study (designated by groups).

As stated previously, test construction is vital to any investigations of feedback as the judgments associated with the efficacy of the feedback are based on test performance. Therefore, the tests utilized in this study were developed using the majority of the standard test development procedures used by the Examinations Institute of the American Chemical Society, Division of Chemical Education (65). The test was constructed with 20 items in four specific content areas and a general design of pairing items based on either algorithmic or conceptual constructs (66). A second test of cloned items was developed with the same parameters, with the items in the same order but with the answer choices in different orders. Therefore, the two tests could be used in either order as repeat tests. Validity and reliability checks were integrated into the study prior to the implementation of the feedback testing of the study (67). Validity measures that included item analysis by classic test theory and expert opinions from multiple chemical educators were utilized to refine items during the trial testing phase. External validity checks were included to examine correlations between student performance and expert assigned complexity (68), other content exams, and student-reported confidence. Reliability was established both internally on one test using a KR-21 index and between the two tests.

The study was specifically designed to examine the immediate timing of feedback. Delayed feedback was provided through posting of performance in subsections and overall performance on a course management system. Based on the earlier discussions, this type of delayed feedback would be considered

non- corrective, and it would be expected to have minimal effects on student performance. An example of this report is shown in Figure 2. Because the access to this feedback required monitoring (to differentiate between students who saw the report and those who did not), it was posted on the course management system as a document where student access to the document is disclosed to the researchers. In order to preserve the confidentiality of the students, a code for each student was generated by the instructor and posted in each student's grade book. The process by which this report was generated was automated by the researchers such that the instructor only had to copy the student responses and practice exam codes into the spreadsheet and the report was generated in the form shown. Additionally, students were surveyed following the second practice exam about their actions based on their report. This process was approved by institutional review boards at each participating university.

Practice Exam Code	Total Score (out of 20)	Topic: Chemical Composition & Formulas (out of 6)	Topic: Gram, Mole, and Molecule Conversions (out of 6)	Topic: Reaction Stoichiometry (out of 4)	Topic: Limiting Reactant Stoichiometry (out of 4)
Univ1_S14_001	18	6	6	4	2
Univ1_S14_002	13	5	5	2	1
Univ1_S14_003	19	5	6	4	4
Univ1_S14_004	11	3	2	3	3

Figure 2. Example of posted student report.

Immediate feedback was provided using the IF-AT® form, both in split-halves testing and full form testing. For the split-halves testing, the students were split into two groups where they completed either 1-10 or 11-20 of the test items using the IF-AT® form and the remainder of the items were answered using a standard bubble form only. For all IF-AT® testing, students completed both the IF-AT® form and their bubble form to capture their first response (on the bubble form) as well as subsequent responses (on the IF-AT® form). Like the delayed feedback group, the immediate feedback group also had access to a score report after the first test. Immediate feedback was only provided for the first of the two tests with a standard delayed feedback test for all second tests regardless of test one group.

Finally, all students also completed a prompt for confidence immediately following each item. To assist the students in managing the mechanics of entering all of this information, bubble forms were designed to integrate all responses from students with the addition of a space on which to place the IF- AT® form. This is shown in Figure 3. These forms were analyzed using a scanning/grading software system with optical mark recognition (OMR). Student performance on the tests and by specific content area was collected. Additionally, a measurement of confidence in test response was collected for all students for all test items. For the students in the groups of immediate feedback, subsequent responses to items were also collected using the IF-AT® form. Student responses to survey items

about using the IF-AT® form and the use of their reports were also collected. Finally, the access of the online report was collected, noting the time at which feedback was initially accessed. In addition, demographic data as well as exam performance and standardized test performance data, such as ACT or SAT scores, was collected for students from institutional research data (per the permission granted through the approved IRB process). Upon completion, this study design will provide researchers with multiple measurements based on timing of feedback.

Place your "Immediate Feedback Assessment Technique (IFAT) form here

	First Answer	Confidence:
1.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
2.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
3.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
4.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
5.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
6.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
7.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
8.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
9.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
10.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident
11.	<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Not at all confident <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Highly confident

Figure 3. Example of student response sheet including prompt for confidence.

Along with the chemistry testing study presented, others have looked at ways to connect this literature to improve student learning in STEM classrooms. In 2013, Slepko published an article on utilizing the IF-AT® forms with integrated testlets for introductory physics course exams (69). A typical testlet contains a physics diagram or description of a physical scenario followed by several independent questions. The integrated testlets contained questions that were sometimes inter-dependent sequentially allowing for transference of knowledge from prior questions. This was accomplished by knowing that students had eventually discovered the correct answers to the previous questions with the IF-AT® form. They suggest that the integrated construction allows for a multiple-choice exam to function like a constructed-response exam with the benefit of immediate corrective feedback.

In 2014, Butler, Marsh, Slavinsky, and Baraniuk published an article on a classroom instructional design for an upper-level electrical and computer engineering course that specifically targeted best practices, including repeated

retrieval practice, spacing, and feedback on electronic homework problem sets (70). Repeated retrieval practice and spacing were achieved by including problems on week two and week three assignments that were included on the assignment from week one. This process was repeated with each consecutive assignment, allowing for new and old material to be practiced. Students entered an open-response answer and then were prompted to input a multiple-choice answer for the same item. The offering of a multiple-choice option made it possible to provide immediate corrective feedback after the assignment was due. Students were required to view each item feedback for credit. These changes to the homework benefited student performance on the midterm and final exam for the course. These three examples illustrate future research directions to connecting testing and feedback literature with improving student learning in STEM education.

### **Concluding Remarks**

Tests should be an important component in evaluating instructional and course design practices. However, they also have the potential to serve as a learning tool for students particularly when appropriate feedback is provided and utilized in a meaningful way. Several key characteristics of feedback were identified through this literature survey. Identifying the correct answer as opposed to simply stating right/wrong is more beneficial for enhancing student performance on repeat testing. Offering more elaborative feedback has had mixed results. The optimal timing of the feedback is still debated but it is clear that students need to engage actively with feedback for it to be effective. The laboratory research supporting delayed feedback illustrates that spaced retrieval of information can be very effective for improving student performance on future tests; however, the classroom research supporting immediate feedback illustrates the difficulty of getting active engagement in delayed feedback. This is perhaps exemplified by the preference of students for immediate over delayed feedback. Future feedback studies utilizing STEM classroom relevant materials and testing styles would add value to this literature base. In addition, it would be important to go beyond looking for complete correctness upon repeat testing and to look for partial incorporation of feedback, particularly as the complexity of some STEM relevant test items increases.

### **Acknowledgments**

This material is based upon work supported in part by the National Science Foundation under Grant No. (DUE-1140914). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are also grateful to many faculty and students who contributed to the chemistry testing study. Last we would like to acknowledge the advice and evaluation insights from Christopher Bauer, who served as the project external evaluator.

## References

1. Gibbs, G.; Simpson, C. *Learn. Teach. Higher Educ.* **2004-2005**, *1*, 3–31.
2. Sadler, D. R. *J. Higher Educ.* **1983**, *54*, 60–79.
3. Slavin, R. E. *Educ. Psychol.* **1978**, *13*, 97–100.
4. Vatale, M. R.; Romance, N. R.; Dolan, M. F. In *Assessment in Science Practical Experiences and Education Research*; McMahon, M., Simmons, P., Sommers, R., DeBaets, D., Crawley, F., Eds; NSTA Press: Arlington, VA, 2006; pp 1–14.
5. McKeachie, W. J. *Teaching Tips: Strategies, Research, and Theory for College and University Teachers*, 10th ed.; Houghton Mifflin: Boston, MA, 1999.
6. Henriques, L. Colburn, A.; Ritz, W. C. In *Assessment in Science Practical Experiences and Education Research*; McMahon, M., Simmons, P., Sommers, R., DeBaets, D., Crawley, F., Eds; NSTA Press: Arlington, VA, 2006; pp 15–30.
7. Scouller, K. *Higher Educ.* **1998**, *35*, 453–472.
8. Zoller, U.; Lubezky, A.; Nakhleh, M. B.; Tessier, B.; Dori, Y. J. *J. Chem. Educ.* **1995**, *72*, 987–989.
9. Barnett-Foster, D.; Nagy, P. *Higher Educ.* **1996**, *32*, 177–198.
10. Cracolice, M. S.; Deming, J. C.; Ehlert, B. *J. Chem. Educ.* **2008**, *85*, 873–878.
11. Dempster, F. N. *Educ. Psychol.* **1987**, *22*, 1–21.
12. Rawson, K. A.; Dunlosky, J.; Sciartelli, S. M. *Educ. Psychol. Rev.* **2013**, *25*, 523–548.
13. Crooks, T. J. *Rev. Educ. Res.* **1988**, *58*, 438–481.
14. Bangert-Drowns, R. L.; Kulik, J. A.; Kulik, C. C. *J. Educ. Res.* **1991**, *85*, 89–99.
15. Maki, R. H.; Serra, M. *J. Educ. Psychol.* **1992**, *84*, 200–210.
16. Crisp, V.; Sweiry, E.; Ahmed, A.; Pollitt, A. *Educ. Res.* **2008**, *50*, 95–115.
17. Oliver, R.; Williams, R. L. *J. Behav. Educ.* **2005**, *14*, 141–152.
18. Knaus, K. J.; Murphy, K. L.; Holme, T. A. *J. Chem. Ed.* **2009**, *86*, 827–832.
19. Butler, A. C. *J. Exp. Psychol.* **2010**, *36*, 1118–1133.
20. Wiggins, G. P. *Assessing Student Performance Exploring the Purpose and Limits of Testing*; Jossey-Bass Publishers: San Francisco, CA, 1993.
21. Butler, A. C.; Karpicke, J. D.; Roediger, H. L., III *J. Exp. Psychol.* **2008**, *34*, 918–928.
22. Kulhavy, R. W.; Anderson, R. C. *J. Educ. Psychol.* **1972**, *63*, 505–512.
23. Angelo, T. A.; Cross, K. P. *Class Assessment Techniques: A Handbook for College Teachers*, 2nd ed.; Jossey-Bass: San Francisco, CA, 1993.
24. Handley, K.; Williams, L. *Assess. Eval. Higher Educ.* **2011**, *36*, 95–108.
25. Bangert-Drowns, R. L.; Kulik, C. C.; Kulik, J. A.; Morgan, M. *Rev. Educ. Res.* **1991**, *61*, 213–238.
26. Dempster, F. N. *Educ. Psychol. Rev.* **1989**, *1*, 309–330.
27. Butler, A. C.; Marsh, E. J.; Goode, M. K.; Roediger, H. L., III *Appl. Cognit. Psychol.* **2006**, *20*, 941–956.

28. Roediger, H. L., III; Marsh, E. J. *J. Exp. Psychol.: Learn., Mem., Cognit.* **2005**, *31*, 1155–1159.
29. Butler, A. C.; Roediger, H. L., III *Mem. Cognit.* **2008**, *36*, 604–616.
30. Pashler, H.; Cepeda, N. J.; Wixted, J. T.; Rohrer, D. *J. Exp. Psychol.: Learn., Mem., Cognit.* **2005**, *31*, 3–8.
31. Marsh, E. J.; Lozito, J. P.; Umanath, S.; Bjork, E. L.; Bjork, R. A. *Memory* **2012**, *20*, 645–653.
32. Hancock, T. E.; Stock, W. A.; Kulhavy, R. W. *Bull. Psychon. Soc.* **1992**, *30*, 173–176.
33. Lhyle, K. G.; Kulhavy, R. W. *J. Educ. Psychol.* **1987**, *79*, 320–322.
34. Pressey, S. L. *Sch. Soc.* **1926**, *23*, 373–376.
35. Butler, A. C.; Godbole, N.; March, E. J. *J. Educ. Psychol.* **2013**, *105*, 290–298.
36. Kulik, J. A.; Kulik, C. C. *Rev. Educ. Res.* **1988**, *58*, 79–97.
37. Skinner, B. F. *Harv. Educ. Rev.* **1954**, *24*, 86–97.
38. Epstein, M. L.; Brosvic, G. M.; Costner, K. L.; Dihoff, R. E.; Lazarus, A. D. *Psychol. Rec.* **2003**, *53*, 177–195.
39. Robin, A. L. *J. Pers. Instr.* **1978**, *3*, 81–87.
40. Kulhavy, R. W.; Stock, W. A. *Educ. Psychol. Rev.* **1989**, *1*, 279–308.
41. Butterfield, B.; Metcalfe, J. *J. Exp. Psychol.* **2001**, *27*, 1491–1494.
42. Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: cognitive domain*; Bloom, B. S., Ed.; David McKay: New York, 1956.
43. Epstein, M. L.; Epstein, B. B.; Brosvic, G. M. *Psychol. Rep.* **2001**, *88*, 889–894.
44. DiBattista, D. *Can. J. High. Educ.* **2005**, *35*, 111–131.
45. Anderson, J. R. *Cognit. Psychol.* **1974**, *6*, 451–474.
46. Epstein, M. L.; Lazarus, A. D.; Calvano, T. B.; Mathews, K. A.; Hendel, R. A.; Epstein, B. B.; Brosvic, G. M. *Psychol. Rec.* **2002**, *52*, 187–201.
47. Dihoff, R. E.; Brosvic, G. M.; Epstein, M. L. *Psychol. Rec.* **2003**, *53*, 533–548.
48. Dihoff, R. E.; Brosvic, G. M.; Epstein, M. L.; Cook, M. J. *Psychol. Rec.* **2004**, *54*, 207–231.
49. Brosvic, G. M.; Epstein, M. L.; Cook, M. J.; Dihoff, R. E. *Psychol. Rec.* **2005**, *55*, 401–418.
50. Brosvic, G. M.; Epstein, M. L. *Psychol. Rec.* **2007**, *57*, 391–408.
51. Persky, A. M.; Pollack, G. M. *Am. J. Pharm. Educ.* **2008**, *72*, 1–7.
52. Butler, A. C.; Karpicke, J. D.; Roediger, H. L., III *J. Exp. Psychol.* **2007**, *13*, 273–281.
53. Retna K. S.; Chong E.; Cavana R. Y. Preliminary analysis of students' perceptions of feedback in a New Zealand University; retrieved from <http://www.oecd.org/edu/imhe/43974843.pdf> (The Organisation for Economic Co-operation and Development).
54. Struyven, K.; Dochy, F.; Janssens, S. *Assess. Eval. Higher Ed.* **2005**, *30*, 331–347.
55. Kang, S. H. K.; McDermott, K. B.; Roediger, H. L., III *Eur. J. Cognit. Recognit.* **2007**, *19*, 528–558.



56. Lee-Sammons, W. H.; Wollen, K. E. *Behav. Res. Methods, Instrum. Comput.* **1989**, *21*, 189–194.
57. FajfarCampitelli, P., G.; Labollita, M. *Aust. J. Psychol.* **2012**, *64*, 169–177.
58. Kettle, K.; Häubl, G. *Psychol. Sci.* **2010**, *21*, 545–547.
59. Zhao, Q.; Zhang, J.; Vance, K. *Learn. Individ. Differ.* **2013**, *23*, 168–171.
60. Metcalfe, J.; Miele, D. B. *J. Appl. Res. Mem. Cognit.* **2014** in press.
61. Mullet, H. G.; Butler, A. C.; Verdin, B.; von Borries, R.; Marsh, E. J. *J. Appl. Res. Mem. Cognit.* **2014** in press.
62. Kornell, N. *Appl. Cognit. Psychol.* **2009**, *23*, 1297–1317.
63. Fyfe, G.; Meyer, J.; Fyfe, S.; Ziman, M.; Sanders, K.; Hill, J. Presented at the Proceedings of the 35th International Association for Educational Assessment Annual Conference, Cambridge, United Kingdom, Sept. 2008.
64. DiBattista, D.; Gosse, L. *J. Exp. Educ.* **2006**, *74*, 311–327.
65. Holme, T. *J. Chem. Educ.* **2003**, *80*, 594–597.
66. Holme, T.; Murphy, K. *J. Chem. Educ.* **2011**, *88*, 1217–1222.
67. Arjoon, J. A.; Xu, X. Y.; Lewis, J. E. *J. Chem. Educ.* **2013**, *90*, 536–545.
68. Knaus, K. J.; Murphy, K. L.; Blecking, A.; Holme, T. A. *J. Chem. Educ.* **2011**, *88*, 554–560.
69. Slepков, A. D. *Am. J. Phys.* **2013**, *81*, 782–791.
70. Butler, A. C.; Marsh, E. J.; Slavinsky, J. P.; Baraniuk, R. G. *Educ. Psychol. Rev.* **2014**, *26*, 331–340.

## Chapter 7

# Exploring the Apparent Motivational Impact of Resurrection Points from Final Exam Performance

Jeffrey R. Raker<sup>1</sup> and Thomas A. Holme<sup>\*,2</sup>

<sup>1</sup>Department of Chemistry Center for the Improvement of Teaching and Research in Undergraduate STEM Education, University of South Florida, 4202 East Fowler Avenue, Tampa, Florida 33620

<sup>2</sup>Department of Chemistry, Iowa State University, 0213 Gilman Hall, Ames, Iowa 50011

\*E-mail: [taholme@iastate.edu](mailto:taholme@iastate.edu).

This chapter examines the motivational impact of resurrection points; a systematic method for encouraging students to earn back missed points on semester exams through performance on the final exam. This method for utilizing hour exams as a type of formative assessment was explored for three courses across three years. Four of the nine individual courses sampled offered resurrection points. A student's grade trajectory based on semester exam performance was used to predict how well a student needed to perform on the final exam to receive a particular course grade. The need to over or under perform based on semester performance was then compared to whether the student actually earned that letter grade. Odds ratios suggest that students in resurrection point courses were more likely to earn a particular course letter grade if they needed to perform better on the final than they had on semester exams. This observation is consistent with an explanation of student behavior during final exams that effort in various courses is rationed based on the perceived value of the exam within the course it is administered.

Instructors are inclined to hope that students prepare equally and at their best level for every exam. Most students are taking more than one course and are continuously making effort decisions about their courses and coursework. This type of learning has been labeled strategic (1) and can be contrasted with categories of “deep” or “surface” learning (2, 3). To the extent that students use strategic approaches in learning, trade-offs in study time are particularly common during final exam periods when students encounter several, often high-stakes, tests. In some sense, for most students, and for those using strategic learning methods in particular, study time is apportioned by the student. Within this premise, the question becomes: Are there ways to garner more student attention during final exams?

This question is important because accurate assessments of learning can be confounded when the extent of student effort can skew measurements. A student who chooses to exert less effort to prepare for an exam in one course in order to save study time for another course affects the measurement in both courses. As a result, a well-constructed test may reflect what the student knows and is able to do at the time of the examination, while at the same time not accurately reflecting the student’s overall ability or proficiency in the material. Nonetheless, the final exam represents the last opportunity for a student to demonstrate understanding, so any chance to further learning in the course is arguably done at that point. It has been argued that some students easily pick up the deficient material when it becomes essential to their success in later courses such as physical chemistry (4). Still, it seems helpful to use every opportunity to help students learn, particularly material that they have found challenging during the course. Providing additional motivation to study for the final exam carries importance, couched as it is in a choice environment where it competes, in the mind of the student, with demands from all the other courses they are taking that semester.

## **Defining the Concept of Resurrection Points**

In Tobias and Raphael’s book *The Hidden Curriculum*, Herschbach describes resurrection points as a method for increasing student motivation for learning or relearning material for the final exam (5). With this pedagogical technique, the student has the opportunity to earn a maximum score on the final examination as well as raise all previous examination scores to the maximum score. If the student earns a higher percentage score on a portion of the final exam than they did on the semester exam that covered the same material, then “resurrection” points are earned. The number of points earned is such that the lower score on the semester exam is essentially replaced by the higher score on the final exam. Practically speaking, the points can be readily calculated in a spreadsheet using the equation:

$$\text{Points} = \text{Max} \left( 0, \left[ \left( \frac{\text{Final \%}}{100} \right) \times (\text{Hour exam max}) - (\text{Hour exam score}) \right] \right)$$

This calculation is repeated for all semester exams. A student, however, is not punished (i.e., a previous exam during the semester score is never lowered) for scoring lower on the given material on the final exam for a given semester exam than they did on the actual semester exam. As an example, consider a course with 3 (mid-term) hour exams each worth 100 points. Say a student scores 85 on exam 1, 68 on exam 2 and 76 on exam 3. The final exam has four sections, each worth 50 points. Section A corresponds to exam 1, B to exam 2, C to exam 3 and D tests material covered after the final hour exam. If the same student scored 41 on part A, 43 on part B and 39 on Part C the resulting resurrection points earned would be 20: 0 points for part A, 18 points for part B and 2 points for part C. Note that they do not receive negative points from part A even though their performance decreases relative to the relatively strong first exam. In terms of raw points, students only benefit from resurrection points. In this scheme, achieving a perfect score on the final exam is equivalent to achieving a perfect score on all semester examinations. The thought is that students will spend time addressing their learning deficiencies determined from their semester performance in preparation for the final examination. Moreover, if a student demonstrates proficiency at the end of the semester, then arguably the student has learned all the information and has earned a grade that reflects that learning.

The use of resurrection points can strike some instructors as unduly generous. It is possible for students to substantially increase a course grade. In over 20 years of implementation (by author TAH) in general chemistry courses the much more common level of grade impact is roughly 1/3 of a grade (e.g. from a B to a B+), and increases of two grades or more (e.g. from a C to an A) have occurred less than a dozen times for students who are replacing actual hour exam performances. Resurrection points also provide a convenient way to manage make-up exams. Excused hour exam absences can be made up via the resurrection points alone. For years one of us (TAH) used both make-up exams and resurrection points for missed exams, but analysis of performances revealed that students obtained resurrection points in over 90% of the cases where make-up exams were given. In other words, even when make-up exams are provided, students who miss exams are often behind (because of illness, for example) in several classes and their performance is less than ideal for the make-up. Using resurrection points alone for make-up exams is not only logistically facile for the instructor, it tends to ease stress for the student returning from an illness by requiring no make-up exam in one of their courses.

For these reasons, the resurrection points concept is potentially an important learning tool. At least in principle, it may provide added motivation for students to study for the final examination and thereby enhance net learning in a course. One aspect of a course that generally motivates student learning is testing (6), so the confluence of customary test-oriented motivation factors and resurrection points may influence studying. If this premise is true there may be measurable ways, based on test performance that can adjudicate the role of resurrection points on student learning. Of course there are other aspects to courses that potentially influence motivation and in chemistry these include relevancy, applications, and current research projects (7–9).

Beyond chemistry education, the role of motivation in learning has received considerable attention. One organizational theory of motivation proposed initially by Deci and Ryan (10) and recently refined (11) places the locus of motivation along a continuum. Thus, student motivation may range from no motivation (amotivation) through varying levels of extrinsic motivation to fully intrinsic motivation. Individuals who lie at different spots along this continuum will tend to be activated by different stimuli. While it might be desirable to have every student in a large classroom with strong intrinsic motivation, the reality that such a class is encountered is rather unlikely. Accordingly, the possibility that a tactic such as resurrection points on the final exam can trigger the extrinsic motivation categories represents a possible mode of action for improved performance in a course. In terms of chemistry specifically, one study found (12), via self-report survey work, that motivation of students in general chemistry tends to lag as the semester progresses, a factor that, if true, would seem to predict lower achievement on final exams. Similar work applied to motivation in organic chemistry (13) found that students with stronger intrinsic motivation factors tended to perform better in the course. Student motivation remains a widely studied construct in educational psychology beyond test-taking factors. For example, several recent studies have sought to parse origins of motivation in terms of several factors such as self efficacy (14), epistemological beliefs (15), extrovertedness, (16), and coping strategies (17). Student self-report motivation instruments have been devised (18) and validated (19) within science contexts. Studies that investigate the role of formative assessment on student motivation (20) and differences between on-line and classroom-based courses (21) have also been described recently.

Given the established importance of motivation in promoting student learning (22, 23), the question of whether or not resurrection points can affect motivation becomes important. At least one study has found evidence that student perception of the value of a test influences motivation for the test (24). Therefore, a potential proxy for understanding motivational factors associated with resurrection points lies in differences in student final exam performances based on the availability of resurrection points. Despite this interest, the ability to devise a quasi-experimental study to investigate the role of resurrection point availability is limited. Teachers who use this method are generally convinced of its utility and thus offering an opportunity to earn resurrection points to some students and others not is an unethical proposition. Given this constraint, the best possible method would be to compare similar courses at a single institution where resurrection points have been used in different ways or not used at all. This is the approach reported here. Thus, we consider connected issues related to the hypothesized effect of resurrection points, based on available empirical data. First, does the availability of resurrection points result in observable differences in student performance on final exams in courses that use them relative to similar courses that do not? Second, does changing when students are aware of the availability of resurrection points influence observable student performances on final exams? This latter question seeks to provide at least preliminary information about whether students “game the system” more if they know about resurrection points from the outset of a course.

Comparisons between final exam performances of students who have the opportunity to earn resurrection points on their final exam with those who do not have such an opportunity can provide insight into how their availability affects student study habits. Ultimately, resurrection points are tied to students improving their performance at the time of the final exam, an effect that represents “over performance” at least relative to test results during the rest of the course. If resurrection points provide measurable motivation, those students with access to them will have a higher odds-ratio of over performing on the final exam to earn a higher grade than those students who do not have the opportunity to earn resurrection points on their final exam. This hypothesis can be tested using logistic regressions of whether a student earns a particular grade (yes or no, binary data) versus a measure of student performance during the semester. Logistic regression has been described in a number of previous articles including from our group (25). Essentially this method provides a way to quantify the difference between the “grade trajectory” of a student during the semester and the ultimate grade after the final exam. If courses with resurrection points behave differently than those without them, this result would be consistent with the hypothesis that this teaching technique increases student motivation. To this end, binary logistic regression odds-ratios were calculated and compared for general chemistry courses where resurrection points were implemented and for courses where no resurrection points were implemented. Additional factors, such as when in the semester students are aware of the availability of resurrection points (from the start or mid-term) and between and within specific general chemistry course types (i.e., one semester course for engineering majors or 1<sup>st</sup> semester of a yearlong course for STEM majors) are also considered in the analysis presented here.

## **Analysis Methodology and Summary Statistics**

De-identified student performance records were obtained with Institutional Review Board approval for three general chemistry courses at Iowa State University across three academic years. One of these courses is a single-semester course for pre-engineering students that covers topics typically covered in both semesters of general chemistry and is titled “survey” throughout the presentation of data. The other two courses are the first- and second-semester of the traditional two-semester general chemistry course for science majors, and are labeled “1<sup>st</sup> Term”, and “2<sup>nd</sup> Term” respectively throughout data presentation. Demographic data were not collected for the students in the study. Table 1 summarizes the nine courses by course type, data collection semester, number of students, notation of when resurrection points were announced to students, number of semester exams, number of points for the final exam, and the total points for the course. For all courses, points from non-examination sources were less than 40% of the total points; therefore, a significant portion of a student’s grade was determined via semester and final exams.

Looking at the information in Table 1, it is clear that there are variations in how the courses were structured. While it is impossible to control for the potential variability associated with this feature of the courses, it is also important

to note that the analysis presented here has access to a large number of student performances. Roughly 1050 students studied had access to resurrection points that they knew about from the start of the course. Another ~1400 students had access to resurrection points, but were not aware of that fact until after the course drop deadline was passed. Finally, over 4400 students were from courses that did not offer resurrection points. This sample includes courses taught by several instructors, some of whom used resurrection points and some who do not.

**Table 1. Summary of General Chemistry Courses Included in the Study**

<i>Course Type</i>	<i>Date of course</i>	<i>N</i>	<i>Resurrect Points</i>	<i># of hour Exams</i>	<i>Final Pts. Avail</i>	<i>Total Pts. Avail</i>
Survey	F 2010	746	Aware from Beginning	4	200	830
Survey	F 2011	908	Aware from Beginning	3	200	800
Survey	F 2012	955	Aware after Drop Date	3	200	800
1 <sup>st</sup> Term	F 2010	902	(Not Awarded)	4	150	800
1 <sup>st</sup> Term	F 2011	1,041	(Not Awarded)	4	150	700
1 <sup>st</sup> Term	F 2012	1,155	(Not Awarded)	4	150	800
2 <sup>nd</sup> Term	S 2011	641	(Not Awarded)	3	150	700
2 <sup>nd</sup> Term	S 2012	697	(Not Awarded)	3	150	700
2 <sup>nd</sup> Term	S 2013	774	Aware after Drop Date	3	150	700

The premise of this analysis is that students will use external motivational factors during finals to apportion their time resources for study. As such, the “payoff” to the student is whether or not a desired, higher grade is obtained. To the extent that this is a measurable goal, it is also inherently binary, either students do, or do not, achieve the higher grade – i.e., change their grade trajectory. To determine if there was a change in grade trajectory at the time of the final exam, for each student, the total number of points prior to the final examination was calculated; this value was utilized in determining how many points the student would then need to earn on the final exam to receive an A (90%), B (80%), C (70%), D (60%), or F (< 60%) for the course. Because the point value for the final

exam differed by course, an average percent needed on the final exam to receive a letter grade was calculated. (Remember, students in resurrection point courses could earn more than the available number of points on the final because they could earn the maximum point value of the final exam and any points missed on semester exams.) It was then determined which letter grades were possible for each student to earn by asking: were enough points available on the final exam for a student to receive that letter grade? For example, in courses that did not offer resurrection points, it was impossible for some students to earn enough points to receive an A or a B; in addition, it was possible for some students to receive 0 points on the final exam and not receive lower than a C for the course. Table 2 summarizes the number of students that could have possibly earned each letter grade in the courses studied. Note that for the engineering “survey” courses and the 2013 2<sup>nd</sup> STEM courses (i.e., resurrection point courses) that a rather large number of students (upwards of 90% of the students) could, in principle, earn an A or B letter grade compared to non-resurrection point courses (around 35% of the students); this observation is to be expected because resurrection point courses allow students to gain back points “lost” during semester exams and thereby present the possibility, if not the probability, that they can get top grades in the course regardless of prior test performance.

**Table 2. Number of Student Able to Earn Each Final Letter Grade in the Courses Analyzed<sup>a</sup>**

<i>Course Type / Date</i>	<i>N</i>	<i>N able to earn “A”</i>	<i>N able to earn “B”</i>	<i>N able to earn “C”</i>	<i>N able to earn “D”</i>
Survey / F10	746	717	494	239	73
Survey / F11	908	801	709	404	162
Survey / F12	955	894	627	281	101
1 <sup>st</sup> Term / F10	902	371	528	319	126
1 <sup>st</sup> Term / F11	1,041	653	552	310	114
1 <sup>st</sup> Term / F12	1,155	529	614	418	199
2 <sup>nd</sup> Term / S11	641	169	378	360	208
2 <sup>nd</sup> Term / S12	697	320	408	312	140
2 <sup>nd</sup> Term / S13	774	677	538	277	115

<sup>a</sup> Note in all tables, “Survey” denotes the 1-semester general chemistry course for engineering students; “1<sup>st</sup> Term” denotes the first semester of a two-semester general chemistry course and “2<sup>nd</sup> term” denotes the second semester of that course.

Next a calculation was devised to estimate the grade trajectory for each student. Thus, the difference between the final exam percentage score needed to earn each possible grade and their average percentage performance on semester



exams was determined. Values for these differences ranged from -99.83% to +73.26% as summarized in Table 3 by possible letter grade. Students who were unable to earn a particular grade were not included in that grade minimum and maximum determination. A negative value is interpreted as the number of percentage points a student could *under perform* relative to their average semester exam performance and still earn that letter grade. A positive value, then, is interpreted as the number of percentage points a student needs to *over perform* relative to their average semester exam performance and earn that letter grade.

**Table 3. Minimum and Maximum Percentage Points Needed on the Final Exam to Earn Each Final Letter Grade**

<i>Course type / Date</i>		<i>To get A</i>	<i>To get B</i>	<i>To get C</i>	<i>To get D</i>
Survey / F10	Min	-46.5	-51.7	-61.0	-61.6
	Max	73.3	61.8	52.1	33.6
Survey / F11	Min	-39.1	-54.4	-59.4	-60.3
	Max	68.4	64.4	54.4	34.4
Survey / F12	Min	-43.9	-54.3	-68.2	-54.7
	Max	57.9	58.9	54.7	33.1
1 <sup>st</sup> Term / F10	Min	-61.5	-80.9	-74.5	-75.0
	Max	20.9	36.8	49.3	47.1
1 <sup>st</sup> Term / F11	Min	-83.3	-90.8	-70.1	-61.5
	Max	30.9	45.8	52.3	72.7
1 <sup>st</sup> Term / F12	Min	-56.8	-77.1	-82.5	-73.8
	Max	23.8	41.1	53.4	67.7
2 <sup>nd</sup> Term / S11	Min	-49.7	-75.0	-78.8	-83.2
	Max	21.3	41.5	47.1	55.9
2 <sup>nd</sup> Term / S12	Min	-58.1	-70.2	-76.7	-70.0
	Max	23.3	39.3	56.6	49.0
2 <sup>nd</sup> Term / S13	Min	-58.3	-72.7	-87.3	-74.9
	Max	51.9	58.7	63.3	44.2

Looking more closely at Table 3 reveals the nature of the binary judgment made in this study. Because of the structure of resurrection point courses, students in them have potential access to higher letter grades than students have in non-resurrection point courses. To do so, however, they must dramatically over perform on the final exam compared to their hour exam average percentage. This

fact can be seen by comparing two different courses. For example, in the “survey” course in Fall 2010, at least one student had the chance to get an A by having a final exam 73.26 percentage points higher than their hour exam performance. By contrast, in the non-resurrection point 2<sup>nd</sup>-term class in Spring 2011, to obtain an A, the largest gap would have been 21.27 percentage points higher. Few, if any, of these students in fact do obtain the A in either course, but the availability of the higher grade is an important factor to consider when looking at the analysis that follows.

Finally, for each student, it was determined what letter grade would be assigned based on the percentage of points the student earned using cutoffs of A (90%), B (80%), C (70%), D (60%), or F (< 60%). Grade distributions were fairly similar among these nine courses, except for the Fall 2011 2<sup>nd</sup> semester STEM course for which the grade distribution was generally lower. Note that this choice of assigning achieved grade removes the possibility of adjustments made by instructors, something that may occur when students fall very close to a grade borderline. Because students are not able to predict what borderline adjustments might be, it is safe to expect that student effort in preparing for the final would have been based off the cutoff scale that is in the course syllabus, which is commensurate with the values used here.

## Results and Discussion

To estimate the impact of resurrection points on the grade trajectory of students, there are several pertinent pieces of information presented here: (1) the number of students receiving resurrection points in the three classes implementing this motivational technique, (2) the number of students under and over performing to receive a particular grade, (3) a graphical representation of binary demarcation of student effort necessary on the final examination versus under or over performing, and (4) the odds of not receiving a particular grade if needing to over perform (the odds of receiving a particular grade would be more appropriate; however, the odds are all less than one and less accessible for interpretation).

Within the four courses implementing resurrection points, students received varying numbers of resurrection points (see Table 4). For this discussion, resurrection points are defined as the number of points added to the previous exam performance to raise their previous exam performance to reflect their performance on the respective exam material on the final exam. For the three “survey” courses for engineers, over 96% of students received some amount of resurrection points with 30 or more points being earned on average (roughly 5% of the total points for the course). The number of resurrection points *available* is determined by the performance on semester exams. On average, the survey courses for engineers scored less than 70% on semester exams. So, while the amount of points available varies by student, in total the class had a considerable opportunity to earn resurrection points. The 2<sup>nd</sup> semester general chemistry course for STEM majors only had about a third of the students receiving resurrection points with an approximate average of 4 points. Considering that the semester

exam average was 82% for this course, the students had less of an opportunity to earn resurrection points than did the students in the “survey” course for engineers. It is also possible that students in the “survey” course for engineers, being aware of the possibility of resurrection points gave *less time* to their chemistry studies during the hour exams, effectively under performing then.

**Table 4. Summary of Resurrection Points Earned in Each General Chemistry Course**

<i>Course Type / Date</i>	<i>N</i>	<i>% Students Earning Resurrection Points</i>	<i>Average Number of Resurrection Points Earned (SD)</i>
Survey / F10	746	99.6	44.0 (23.8)
Survey / F11	908	97.3	34.0 (22.6)
Survey / F12	955	96.3	30.3 (19.9)
1 <sup>st</sup> Term / F10	902	n/a	n/a
1 <sup>st</sup> Term / F11	1,041	n/a	n/a
1 <sup>st</sup> Term / F12	1,155	n/a	n/a
2 <sup>nd</sup> Term / S11	641	n/a	n/a
2 <sup>nd</sup> Term / S12	697	n/a	n/a
2 <sup>nd</sup> Term / S13	774	37.3	3.97 (8.56)

A student’s need to under or over perform for a particular grade was determined as is presented in Table 5. The number of students receiving a particular grade (i.e. Got A, Got B, Got C, or Got D) by under or over performance is summarized in this table. By visual observation, it can be seen that students in the survey course for engineers more commonly over perform and receive a particular grade than students in either the 1<sup>st</sup> semester or 2<sup>nd</sup> semester general chemistry courses for STEM majors. The numbers are not, however, particularly large in any course.

A graphical depiction of these results emphasizes the binary nature of the analysis and is provided in Figures 1 and 2. The y-axis is marked as “1” receiving the grade or “0” not receiving the grade. The x-axis is the effort score, the difference between average performance during the semester and performance needed on the final examination to earn the particular grade. For ease of interpretation, a vertical red line marks 0 effort (i.e. average semester performance equal to performance needed on the final exam). Additionally, those needing to under perform are marked as blue dots; those needing to over perform are marked as red dots.

**Table 5. Number of Students Under or Over Performing on the Final Exam for Each Final Letter Grade**

<i>Course Type / Date</i>	<i>N</i>	<i>n Students Under Performed and</i>				<i>n Students Over Performed and</i>			
		<i>Got A</i>	<i>Got B</i>	<i>Got C</i>	<i>Got D</i>	<i>Got A</i>	<i>Got B</i>	<i>Got C</i>	<i>Got D</i>
Survey / F10	746	231	246	162	41	19	11	4	0
Survey / F11	908	159	288	230	99	27	30	12	8
Survey / F12	955	284	332	170	66	35	23	10	0
1 <sup>st</sup> Term / F10	902	209	327	227	91	2	4	1	4
1 <sup>st</sup> Term / F11	1,041	365	327	217	83	1	1	1	3
1 <sup>st</sup> Term / F12	1,155	299	386	259	152	7	0	0	1
2 <sup>nd</sup> Term / S11	641	49	190	176	169	1	0	3	2
2 <sup>nd</sup> Term / S12	697	145	220	186	111	0	0	3	1
2 <sup>nd</sup> Term / S13	774	221	274	162	79	0	1	1	0

To quickly assess the information contained in these graphs, the red dots in the upper right position represent students who obtain the higher grade by over performing on the final exam. Blue dots on the upper line obtain the higher grade, but were able to do so without over performing, on average, on the final exam. The lower line plots students who did not receive the higher grade. The graphical representation (i.e., Figures 1 & 2) readily show that many more students over perform and receive the grade (i.e., red dots on the upper line) in the survey courses for engineers than the other courses. Nonetheless, even in these courses, the number of students who over perform and receive the higher grade is smaller than those who over perform and do not (red dots on the lower line.) This evidence suggests that resurrection points are not inherently over-generous to students. Another value of these visual representations is that one can observe that many of the over performers were relatively close to the 0 effort mark (i.e., performance needed on the final examination was close to equivalent to performance on semester exams). In addition, looking at these graphs for the different grades, the majority of over performance resulting in a high grade occurs for grades of A or B. Relatively few students over perform to receive a C, for example, even when resurrection points are available. The idea that students who struggle with the material may gain less in measured performance from the use of this motivation tool is consistent with previous studies (12).

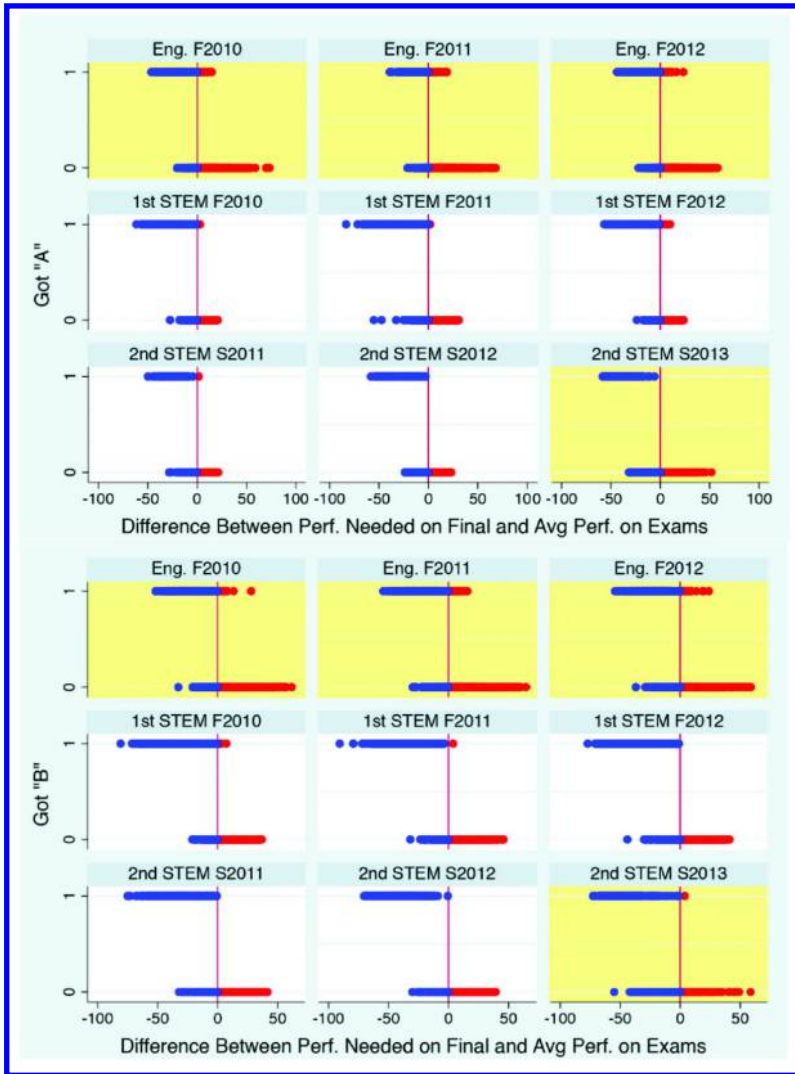


Figure 1. Graphical depiction of instances of under performance (blue dots) and over performance (red dots) as a function of course and grade level for grades of “A” and “B”. Courses with a yellow background included resurrection points. For vertical axis, 0 means the grade depicted as not achieved and 1 means the grade was achieved.

As can be inferred from the numbers in Table 5, an attempt to *quantify* the possibility of over performance – the odds of over performing and receiving the particular grade – leads to a value less than one for all grades and all courses. Comparison of these numbers is possible, but the inverse ratio provides the same

information in a more readily digestible form. Thus, the odds of “not receiving the grade when the student must over perform on the final examination to receive it” are reported in Table 6. Across all nine courses, odds range from 17.22 to 817.50; odds ratios are only reported only if they are statistically significant ( $p < 0.05$ ).

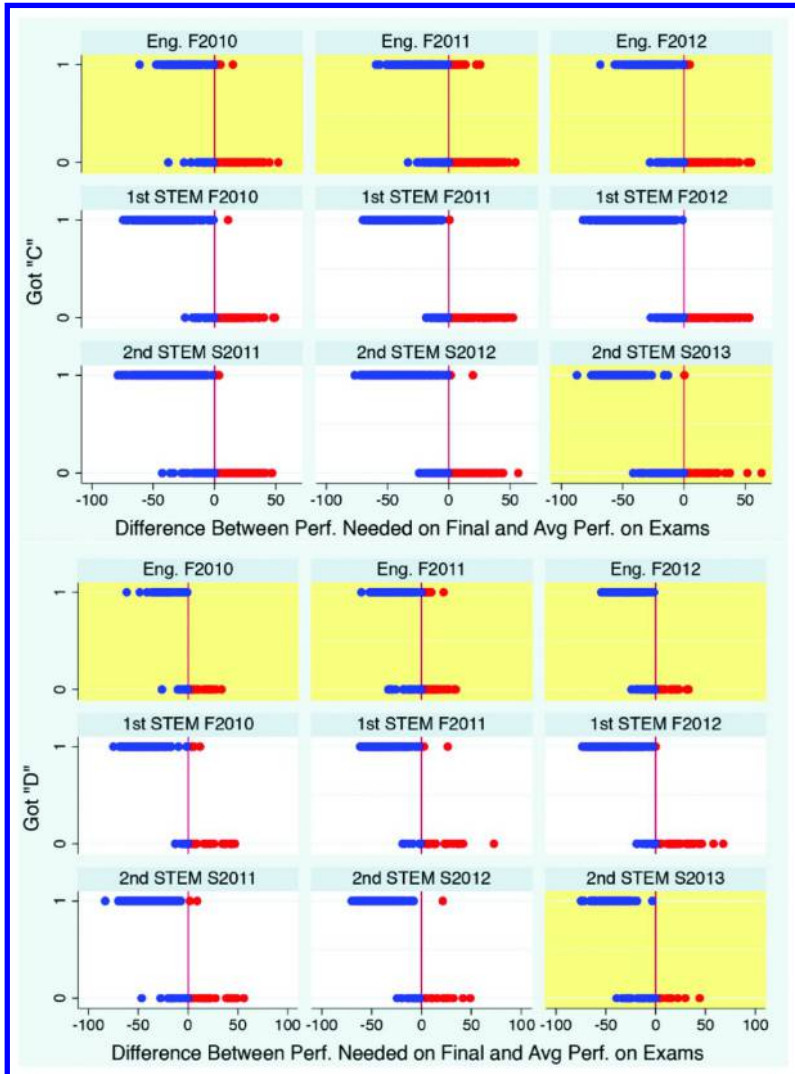


Figure 2. Graphical depiction of instances of under performance (blue dots) and over performance (red dots) as a function of course and grade level for grades of “C” and “D”. Courses with a yellow background included resurrection points.

**Table 6. Odds of Not Getting a Final Letter Grade if Needed to Over Perform on the Final Exam**

<i>Course Type / Date</i>	<i>N</i>	<i>Odds of “Not Getting” the Grade if Needed to Over Perform (p &lt; 0.05 for reported odds)</i>			
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
Survey / F10	635	44.6	39.3	88.0	*
Survey / F11	908	33.9	25.2	30.9	17.2
Survey / F12	955	33.2	24.4	24.9	*
1 <sup>st</sup> Term / F10	902	205.	240.	712.	65.4
1 <sup>st</sup> Term / F11	1,040	715.	817.	581.	69.2
1 <sup>st</sup> Term / F12	1,160	116.	*	*	431.
2 <sup>nd</sup> Term / S11	641	70.0	*	124.	139.
2 <sup>nd</sup> Term / S12	697	*	*	124.	96.2
2 <sup>nd</sup> Term / S13	774	*	140.	55.3	*

These large numbers indicate that the odds are generally against, often strongly against, observing student over performance on a final exam to obtain a higher grade. In some courses the odds against over performing enough to earn a higher grade may be greater than 500 to 1. Many experienced instructors will see this data as confirmation for more anecdotal observations of students who convince themselves that they can “save” their grade via the final exam, often with unsuccessful results. Nonetheless, despite the overall large numbers present, there are important differences between courses that include resurrection points and those that do not. The largest odds against over performance are prevalent in the courses that do not provide the opportunity for resurrection points. In the resurrection points courses the odds tend to still be smaller, an average of 46 to 1 against making the higher grade, but less daunting. It is impossible to argue without qualitative, interview-style data whether students consciously choose an effort level based on their prospects for achieving a desired grade. Neither is it possible to adjudicate if students make accurate judgments about such prospects. Nonetheless, the difference in odds ratios obtained in the analysis summarized here is such that the difference in student performance is clear. The difference is not tied to individual instructors, as the courses studied have several instructors. It is not possible to prove that motivation associated with the availability of resurrection points are responsible for the difference, but the results noted suggest this as a plausible explanation.

This study did not have access to student performance in any other courses than chemistry, so it is not possible to determine the relative performance on the chemistry final compared to other topics taken by students in a given semester. Nevertheless, the overall message of this set of data appears to be that student

ability to over perform during finals is limited. At the same time, a technique like resurrection points seems to result in having occurrences of over performance become more common.

## Conclusions and Implications

The hypothesis for this study was that the perceived motivational interest in earning resurrection points may lead to enhanced student motivation to study for a chemistry final. Extra studying would, in turn, lead to students being more likely to over perform on the final exam compared to their average semester exam performance; in other words change their grade trajectory. The current study does not preclude the design of an ideal experimental or quasi-experimental study related to this question but such a study would require instructors who believe that resurrection points are useful to deny this method to some fraction of their students, which is ethically troublesome. The study presented here does benefit from having data collected in three different general chemistry courses across three years and thereby provides several comparative implementations of resurrection points. Specifically, one course did not offer resurrection points ever (i.e., 1<sup>st</sup> semester STEM), one course offered resurrection points only in the last year of the study (i.e., 2<sup>nd</sup> semester STEM), and one course offered resurrection points for each year (i.e., Survey General Chemistry for Engineering Students). In addition, the 2<sup>nd</sup> semester course for science majors had different implementation style for resurrection points with students first learning about resurrection points in the middle of the semester in the last year of the study. Overall, regardless of the course or specific implementation, numerical results of course performance suggest that students were more likely to over perform and receive a particular grade in a resurrection-point course than a non-resurrection-point course. This result is consistent with the argument that at least some fraction of the students were differently motivated in resurrection point courses to relearn material missed on semester exams and study for the final exam.

This study is limited in several ways: First, student effort in preparing for the final examination was not actually measured. Measures of such constructs routinely involve self-report data, and such data is difficult to calibrate among students, so direct measures across difference courses would be more difficult to obtain and use. The idea of using performance during the semester to establish a grade trajectory to which is compared to performance on the final examination relative to that trajectory is argued to be a proxy for motivation. This choice encompasses an assumption that higher performance on the final exam likely reflects extra study efforts. It is certainly true that other unmeasured factors such as the number of finals a student had on the day of the general chemistry final exam may have affected student performance as well. Personal influences such as medical issues or family emergencies also may have an impact on student performance. Finally, the nature of final exams themselves, as arbiters of content knowledge likely plays a role in the results presented here. Comprehensive finals present challenges to students, but many teachers would agree that individual items on such exams are often less complex than those asked during hour exams.



This change in structure is practically important because of the relative size of the knowledge domain covered on the final exam. Even with this change, it is noteworthy that student over performance is uncommon in the empirical data presented here. These, and other, confounding variables have not been considered and could be the pursuit of future work in evaluating the educational impact and worth of resurrection points or any other pedagogical strategy as a motivational tool.

Second, despite the sample including three courses across three years, a relatively large available sample (over 7,000 students), with necessary variations of resurrection points versus non-resurrection points, the number of trials is ultimately small. Furthermore, most instructors perceive that individual classes have something like a “personality”, and in some cases there is no obvious reason why one group of students seems to struggle with course materials more than the students in courses before or after them. There may be ways to control for some of the variables noted above, or at least measure them to investigate their potential impact relative to whether or not resurrection points are implemented in a class. Nonetheless, it is important to recognize that some instructors have used this method for over 20 years, and the ratio of student comments lauding resurrection points to those expressing concern (in venues such as course evaluations) is overwhelmingly towards the positive. Even if the odds of over performing are not particularly strong even with the availability of resurrection points, students tend to appreciate the opportunity to improve their grades in a systematic way.

## References

1. Entwistle, N. J.; Ramsen, P. *Understanding Student Learning*; Croom Helm: London, 1983.
2. Biggs, J. *Teaching for Quality Learning at University*; Open University Press: Buckingham, 2003.
3. Marton, F.; Saljo, R. *Br. J. Educ. Psychol.* **1976**, *46*, 4.
4. Rieck, D. F. *J. Chem. Educ.* **1998**, *75*, 850.
5. Herschbach, D. In *The Hidden Curriculum - Faculty-Made Tests in Science: Part I: Lower-Division Courses*; Tobias, S., Raphael, J., Eds.; Plenum Press: New York, 1997.
6. Ward, R. J.; Bodner, G. M. *J. Chem. Educ.* **1993**, *70*, 198.
7. Woodburn, J. H. *J. Chem. Educ.* **1977**, *54*, 763.
8. Foote, J. *J. Chem. Educ.* **1981**, *58*, 198.
9. Holme, T. *J. Chem. Educ.* **1994**, *71*, 919.
10. Deci, E. L.; Ryan, R. M. *Psychol. Inquiry* **2000**, *11*, 227.
11. Kusurkar, R. A.; Ten Cate, T. J.; Vos, C. M. P.; Westers, P.; Croiset, G. *Adv. Health Sci. Educ.* **2013**, *18*, 57.
12. Zusho, A.; Pintrich, P. R.; Coppola, B. P. *Int. J. Sci. Educ.* **2003**, *25*, 1081.
13. Lynch, D. J.; Trujillo, H. *Int. J. Sci. Math. Educ.* **2011**, *9*, 1351.
14. Prat-Sala, M.; Redford, P. *Br. J. Educ. Psychol.* **2010**, *80*, 283–305.
15. Cavallo, A. M. L.; Rozman, M.; Blinkenstaff, J.; Walker, N. *J. Coll. Sci. Teach.* **2003**, *33*, 18–23.

16. Liang, C.; Chang, C-C. *Learn. Individ. Differ.* **2014**, *31*, 36–42.
17. Moneta, G. B.; Spada, M. M. *Pers. Individ. Differ.* **2009**, *46*, 664–669.
18. Glynn, S. M.; Koballa, T. R., Jr. In *Handbook of College Science Teaching*; Mintzees, J. J., Leonard, W. H., Eds.; National Science Teachers Association Press: Arlington, VA, 2006; pp 25–32.
19. Glynn, S. M.; Taasoobshirazi, G.; Brickman, P. J. *Res. Sci. Teach.* **2009**, *46*, 127–146.
20. Yin, Y.; Shavelson, R. J.; Ayala, C. C.; Ruiz-Primo, M. A.; Brandon, P. R.; Furtak, E. M.; Tomita, M. K.; Young, D. B. *Appl. Meas. Educ.* **2008**, *21*, 335–359.
21. Yen, H. C.; Tuan, H. L.; Liao, C. H. *Res. Sci. Educ.* **2011**, *41*, 211–224.
22. Dweck, C. S. *Am. Psychol.* **1986**, *41*, 1040.
23. Elliott, E. S.; Dweck, C. S. *J. Pers. Soc. Psychol.* **1988**, *5*.
24. Hong, E.; Peng, Y. *Learn. Instr.* **2008**, *18*, 499.
25. Emenike, M.; Raker, J. R.; Holme, T. J. *Chem. Educ.* **2013**, *90*, 1130–1136.

## Chapter 8

# Use of Student Self-Assessment of Exams To Investigate Student Learning in Organic Chemistry Classes

Andrew G. Karatjas\*

Department of Chemistry, Southern Connecticut State University,  
501 Crescent Street, New Haven, Connecticut 06515, U.S.A.

\*E-mail: karatjasa2@southernct.edu.

Post examination self-assessment surveys were utilized to explore student performance on examinations in organic chemistry courses. This study of student self-perception looked at the application of the Kruger-Dunning effect in organic chemistry courses. The results include a comparison of student performance to expectations and the amount of time spent preparing. Results for poorer performing students indicate a lack of connectivity between perception and actual results.

### Introduction

When students perform poorly on examinations or coursework, instructors look for reasons to explain such performance. One commonly held belief is that poor results are due to lack of effort or ability. In this study, two often overlooked reasons for poor student performance are explored. This study seeks to make a connection by looking at student self-assessment data of pre and post-examination performance and reported study time with actual examination performance.

After results on an examination in an organic chemistry course were below expectations, a study was initiated to probe reasons for the level of performance. The purpose was to investigate the performance and to find ways to help students improve. The study utilized a post-examination reflection developed by Dexter Perkins (*1*). The survey was modified slightly from the one developed by Perkins (modifications were non-substantive and only served to make the survey relevant to an organic chemistry course) and included such questions as:

After studying for this exam, what grade (out of 100) did you expect to earn?

After completing the exam, what grade (out of 100) did you expect to earn?

Approximately how many hours did you spend studying for the exam?

Approximately how many hours do you spend in general studying for this course each week?

Did you study enough?

Work by in psychology by Kruger and Dunning (2) suggests that those who are weakest at a task often believe that their abilities are at a higher level. They found the level of disconnect increased as skill level decreased. They found that most people believe that they are above average when asked to self-assess their skill level (2). They showed that when one is not proficient in an area that they lack the ability to recognize that fact.

This type of self-assessment has seen the most study in psychology (2), and other fields (3–7), but studies in chemistry are limited (8, 9). Bell and Volkmann used surveys to assess students' learning in a general chemistry class. Their study (8) indicated that the Kruger–Dunning effect was seen on the final examination in their general chemistry course. However, no studies in higher level chemistry courses have been done. Data from general chemistry and organic chemistry have the potential to be very different. One reason is that students may be able to self-assess better when they are already familiar with material. According to The National Center for Education Statistics, in 2009, 70.4% of high school graduates had taken a chemistry course in high school (10). The overall effect is unknown, but as these students have already taken examinations and been graded in a general chemistry course this may affect their self-assessment. Because organic chemistry is not commonly taken by high school students, familiarity with course material as well as prior assessment should be minimized. Based on a survey of these students, the only students that had previously taken organic chemistry were those that were repeating the course. At the start of each semester, students are surveyed as to their chemistry background. The only background in organic chemistry that was found was if a student is repeating the course (and that students taking Organic Chemistry II have already taken Organic Chemistry I. Organic chemistry was also of interest as poorer performing students in general chemistry would be less likely to take organic chemistry. It is hypothesized that students taking organic chemistry would be the students able to predict most accurately their performance in general chemistry (i.e. the students that did well in general chemistry). Would those students who performed well in general chemistry (and likely would have been able to predict accurately) and then performed poorly in organic chemistry, still predict their performance accurately?

Accurate self-assessment in a course such as organic chemistry is critical. Many of the students in this course are taking it due to aspirations for the medical field. However, the ability to recognize one's limitations is not limited to the medical field, as success in any field requires this. One who has a lower level of knowledge but believes that they have a higher level may be less likely to put in the effort needed to succeed at a task associated with using that knowledge.

In addition, amount of preparation time was also examined. The amount of preparation time can be directly related to self-perception. If a student believes that they are well prepared, they may see little need to study further. This leads to the potential that inaccurate self-assessment by poor performing students may lead to insufficient effort. Students were asked how many hours they had spent preparing for the examination as well as whether they had spent an appropriate amount of time studying. If it is found that poorer performing students are reporting similar numbers of study hours to the top students, then the reasons for poor performance are likely more complicated, and at least partially related to the quality of study time and not the quantity.

While common sense indicates that a student will not be successful in an organic chemistry course without studying, previous work indicates a mixture of results on whether there is a direct correlation between amount of time spent studying and performance in the course. There is a limited body of work in chemistry exploring the connection between amount of time spent studying and course performance. Jaisen found a direct correlation between passing an organic chemistry course and the amount of time that students spent studying (11). There have been studies in other fields looking at this correlation, mainly in psychology. Landrum, et al. found an inverse correlation – better performing students often spent less time studying than students earning “C”, “D”, & “F” grades (12). A number of other studies have found a weak (or no correlation at all) between time spent studying and course performance (13–16).

This study sought to explore these two topics further (student self-assessment of examination preparation and the correlation of hours of study and course performance) in an organic chemistry classroom. This was also of interest as the previous study in chemistry by Jaisen seems to disagree with most published work, as Jaisen found a direct correlation between the amount of time studying and passing an organic chemistry course while studies in other fields have found at best a weak correlation between time studying and course performance.

## Methodology

The survey was designed to ask students to reflect on their performance and their preparation to help them improve on future coursework. Exam surveys based on the one developed by Perkins were distributed to students when they received their graded exams. Students received a small number of bonus points based only on completion of the survey. The data reported here (n = 187) are combined from three courses (Organic Chemistry I, Summer 2012; Organic Chemistry I, Fall 2012; Organic Chemistry II, Spring 2012).

The spring and fall courses consisted of 2.5 hours of lecture each week for 15 weeks while the summer course has 7 hours of lecture each week for 5 weeks. The prerequisite for Organic Chemistry I is General Chemistry II and the prerequisite for Organic Chemistry II is Organic Chemistry I. Courses generally consist of 3 exams during the semester plus a cumulative final exam (no surveys are done for

the final examination), and the author was the sole instructor for the lectures. The data presented here is based on all three of the semester examinations from each course.

The study and use of the data was approved by the IRB at Southern Connecticut State University.

## Results and Discussion

### Pre-Exam Perception

Table 1 shows the comparison between students' expected performance on their examinations and their actual performance (these surveys are collected from all examinations excluding the final during the given semesters). As the work by Kruger and Dunning suggests, students who performed worst have predictions that were the least accurate. Students that earned "A" and "B" grades on exams were highly accurate in their predictions. Students who earned an "A" tended to underestimate their scores slightly, while "B" students tended to slightly overestimate. However for both groups, the mean difference is less than three points. As scores decreased, the accuracy of the predictions also decreased. Students who earned "C" grades expected "B" grades based on their preparation, students who earned "D" and "F" grades expected "C+" grades, with "F" students being more than 30 points below their predicted score on average.

As Kruger and Dunning indicated (2), most people believe that they are above average for a given task. This data here are consistent with this. In this group, 171 out of 187 students (91.4%) predicted that they would score above the overall mean for these examinations. As further illustration of the lack of awareness of their own preparation, 62% of "A" students under-predicted their examination scores while 92% of "C", "D", and "F" students combined overestimated their scores. Statistical tests (f-tests and t-tests) to verify the validity of the data were performed and showed each of the samples to be independent (17).

Performance was then broken down into smaller grade groups (Table 2). The grading scale used in this course is found in Table 3. This breakdown revealed some additional trends. Although the sample size was small, students who received the highest grades ("A+", higher than 96%) had the highest under-prediction of any group – underestimating by a full letter grade. Students who earned an "A" and "A-" under-predicted but by smaller margins. Most "B" students had small over-predictions, but there was little difference between "B+", "B", and "B-" students. From "C+" grades through "D" grades the level of over-prediction ranged from 9 to 14 points. Finding this level of consistency within this group of five grade categories was initially surprising given the data seen in Table 1. However, it was explained by the fact that the "D-" students were more similar to "F" students than either "D" or "D+" students. This resulted in the difference seen between "C" and "D" students in Table 1.

**Table 1. Comparing Students Pre-Examination Prediction to Their Actual Score. (Reproduced with permission from reference (18). Copyright 2013 American Chemical Society.)**

<i>Group of Students</i>	<i>Number of Students</i>	<i>Expected Examination Grade after Studying (Mean) (%)</i>	<i>Actual Examination Grades (Mean) (%)</i>	<i>Difference of Means (%)</i>
Received an A	37	87.56	90.49	-2.93
Received a B	41	82.08	79.46	2.62
Received a C	53	78.51	67.21	11.30
Received a D	22	70.18	54.68	15.50
Received an F	34	71.67	39.80	31.87
All students	187	79.07	68.36	10.71

**Table 2. Comparing Students Pre-Examination Prediction to Their Actual Score (Detailed)**

<i>Group of Students</i>	<i>Number of Students</i>	<i>Difference of Means (%)</i>
Received an A+	5	-12.5
Received an A	13	-3.9
Received an A-	19	-0.6
Received a B+	14	1.8
Received a B	16	3.4
Received a B-	8	1.6
Received a C+	15	9.6
Received a C	18	10.2
Received a C-	20	13.0
Received a D+	7	9.2
Received a D	7	13.9
Received a D-	8	25.7
Received an F	34	32.3
All students	187	10.7

**Table 3. Grading Scale Used. (Reproduced with permission from reference (18). Copyright 2013 American Chemical Society.)**

<i>Letter Grade</i>	<i>Percentage</i>	<i>Letter Grade</i>	<i>Percentage</i>
A+	96–100	C	66–69
A	91–95	C–	62–65
A–	86–90	D+	58–61
B+	82–85	D	54–57
B	78–81	D–	50–53
B–	74–77	F	<50
C+	70–73		

### **Post-Exam Perception**

Upon switching to students' post-examination predictions, a marked difference was seen in all but one group (Table 4). While "A" students on average under-predicted before taking the exam, the degree to which they had done so was much larger after the exam. After completing the exam, "A" students predicted a score that was almost 9 points lower than what they actually earned. Students who earned a "B" had gone from a slight over-prediction to a slight under-prediction. Again, on average, "B" students were the most accurate. The biggest difference was in the "C" and "D" students. For these students who both had overestimations in their predictions before the examinations, they now had a much smaller overestimation (3-4 points). This suggests that these students were able to self-assess their performance accurately after completing the examination, but were unable to do so before the examination. The students that earned "F" grades on their exams were still more than 20 points high. However, this was closer than what they predicted before the exam (pre-exam prediction: "C+", post-exam prediction: "D+"). Statistical tests (F-tests and T-tests) to verify the validity of the data were performed and showed each of the samples to be independent (17).

As with the pre-examination predictions, this data was also broken up into more specific categories (Table 5), with similar patterns. Again, "A+" students were the largest underestimators of their performance at 14.5 points. "A" and "A–" students had a larger overestimation as well. "B" students had a small underestimation from 3 to 5 points. Students in the range of "C+" to "D" range had a small overestimation of their grades. Students who earned a "D–" had a larger overestimation (8 points). As with the pre-exam estimations, without the "D–" students, the "C" and "D" students would have nearly identical data. This suggests that these students (50-53%) behaved more like students that fail even though their actual scores put them in a passing category.



**Table 4. Comparing Students Post-Examination Prediction to Their Actual Score. (Reproduced with permission from reference (18). Copyright 2013 American Chemical Society.)**

<i>Group of Students</i>	<i>Number of Students</i>	<i>Expected Exam Grade after Taking the Exam (Mean) (%)</i>	<i>Actual Examination Grades (Mean) (%)</i>	<i>Difference of Means (%)</i>
Received an A	37	81.75	90.49	-8.74
Received a B	41	76.36	79.46	-3.10
Received a C	53	70.87	67.21	3.66
Received a D	22	58.71	54.68	4.03
Received an F	34	61.15	39.80	21.35
All students	187	71.33	68.36	2.97

**Table 5. Comparing Students Post-Examination Prediction to Their Actual Score (Detailed)**

<i>Group of Students</i>	<i>Number of Students</i>	<i>Difference of Means (%)</i>
Received an A+	5	-14.5
Received an A	13	-7.5
Received an A-	19	-8.3
Received a B+	14	-2.8
Received a B	16	-3.2
Received a B-	8	-5.4
Received a C+	15	2.8
Received a C	18	5.1
Received a C-	20	5.3
Received a D+	7	4.2
Received a D	7	2.1
Received a D-	8	8.2
Received an F	34	22.1

The results from this study suggest that improving student performance when examinations are a major assessment method requires that students can understand their own level of mastery before taking the examination. The students that performed poorly may have benefited from preparation in which they attempted problems under examination conditions. Given that “C” and “D” students had a

higher degree of accuracy after taking the examination indicates that under the right conditions, many lower-performing students can realize that their level of preparation is inadequate. Therefore it is desirable to get students to modify their study habits to allow for accurate self-assessment before taking an examination.

### Time Spent Studying

Given the evidence of the Kruger-Dunning effect, one possible explanation for student performance is that they may not be putting in the necessary time to succeed. If a student thinks that they are well prepared for an examination they may have little reason to continue preparing. Therefore, their inaccurate self-assessment may prevent them from putting in the necessary effort. On the same post-examination survey, students were asked to indicate the amount of time that they had used to study for the exam as well as whether they believed that it was sufficient. The results for mean hours of study are seen in Table 6. Two groups reported mean study times above the mean for the course. These were students that received “A’s” and “F’s”. The lowest amount of time spent studying were the students that received “B”’s. However, the amount of time for “B”, “C”, and “D” students only differed by 0.62 hours (or less than 40 minutes). This suggests that there is little correlation between the amount of time that students spend studying and their actual performance.

**Table 6. Comparing the Number of Hours Students Studied to Their Examination Performance**

<i>Group of Students</i>	<i>Number of Students</i>	<i>Time Spent Studying For Examination (Mean) (Hours)</i>
Received an A	37	16.7
Received a B	41	14.0
Received a C	53	13.8
Received a D	22	14.4
Received an F	34	19.3
All students	187	15.4

Students were also asked if they believed they had studied enough for the examination (Table 7). Not surprisingly, “A” students had the highest percentage of students who were satisfied with the amount of time they had used to prepare. However, even among the “A” students, only half said that they had spent enough time preparing for the exam, and of the students who had received grades above 96% (“A+”) only 60% indicated satisfaction. This continues to provide insight into why these students are the top students – even after doing extremely well, they are aware of their limitations and believe that they can achieve more. The

lowest percentage of students who were satisfied with the amount of time they spent studying were students who received “D” grades. It was surprising that students who failed (received F grades) the exams were the second highest group to say that they had studied enough.

This has some interesting intersection with the pre- and post-exam predictions. Students who earned “C” and “D” grades both indicated in high percentages that they could have spent more time studying. Both groups were far off in their pre-examination prediction, but accurate in their post-examination predictions. This could indicate that after these students had done poorly they realized that there may have been problems with their preparation. “F” students were highly incorrect in both pre-exam and post-exam predictions. They also had the second highest percentage of students that said they had studied enough. Therefore, there are multiple issues in getting these students to perform better. First, while they report that they are spending the most time studying, the “F” students received the poorest grades. This suggests that they might not use their studying time effectively, although this does not take into account performance in prior courses (e.g. general chemistry). Second, they are unable to perceive that there is a problem with their mastery of material either before or after the examination. Until this group is able to realize what they do and do not know, and that they are not studying efficiently, it will be difficult for them to improve. However, this offers the most hope for students in the “C” and “D” range. These students do realize that they are not performing well in a post-examination prediction, and also understand they have not spent enough time studying. Therefore if one can get these students to realize that they are not prepared earlier, there is the potential for significant improvement with this group.

**Table 7. Comparing the Students’ Self-Assessment As To Having Spent an Adequate Amount of Time Preparing**

<i>Group of Students</i>	<i>Number of Students</i>	<i>Percentage of Students That Said They Had Spent Enough Hours Studying</i>
Received an A	37	51.4
Received a B	41	32.5
Received a C	53	21.8
Received a D	22	15.8
Received an F	34	38.7
All students	187	32.4

Reliability of self-reported data is always a concern for any research study that relies on it. Students may want to report in ways that make them look better. Work by Ehrlinger, Johnson, Banner, Dunning, and Kruger (18) gives a higher degree of confidence in this type of data. They indicated that when people were offered a

monetary reward for accurate self predictions, the accuracy did not improve over when no reward was offered. Studies such as this provide increased confidence in the use of self reported data.

What needs to be explored next is the quality of the students' time spent studying. Based on the data it does not show a difference in how much students are studying (with the exception of the "F" students who recorded the most hours studying) compared to the grades that they receive. Preliminary data in that area seems to indicate a difference from the data presented here; however more data is being collected (19). If true then the next step is to explore how students are spending that time.

Future work will explore more details regarding how students are spending that time, and attempting to contrast between groups. If "F" students are putting in the most time, why are they not seeing positive results? Additional survey questions will be analyzed to attempt to find information about this question. In addition, both the pre-examination and post-examination studies are being extended to larger groups of students to examine these phenomena in greater detail.

## Conclusions

The data here clearly indicate two things. First, there is a clear pattern of the Kruger-Dunning effect in these organic chemistry courses. Students that are poorly performing on examinations are unaware prior to taking the exam that they are poorly prepared. After completing the examination, most students are more aware of how they have done. Students at different performance levels are spending similar amounts of times studying. This contrasts with Jaisen's work which showed a positive correlation between amount of study time and course performance. The work here suggests that for many poorer performing students the primary issue may be the quality of time studying and not a lack of it.

## Acknowledgments

The author wishes to thank Nicholas Karatjas for helpful discussions on the surveys and suggestions on statistical analysis.

## References

1. Perkins, D. *Reflection After Exam #1*; <http://serc.carleton.edu/NAGTWorkshops/metacognition/activities/28500.html> (accessed Mar 2012). Please contact the author if interested in a copy of the survey as used in the organic chemistry courses.
2. Kruger, J.; Dunning, J. J. *Pers. Soc. Psych.* **1999**, *7*, 1121–1134.
3. Jordan, J. J. *Stat. Educ.* 2007, *15*, <http://www.amstat.org/publications/jse/Jordan.pdf> (accessed Mar 2013).

4. Wirth, K.; Perkins, D. *Knowledge Surveys: An Indispensable Course Design and Assessment Tool*; <http://www.macalester.edu/geology/wirth/WirthPerkinsKS.pdf> (accessed Jan 2013).
5. Bowers, N.; Brandon, M.; Hill, C. D. *Cell Biol. Educ.* **2005**, *4*, 311–322.
6. Grimes, P. J. *Econ. Educ.* **2002**, *33*, 15–30.
7. Austin, Z.; Gregory, P. A. M. *Am. J. Pharm. Educ.* **2007**, *71*, 89.
8. Bell, P.; Volckmann, D. J. *Chem. Educ.* **2011**, *88*, 1469–1476.
9. Potgieter, M.; Ackermann, M.; Fletcher, L. *Chem. Educ. Res. Pract.* **2010**, *11*, 17–24.
10. National Center for Education Statistics. *Percentage of public and private high school graduates taking selected mathematics and science courses in high school, by sex and race/ethnicity: Selected years, 1982 through 2009*; [http://nces.ed.gov/programs/digest/d12/tables/dt12\\_179.asp](http://nces.ed.gov/programs/digest/d12/tables/dt12_179.asp) (accessed June 2014).
11. Jaisan, P. G. *Chem. Educ.* **2003**, *8*, 155–161.
12. Landrum, R. E.; Turrisi, R.; Brandel, J. M. *Psychol. Rep.* **2006**, *98*, 675–682.
13. Plant, E. A.; Ericsson, K. A.; Hill, L.; Asberg, K. *Contemp. Educ. Psychol.* **2005**, *30*, 96–116.
14. Allen, G. J.; Lerner, W. M.; Hinrichsen, J. J. *Psychol. Rep.* **1972**, *30*, 407–410.
15. Schuman, H.; Walsh, E.; Olson, C.; Etheridge, B. *Soc. Forces* **1985**, *63*, 945–966.
16. Gortner Lahmers, A.; Zulauf, C. R. *J. Coll. Stud. Dev.* **2000**, *41*, 544–556.
17. Karatjas, A. G. *J. Chem. Ed.* **2013**, *90*, 1096–1099.
18. Ehrlinger, J.; Johnson, K.; Banner, M.; Dunning, D.; Kruger, J. *Organ. Behav. Hum. Decis. Processes* **2008**, *105*, 98–121.
19. Karatjas, A. G. Unpublished results.

## Chapter 9

# The Role of Non-Content Goals in the Assessment of Chemistry Learning

Jessica J. Reed and Thomas A. Holme\*

Department of Chemistry, 1105 Gilman Hall, Iowa State University,  
Ames, Iowa 50011

\*E-mail: [taholme@iastate.edu](mailto:taholme@iastate.edu).

As technology continues to make information and facts readily accessible, the importance of understanding the context of the information and demonstrating how to use it appropriately will provide better indications of learning than factual recall. This chapter examines the manner in which curriculum and assessment reforms are moving toward promotion of student skill development beyond traditional content knowledge recall. A discussion of the current state of non-content skill assessment in chemistry is presented noting in particular that instructor interest in non-content aspects of learning appears to outpace the measurement of them. Additionally, the chapter presents data from a national survey. These data were used to understand the relative importance of non-content goals and skills in the general chemistry classroom. How these data will inform future efforts to create appropriate formative and summative assessments of goals and skills beyond content knowledge is also discussed.

### Introduction

In a world where facts are accessible with a click of a button, simple factual recall is no longer the appropriate principle indicator of learning. Rather the context of the knowledge and the ability to use it appropriately are of greater importance. Official reports that use this premise to call for various education

reforms have been prominent components of policy debates (1). Not surprisingly, calls for curriculum reform in chemistry often echo these sentiments. One theme for implementation of suggestions such as these notes the need for data-driven and evidence-based curriculum and assessments (2–5).

Beyond the policy calls, and at least partly in response to them, several efforts to revise science curricula have arisen. Among the most ambitious are the recent changes in both the curriculum and tests associated with Advanced Placement (AP)<sup>®</sup> courses in several sciences, including chemistry (6). In this case, developers at College Board have shifted to an evidence-based approach to curriculum design that utilizes Evidence Centered Design (ECD) (7) along with principles of “backwards design” (8, 9). In this model for curriculum design, learners are expected to master not only content essential to the understanding of scientific concepts, but additionally meet expectations about what they should be able to do with that knowledge (10). In order for ECD to accomplish its goals, assessments need to be carefully constructed in order to measure whether a learner has successfully achieved all of the desired outcomes for the course beyond recall of factual knowledge. The current state of this curriculum development process is described in the re-designed AP chemistry curriculum by College Board (11). A key component of this approach lies in the definition of learning objectives (LOs) that were specifically created to integrate “essential knowledge” (content) and “science practices.”

The Next Generation Science Standards (NGSS) are designed in similar fashion to the reforms of AP courses at the high school level. The ultimate goal of the NGSS is to aid science education at the K-12 level by describing what all students should know and be able to do by certain grade levels (12). While there is no standardized curriculum or assessment associated with the NGSS, the interconnectedness of core content, practices, and crosscutting concepts implies that assessments will need to measure all three cohesively.

Regardless of the intended audience of the reform effort, it is evident that attempts to move beyond simple factual recall assessments to develop rich assessments that measure the development of student skills and practices are becoming increasingly commonplace. The effects of such efforts promise to change how chemistry is taught and assessed at the post-secondary level, as future generations of college students may enter the classroom prepared to engage with the content in different ways. With this potential future in mind, the goals of general chemistry instruction and assessment at the collegiate level should be prepared to consider the development of content knowledge and to encompass development of skills and practices that students can transfer to other courses and disciplines.

What such a curriculum and assessment regime might look like in practice is not yet established in the literature. The concept of considering curriculum development in conjunction with assessment reform has been proposed (13) where assessment design is driven by curriculum prerogatives, and assessment data informs changes in curriculum. This is not to say that multiple modes of assessment have not already been developed within chemistry. Nonetheless, evidence suggests that many chemistry faculty members are aware of a relatively small number of assessment methods and instruments (14–16).

Currently, efforts within the chemistry education research community are seeking to provide means for assessment of student performance beyond content. Assessment instruments used in chemistry education include several that are not directed strictly at content knowledge measurement. For example, an instrument to measure student attitudes about learning chemistry, Attitude toward the Subject of Chemistry Inventory (ASCI), was created by Bauer (17). The instrument measures students' attitudes by asking students to select the position on a semantic differential that most closely relates to their perceptions of chemistry. Xu and Lewis later refined the instrument to a shorter version which measures fewer constructs than the original (18). Other instruments such as CHEMX (19) and CLASS (20) focus on students' expectations and beliefs about chemistry. The CHEMX instrument aims to compare student expectations of the chemistry learning environment to those of faculty within the context of a specific chemistry course, including the laboratory. The CLASS instrument compares student beliefs about chemistry in general to those of experts. While some of the constructs measured by the two instruments overlap, each instrument measures a unique piece of the chemistry experience from the perspective of students. Additionally, the Metacognitive Activities Inventory (MCAI) measures students' metacognitive awareness and how that awareness influences chemistry problem solving skillfulness (21, 22). It is important to note that this summary highlights only a small fraction of the number of published instruments available for use in chemistry instruction. While these assessment instruments do not specifically intertwine the measurement of chemistry content knowledge with content independent skills, they are important for use in classroom contexts to understand better the development of specific attitudes and skills by students.

The number and variety of assessment instruments that have been developed illustrates the apparent demand for assessment measures that go beyond content knowledge. To some degree, however, instrument development has tended to result in only modest implementation. In other words, the number of times in which non-content aspects of learning have been explored in a preliminary way via instrument development is growing, but day-to-day usage of such tools has shown a less robust pattern, at least in terms of literature (23). This does not imply an outright lack of interest in the measurement of non-content learning goals. Indeed, usage of assessment tools in classrooms that go unreported in the literature may be common. Nonetheless, from the literature base alone, it is not easy to ascertain the key non-content characteristics chemistry instructors feel are important to measure. Therefore, it is important to 1) understand what skills and practices general chemistry instructors value for students to develop and 2) think about how future assessment designs might incorporate essential content with skill assessment.

### **Arguing the Importance of Non-Content Assessment in Chemistry**

Beyond the impetus from emerging curriculum development and studies within chemistry education research, there are two important aspects of chemistry instruction that suggest the measurement of non-content goals may be important. First, theories of how people learn have repeatedly included key components



that are not formally related to content knowledge alone. Second, for many forms of pedagogical improvement, an increase in non-content components of learning may be important. In this sense, the potential importance of measuring non-content goals follows a familiar theory and practice breakdown that can be elaborated further.

### *Theories of Learning and the Role of Non-Content Assessment*

Novak's Theory of Education, Human Constructivism, is integral to the design and analysis of this research (24, 25). Novak draws heavily upon the ideas of psychologist and philosopher David Ausubel's *assimilation theory* which describes the differences between rote and meaningful learning, outlines the conditions necessary for meaningful learning, and suggests that meaningful learning occurs when the learner is afforded experiences in each of the three learning domains (cognitive, affective, and psychomotor) (26). Meaningful learning is achieved only when all three components are present.

Novak's theory asserts that knowledge is a human construction, and thus it is incumbent upon the educational system to support learners as they construct knowledge (24). Additionally, meaningful learning empowers students to commit and be responsible for learning by integrating thinking, feeling, and acting. Therefore, this framework provides a unique lens to analyze the learning goals of general chemistry instructors because it establishes a basis to understand how the learning goals provide an opportunity for meaningful learning in a general chemistry course (27).

It is also important to consider that the general chemistry classroom provides experiences that are unique to the discipline of chemistry. That is to say that the learning that occurs within the general chemistry classroom is situated within the context of a chemistry community. Thus it is useful to consider that activity, concept, and culture found within the chemistry classroom are interdependent. The theory of situated cognition provides an additional lens for understanding the role of activity to develop skill and concept creation within the realm, or culture, of general chemistry (28). It is posited that even though students acquire tools, or skills, they will not know how to use them appropriately if not given opportunities to use them within the context of the discipline (28). This suggests that even though opportunities for meaningful learning may be presented to students, the knowledge and skills acquired may remain decontextualized, and even inert, unless students are presented with insight about how those concepts and skills are actually used within chemistry and how to transfer them to applicable real-life situations (29).

Additionally, the importance of the interconnection of content knowledge and procedural skills in understanding learning is shown within the Unified Learning Model (ULM) (30). The ULM provides a model of how people learn, and a resultant model of teaching and instruction, by drawing on the principles of cognitive science and psychology. In this model, working memory, knowledge, and motivation are central to understanding how all people learn. Knowledge in this case refers not only to concepts or facts (declarative knowledge), but also to

the skills, behaviors, and thinking processes that an individual knows (procedural knowledge). Learning is then influenced by the individual's working memory capacity, the concepts and skills he or she already knows (prior knowledge), and the goals that drive him or her to put forth effort. In this model the instructor aids the learner by directing the student's attention (working memory) to the concept or skill to be learned, providing opportunities for the creation of new connections between prior knowledge and the new concept or skill, and creating goals to support the motivation of the student to learn. In this sense the instructor serves as a mere facilitator of individual learning, yet guides the course of the learning experience by influencing the content and skills developed through specific course goals and objectives.

### *Practical Implications of Measuring Non-Content Learning in the Classroom*

There is little question that content knowledge gains represent the main goal of any course, and chemistry courses are no exception. However, it is also true that understanding just how teaching methods influence the efficiency of learning often hinges on non-content aspects. In particular, the concepts of student engagement, student motivation or student persistence have received considerable attention in research studies regarding how to promote learning success in chemistry (31–33). Perhaps just as importantly, the measurement of non-content variables is often measured as a part of formative assessment during attempts at curriculum or pedagogical innovation. Determining whether or not students “like” a new approach, is often reported – but it is arguable that non-content learning can be parsed with significantly more resolution than this construct.

Several teaching methodologies have emerged with an intention to improve content learning and provide non-content gains as well. Within chemistry, Process Oriented Guided Inquiry Learning (POGIL) is perhaps the most prominent example (34–37). For this teaching method, the process-orientation component is focused on enhancing the development of generalizable process skills that allow students to gain more content knowledge. Other teaching methods such as problem based learning (38), case-based historical development of chemical concepts (39) and active learning via a “flipped” classroom (40) all include aspects that relate to student engagement and non-content skill development.

While a number of research questions related to the assessment of the non-content components of these emerging methodologies still remain, the methodologies themselves serve to exemplify the practical nature of enhancing student skills in addition to content knowledge.

Before researchers can address creation of assessment materials for measurement of non-content goals and skills, it is necessary to understand what are the goals and skills that chemistry instructors value. The survey and data presented here aim to inform the community about the types of goals and skills that are valued in the general chemistry curriculum.

## Methodology for the Study

### *Survey Development*

Quantitative survey items were developed from themes present in qualitative interviews conducted with chemistry instructors about the learning goals present in introductory general chemistry courses. The semi-structured interviews were conducted with 18 general chemistry instructors from high schools, community colleges, and state-funded universities. Participants were asked open-ended questions that progressively became more specific depending on a participant's response, such as "What are the learning goals you have for your general chemistry course?" to "What are the non-content goals you have for students in your course?" (41). The interviews were then transcribed and open-coded using a Grounded Theory approach (42). Additionally, learning goals were labeled according to the primary domain (cognitive, affective, or psychomotor) associated with the goal. Interestingly, participants often discussed incorporating a variety of goals into their courses, but felt that students did not meet the often implicit expectations associated with these goals even though they did not formally assess their non-content goals (41). In order to obtain more generalizable results about the status of non-content learning goals, the ten most frequently discussed non-content goals from the interviews were transformed into survey items. The survey items were part of a national online survey from the ACS Examinations Institute about conceptual understanding in general chemistry.

The major non-content goals surveyed were: appreciation of chemistry in everyday life, development of communication skills, laboratory skills, graphing of data, interpreting and drawing conclusions from data, life skills (e.g., study skills, responsibility, time management), problem solving skills, nature of science (i.e., how science works and has developed), critical thinking, and conceptual understanding of traditionally algorithmic problems.

### *Survey Items*

Three questions on the survey related to non-content goals and each question evaluated all ten non-content skills identified as common themes amongst qualitative interview participants.

The first question related to learning goals asked participants to indicate how often they intentionally and explicitly incorporated the learning objectives into their course. Response choices ranged from "I do not incorporate this" to "Every class period," with options of "Once or twice per semester," "Once per month," and "Once per week" in between. Participants were only able to select one response choice per learning goal.

The second question in the set related to how the learning goals were assessed in the course. Participants were asked to select all modes of assessment that applied to each learning goal. Methods of assessment surveyed were clickers (student response systems), exams, homework, laboratory reports, and quizzes.

Additionally, response options were available for participants that did not assess or did not incorporate a goal in the course.

The final question related to learning goals asked participants to describe, on average, how well they felt that students met their expectations for these learning goals. Respondents were allowed to choose one response from five choices ranging from “Below my expectations” to “Exceeds my expectations.”

### *Sample*

The sample consisted of chemistry instructors and faculty at community colleges, four-year colleges, and universities in the United States who had taught a general chemistry course within the past five years. Institutional classifications were based upon the self-reported highest degree offered in chemistry at the participant’s institution. The sample excluded instructors of special topics courses and General, Organic, and Biochemistry (GOB) courses. For analysis purposes, only participants who completed all questions relating to learning goals were considered as part of the sample (N=1,075). Table 1 shows participant distribution by institution type. General chemistry teaching experience of participants ranged from one year to 40 years experience, with an average of approximately 15 years. Additionally, 84% of the sample had taught a full-year (two-semester) general chemistry course and 75% were responsible for teaching both a lecture and laboratory component of the course.

**Table 1. A Description of Quantitative Survey Participants by Institution Type**

<i>Survey Participant Demographics</i>		
<i>Institution Type</i>	<i>Participants</i>	<i>Percent of Sample</i>
Community College	170	15.8
Bachelors Institution	513	47.7
Graduate Institution	392	36.5
Total	1,075	100

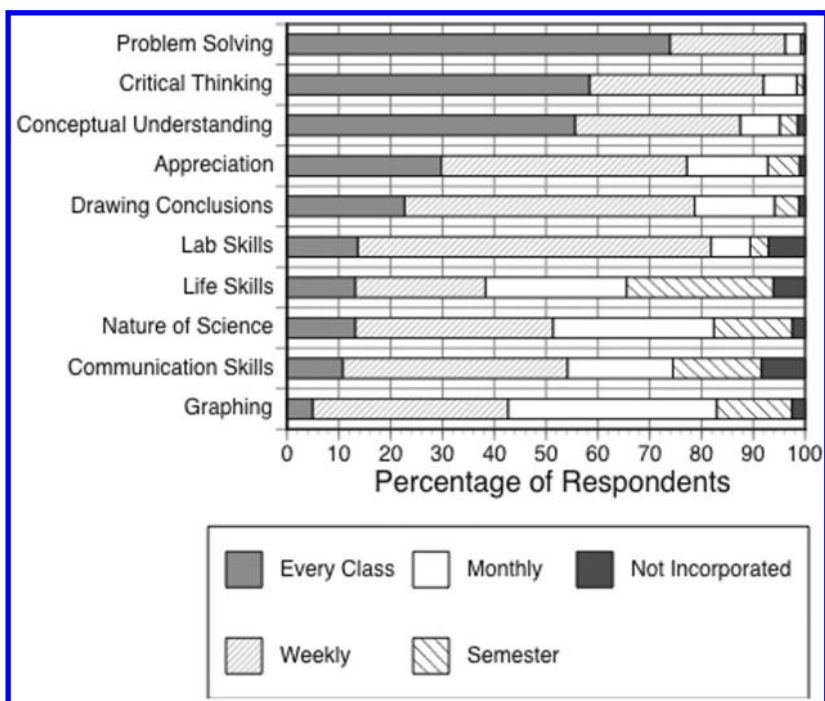
## **Results**

### *Quantitative Survey Results and Discussion*

Results from the survey provided insight into chemistry instructors’ values of non-content goals and skills.

Responses to the first question about frequency of intentional incorporation of non-content learning goals were as expected. Skills traditionally associated with chemistry courses, such as conceptual understanding, critical thinking, and

problem solving, were reported to be incorporated into every class period by a majority of instructors. Figure 1 displays the frequency of incorporation of the non-content goals and skills as self-reported by instructors. Problem solving appeared to have the highest frequency of incorporation. Approximately 74% of instructors reported incorporating problem solving into every class period, and an additional 22% reported incorporating it on a weekly basis. Less than 1% (0.28%) of instructors reported not incorporating development of problem solving skills as a goal of their general chemistry course. Critical thinking and conceptual understanding also had a majority of respondents indicate that they incorporate those skills into every class period with 58% and 56%, respectively. Additionally, nearly 70% of instructors reported incorporating laboratory skills on a weekly basis. This is consistent with the typical general chemistry course design, which includes a weekly laboratory section. Other goals, such as development of communication skills, showed a broader range of reported incorporation.



*Figure 1. General chemistry instructors' self-reported incorporation of non-content goals and skills. Incorporation ranges from every class meeting to not incorporated at all.*

While these statistics are not surprising due to the nature of general chemistry coursework, it is important to note that these data are self-report so we cannot ascertain for certain whether instructors are actually incorporating these goals in the manner in which they claim. For example, while over 95% of instructors claim to incorporate problem solving into their course at minimum on a weekly basis, it is unclear as to whether participants in this survey were differentiating the nature of problem solving, such as how the course activities might be compared with students performing learning exercises (43). Such distinctions are not wholly necessary for this study because these data were not meant to assert sweeping observations about the condition of the collegiate general chemistry classroom. Rather, the objective is to understand the types of goals and skills that are valued by general chemistry instructors in an effort to understand better the types of non-content skills that future formative and summative assessments could be designed to measure. In this context, it is considered that an instructor who makes an effort to incorporate a goal or skill more frequently likely values that skill more and desires to develop it in students more so than goals that are incorporated on a less frequent basis.

The frequency with which instructors reportedly incorporate non-content goals and skills into their courses provides an indication of the types of skills they hope to develop in their students. Yet, incorporation of a goal into a curriculum does not imply that students successfully develop that skill. Assessment plays a key role in understanding and rating student skill development. In order to understand better how future assessments might be designed to measure content independent learning goals, it was important to elicit how instructors assess non-content goals within their general chemistry courses. Again, these are self-reported data intended for use to understand how instructors perceive these learning goals to be assessed. Respondents were allowed to select multiple modes of assessment for a single learning goal. The modes of assessment were selected from the most frequent responses collected in qualitative interviews, and included clickers, exams, homework, lab reports, and quizzes. Respondents were also allowed to indicate that a particular learning goal was not assessed in their course.

Instructors' responses regarding modes of assessment used can be seen in Figure 2.

For ease of interpretation, responses have been combined to reflect summative assessments (exams and quizzes), formative assessments (homework and clickers), laboratory reports, and responses indicating a goal was not assessed. It is of interest to note that laboratory reports were the most frequent response for assessment of communication skills, laboratory skills, graphing of data, and drawing conclusions from data, whereas problem solving skills, critical thinking about concepts or problems, and conceptual understanding of problems traditionally solved algorithmically are reported as most commonly assessed by exams and quizzes.

Other methods of assessment were not selected as frequently. For example, clickers make up a smaller fraction of the formative assessment category compared to homework. Clickers had minimal use in assessment of the non-content goals except for problem solving. This result may not be surprising in light of previous research about clicker usage among chemistry instructors (44). Goals related to

development of an appreciation of the subject of chemistry, understanding of the nature of science (NOS), and life skills were reported as most frequently not assessed in any fashion.

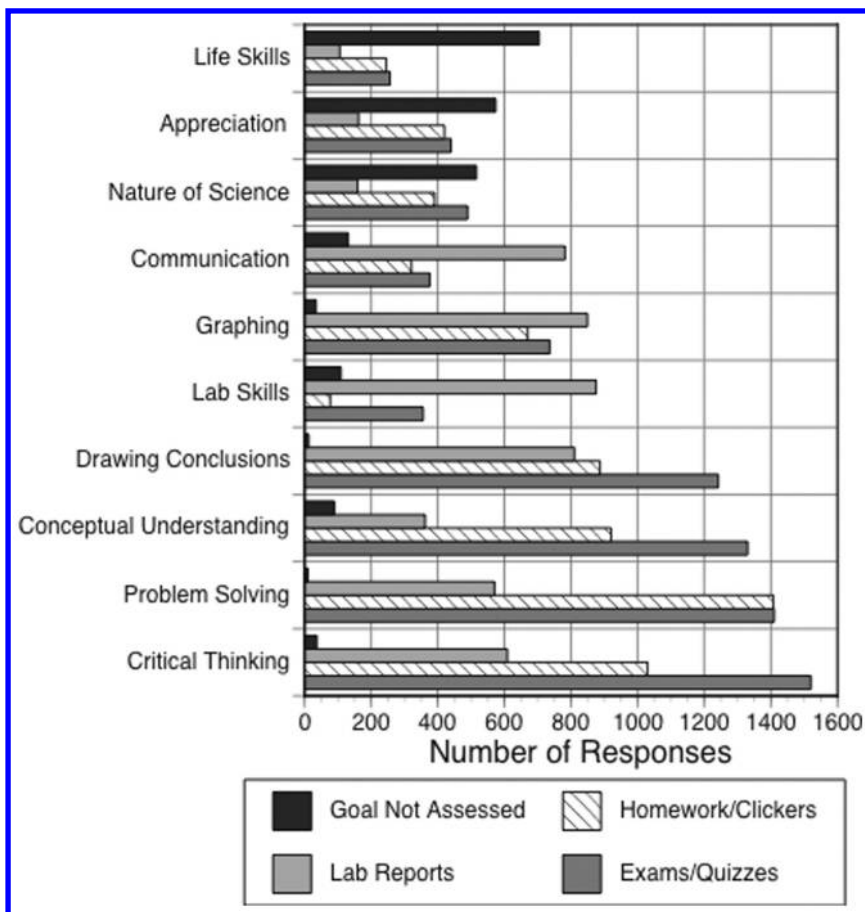


Figure 2. Instructors' self-reported methods of assessment of content independent goals and skills in general chemistry courses.

Instructors reported use of assessments gives insight into how opportunities for meaningful learning are being evaluated in the classroom. Skills that lie predominantly in the cognitive domain (problem solving, conceptual understanding, and critical thinking) are reported as most frequently assessed by exams, whereas skills that lie predominantly within the psychomotor domain, with some overlap of the cognitive domain, such as laboratory skills, communication skills, and graphing are measured by laboratory reports. Affective goals such as appreciation of chemistry and life skills are reported as not assessed at all. While

it is not surprising that there is a disconnect between the methods of assessment (or lack thereof) for each domain, it is indicative of the challenge faced by assessment designers to incorporate more than one domain within a single format of assessment.

Regardless of how the learning goals are purportedly assessed, there appears to be room for improvement in student performance. Instructors were asked to evaluate how students met expectations regarding successful development of these learning goals, and their responses can be seen in Figure 3. Although the percentage of students meeting the expectations of their instructors for development of these non-content goals was generally over half, a sizable fraction of students appear to have fallen short in the estimation of the participants in this survey. Indeed, more instructors rated student performance as “Does not meet expectations” than “Exceeds expectations,” suggesting that there is room for improvement in student performance in non-content aspects of learning. It is important to remember, however, that assessment methods that instructors have indicated are used for non-content goals tend to be more informal. As such, the impressions they form (which presumably inform their answers to this survey item) may lack quantitative rigor. Thus, the expectations reported here, while informative about future challenges related to assessment of non-content learning, should not be considered a rigorous judgement of student non-content learning.

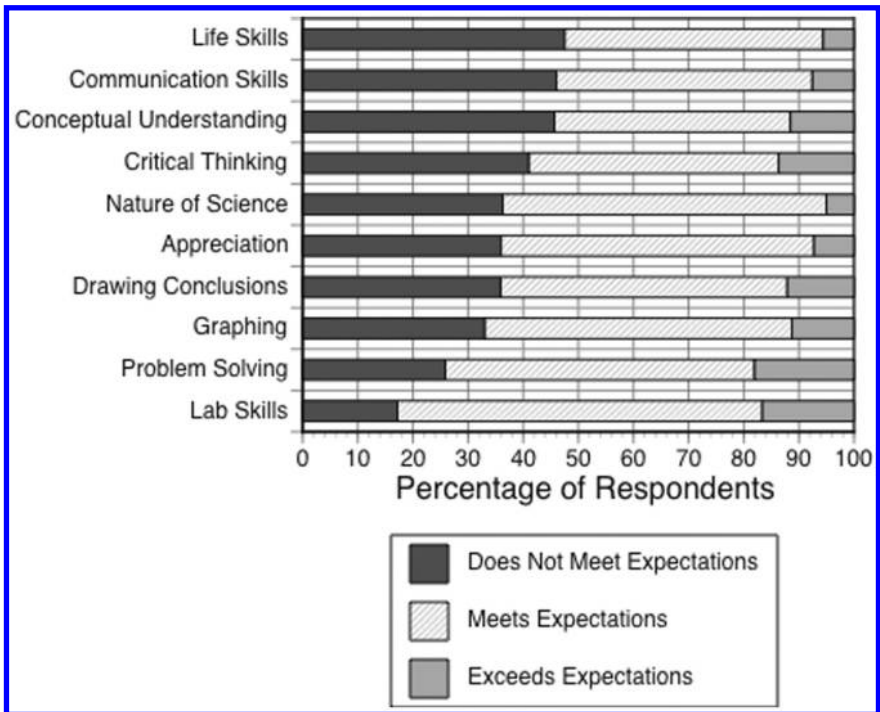


Figure 3. Instructors' evaluation of student performance on achievement of non-content learning goals.



## Summary and Implications

Although it may not be routinely articulated by chemistry instructors, the development of skills beyond the scope of content knowledge in chemistry courses is important and most instructors view it as such. Curriculum reform efforts often influence non-content learning outcomes but without a more rigorous effort to enhance assessment it may be argued that these changes essentially resort to a “hope for the best” approach. The survey research presented here provides evidence that non-content learning goals are valued by the chemistry education community. As such, assessments are needed to measure the development of students’ skills beyond typical content exams.

Calls for changes in the chemistry curriculum focus on the need for evidence-centered and data-driven reform efforts (2–5), beyond measuring whether students “like” an activity. Instruments have been developed to measure student skills beyond the domain of chemistry content knowledge; however, these instruments appear to be underutilized by the traditional chemistry community, perhaps due to a lack of awareness of these instruments. Additionally, these instruments tend to be quite specific and measure only specified constructs. This means that to gain a whole picture of the classroom environment, an instructor would likely need to devote significant effort to administering and analyzing survey instruments. This level of effort may not be practical in the typical general chemistry classroom.

Ultimately, the most attractive trajectory for addressing the need for non-content assessment may lie in finding ways to incorporate it more closely within traditional content assessments. Efforts to devise such assessment are part of the high profile developments in AP Chemistry (6–11) and the Next Generations Science Standards project (12). In order to guide such development the current work suggests an iterative process may be particularly helpful to determine what non-content skills are most important to assess in this way. Instructors appear to be interested in gaining better information about student learning, but it seems reasonable to expect that initial attempts to measure non-content aspects may require refinement. Thus, the collaboration between curriculum reform efforts and assessment development efforts (13) will take on ever more importance as chemistry education moves forward over the next few years.

## References

1. Hess, F. M.; Kelly, A. P.; Meeks, O. *The Case for Being Bold: A New Agenda for Business in Improving STEM Education*; Institute for a Competitive Workforce, Washington, DC, 2011.
2. Lloyd, B. W.; Spencer, J. N. *J. Chem. Educ.* **1994**, *71*, 206–209.
3. Cooper, M. M. *J. Chem. Educ.* **2010**, *87*, 231–232.
4. National Research Council. *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*; Singer, S. R., Nielsen, N. R., Schweingruber, H. A., Eds.; The National Academies Press: Washington, DC, 2012.
5. Cooper, M. M. *J. Chem. Educ.* **2013**, *90*, 679–680.

6. College Board. *The AP<sup>®</sup> Chemistry Curriculum Framework 2013-2014*; College Board: New York, 2011.
7. Mislavy, R. J.; Steinberg, L. S.; Almond, R. A. *Meas. Interdis. Res. Perspect.* **2003**, *1*, 3–67.
8. Brennan, R. L. *Appl. Meas. Educ.* **2010**, *23*, 392–401.
9. Huff, K.; Steinberg, L.; Matts, T. *Appl. Meas. Educ.* **2010**, *23*, 310–324.
10. Mislavy, R. J. *CRESST Report 800*, “Evidence-Centered Design for Simulation-Based Assessment”; National Center for Research on Evaluation, Los Angeles, CA, 2011.
11. College Board. *AP<sup>®</sup> Chemistry: Course and Exam Description*; College Board: New York, 2013.
12. Achieve. *Next Generation Science Standards*; National Research Council: Washington, DC, 2013.
13. Holme, T. A.; Bretz, S. L.; Cooper, M.; Lewis, J.; Paek, P.; Pienta, N.; Stacy, A.; Stevens, R.; Towns, M. *Chem. Educ. Res. Pract.* **2010**, *1*, 92–97.
14. Towns, M. H. *J. Chem. Educ.* **2010**, *87*, 91–96.
15. Raker, J. R.; Emenike, M. E.; Holme, T. A. *J. Chem. Educ.* **2013**, *90*, 981–987.
16. Emenike, M. E.; Raker, J. R.; Holme, T. A. *J. Chem. Educ.* **2013**, *90*, 1130–1136.
17. Bauer, C. F. *J. Chem. Educ.* **2008**, *85*, 1440–1445.
18. Xu, X.; Lewis, J. *J. Chem. Educ.* **2011**, *88*, 561–568.
19. Grove, N. P.; Bretz, S. L. *J. Chem. Educ.* **2007**, *84*, 1524–1529.
20. Adams, W. K.; Wieman, C. E.; Perkins, K. K.; Barbera, J. *J. Chem. Educ.* **2008**, *85*, 1435–1439.
21. Cooper, M. M.; Sandi-Urena, S.; Stevens, R. *Chem. Educ. Res. Pract.* **2008**, *9*, 18–24.
22. Cooper, M. M.; Sandi-Urena, S. *J. Chem. Educ.* **2009**, *86*, 240–245.
23. Arjoon, J. A.; Xu, X.; Lewis, J. E. *J. Chem. Educ.* **2013**, *90*, 536–545.
24. Novak, J. D. *A Theory of Education*; Cornell University: Ithaca, NY, 1977.
25. Bretz, S. L. *J. Chem. Educ.* **2001**, *78*, 1107.
26. Ausubel, D. P. *Educational Psychology: A Cognitive View*; Holt, Rinehart, and Winston: New York, 1968.
27. Bretz, S. L.; Fay, M.; Bruck, L. B.; Towns, M. *J. Chem. Educ.* **2013**, *90*, 281–288.
28. Brown, J. S.; Collins, A.; Duguid, P. *Educ. Res.* **1989**, *18*, 32–42.
29. Roth, W. M.; McGinn, M. K. *Res. Sci. Educ.* **1997**, *27*, 497–513.
30. Shell, D. F.; Brooks, D. W.; Trainin, G.; Wilson, K. M. *The Unified Learning Model*; Springer: Netherlands, 2010.
31. Seymour, E.; Hewitt, N. *Talking about leaving: Why undergraduates leave the sciences*; Westview Press: Boulder, CO, 1997.
32. Zusho, A.; Pintrich, P. R.; Coppola, B. *Int. J. Sci. Educ.* **2003**, *25*, 1081–1094.
33. Sadler, P. M.; Sonnert, G.; Hazari, Z.; Tai, R. *Sci. Educ.* **2012**, *96*, 411–427.
34. Farrell, J. J.; Moog, R. S.; Spencer, J. N. *J. Chem. Educ.* **1999**, *76*, 570–574.
35. Minderhout, V.; Loertscher, J. *Biochem Mol. Bio. Educ.* **2007**, *35*, 172–180.
36. Hein, S. M. *J. Chem. Educ.* **2012**, *89*, 860–864.

37. Chase, A.; Pakhira, D.; Stains, M. *J. Chem. Educ.* **2013**, *90*, 409–416.
38. Overton, T. L.; Byers, B.; Seery, M. K. *Innovative Methods of Teaching and Learning Chemistry in Higher Education*; Elks, I., Byers, B., Eds.; Royal Society of Chemistry: London, 2009; pp 45–61.
39. Obenland, C. A.; Munson, A. H.; Hutchinson, J. S. *Chem. Educ. Res. Pract.* **2013**, *14*, 73–80.
40. Fakayode, S. O.; Yakubu, M.; Adeyeye, O. M.; Pollard, D. A.; Mohammed, A. K. *J. Chem. Educ.* **2014**, *91*, 662–665.
41. Reed, J. J.; Holme, T. A. Unpublished work, 2014.
42. Glaser, B. G.; Strauss, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, 7<sup>th</sup> ed.; Transaction Books: Chicago, IL, 2009.
43. Bodner, G. B. *J. Chem. Educ.* **1987**, *64*, 513–514.
44. Emenike, M.; Holme, T. *J. Chem. Educ.* **2012**, *89*, 465–469.

## Chapter 10

# Classroom Salon – An Innovative Method for Investigating Student Learning

Anja Blecking\*

Department of Chemistry and Biochemistry, University of Wisconsin-Milwaukee, 3210 N. Cramer Street, Milwaukee, WI 53201

\*E-mail: [blecking@uwm.edu](mailto:blecking@uwm.edu).

Classroom technology is an integral part of modern education and has long been used to facilitate student access to course material like textbooks and homework sites. This chapter introduces technology that combines the qualities of a social media application and analytical functions that can be utilized to assess student engagement and learning through textual annotations. This chapter also explores the possibilities the platform offers to increase student success in large preparatory chemistry classes through content-focused student interaction and formative instructor feedback.

### Introduction

Teaching and learning depend greatly on communication between teachers and learners. In an ideal world, instructors would have detailed information about their students' preferred learning style, level of understanding, and existing misconceptions prior to instruction. Unfortunately this level of personal connection is very difficult, even unrealistic, to achieve in high enrollment courses. Students in large classes often feel anonymous and very distant to the course instructor (1–3). This might not pose a real problem for high-performing students but can have a serious effect on struggling students (3). Learning is known to be an active process, in which the construction of new knowledge is based on prior knowledge (4), and therefore identification of the latter is crucial to promote and support learning processes. Students who are unable to make sense of instruction due to their misconceptions might not be able to learn the material

at all. Other students with misconceptions might be able to make sense of the instruction by incorporating the learned material into their own way of thinking (5), which also does not result in learning as intended by the instructor.

Assessing incoming student knowledge can be accomplished in many ways. It is not the author's intention to discuss all different facets of this topic as assessment tools focus on many specific aspects of knowledge, learning, or behavior. It is rather the author's intent to introduce a newly developed social on-line platform that shows potential to impact positively student learning through textual annotations.

In 2011 Classroom Salon or "CLS" ([www.classroomsalon.org](http://www.classroomsalon.org)) was introduced to instructors and students at UW-Milwaukee (UWM), a large urban research and doctoral university, in the context of a research project unrelated to the research described in this chapter. CLS's design as a social, collaborative learning platform presented opportunities to open up new lines of communication through annotations and online discussions between students and instructors in high enrollment courses. CLS has since been implemented into large-enrollment preparatory chemistry classes at UWM on a regular basis. As a bonus, the program is free of cost for educational institutions and students, easy to incorporate and delivers an abundance of data, which can be quickly analyzed using the program's analytical functions to deliver meaningful results.

The initial purpose of increasing communication quickly expanded and CLS is now being utilized to reveal other aspects of learning, such as student misconceptions, questions about chemical concepts, level of use of chemical terminology, student attitudes toward the subject matter, and student engagement with the subject matter.

## **Background**

### **The Online Platform – Classroom Salon (CLS)**

CLS, the Facebook-like technology, had initially been developed by educators and researchers at Carnegie Mellon University (CMU) for use in writing classes (6) in 2010. Since then, CLS has been integrated into courses of many different disciplines in higher education and K-12 school districts. CLS allows its users to create online, collaborative, social learning communities.

In CLS, instructors can create course-specific communities (referred to as "Salon") which function as an online discussion forum. Community members are usually the course instructor and students in the class. Both parties are able to post documents and videos that are then visible to all community members. In other words, a Salon is a collection of users and documents.

Members of a Salon can read documents or watch videos, annotate, comment, and ask questions about content, which initiates an open, on-line discussion with other Salon members. The creative team at CMU proposes that visualization increases student engagement in annotation. Students become aware that their own investment in the text increases their investment in the social community and therefore in the class. Furthermore, through annotation and interaction with other Salon members, students can explore whether or not their own points of view are

reflected in other member's views. This open and transparent concept is meant to lead to deeper text analysis and self-reflection than what can be achieved through traditional reading.

Instructors ("Salon Owners") generally see the same documents and comments as the students, but they also have access to analytic features that allows for convenient management of large numbers of students and comments. These features will be explained in detail under "Classroom Salon Analytics" below, but let's have a look what CLS looks like from a student's perspective. Once logged into the program ([www.classroomsalon.com](http://www.classroomsalon.com)), Salon members gain access to their Salon home page by either following a link provided by the instructor or by simply searching for the Salon name. On their Salon course homepage, students then have access to posted documents, view instructor messages, and also other members that are registered for the Salon (Figure 1).

It is important at this point to mention that CLS offers a variety of settings and analytics for student and instructor use. However, this chapter will only focus on some important key features for Discussion Salons. For a complete description of all available features, see [www.classroomsalon.org](http://www.classroomsalon.org). CLS offers *two primary working modes* for Salon members in Discussion Salons:

1. The *Individual Mode* (Figure 2) allows Salon members to read, highlight, and annotate selected parts of a document. As of now, students can view the comments of their classmates while annotating but soon the program will also offer the option to hide these comments in this mode. It will then be the instructor's discretion to choose the setting that is most suitable for the course. Currently all comments are displayed on the right side of the screen, and the course text is visible on the left.

While annotating, students are encouraged to specify the nature of their comments using tags. Tags can be pre-made by the instructor, for example "This is Important", "Can we discuss this in class", or "I do not understand", which allows the instructor to quickly identify questions and problems when sorting comments by tags. Students can also use the "General" tag when annotating, meaning that there is no particular message attached to their annotation. Additionally, CLS allows students to create their own tags for annotation so that they are not limited to the instructor choices.

Filter functions (not shown) allow students to follow selected students or selectively view their annotations. Users are able to filter annotations by selecting a text region, or by user name or tags.

2. The second mode is called *Discussion Mode*. It encourages students to respond to classmates' annotation and join in open discussions about class content (Figure 3). Student's comments are located on the right side of the screen. The text passage each comment is linked to appear highlighted once members click "See Context" in the comment box. If Salon members wish to reply to student comments, they will open up a textbox by hitting the "Reply" button on the lower left side in the comment box.

CLS also offers an additional feature in which discussion threads can be seen in what is referred to as "Tree View" (not shown). In this mode, the initial comment and replies are visible in the order they were made so that students can more easily follow and join in an ongoing content discussion.

**Instructions**

**Posted documents**

**CLASSROOM SALON**

You are the owner of this Salon.

[Instructor] Chemistry 100 Class, Spring 2014

**About this Salon**

Name: [Instructor] Chemistry 100 Class, Spring 2014  
 Description: Hello Students and welcome to Classroom Salon! You will be able to access your reading assignments here. The reading assignments should be worked on before before we discuss the topics in lecture. Classroom Salon allows to communicate and share your views, concerns and questions with your classmates. In order to increase communication, please make at least 5 comments on the text in addition to answering the questions. Please also reply at least 3 times to comments of their classmates. - Thank you!  
 Owner: UWM Professor  
 ID: 1960  
 Access: This Salon is open to all

**Members of this Salon (158)**

Subject  
 Enter message to all members of this Salon here...

**Documents in this Salon**

Document Title	Views	Comments	Date
Reading Assignment #9 - Ideal Gas Law	404	68	2014/04/23
Reading Assignment #8 - Limiting Reactants	666	83	2014/04/09
Reading Assignment #7 - Solutions	758	85	2014/03/24
Reading Assignment #6 The Mole	828	97	2014/03/13
Reading Assignment #5 - Covalent Bonding	986	109	2014/02/17
Reading Assignment #4 - Ionic Bonding	948	121	2014/02/17
Reading Assignment #3 - The Modern Model of the Atom	992	129	2014/02/02
Reading Assignment #2 - Atomic Theory	1015	122	2014/01/27
Reading Assignment #1	1233	139	2014/01/21

**Salon members**

Figure 1. CLS screenshot of Salon Course Homepage for a Preparatory Chemistry course. Salon members, instructor messages, and documents are visible.

Document (with text passage highlighted)

Tag (e.g. "Important")

The screenshot displays the Classroom Salon interface. At the top, the header reads "CLASSROOM SALON" with navigation icons for "SALONS", "DOCUMENTS", and "ME". Below the header, the page title is "[Instructor] Chemistr... Reading Assignment #3 ...". The main content area shows a document with highlighted text. A dashed box highlights a section of text, and a yellow tag labeled "Important" is placed over it. To the right, a "Comment Box for comments or questions" is visible, containing a text input field and buttons for "Cancel", "Save as draft", "Post", and "Post as anonymous". Below the comment box, there are several discussion threads, each with a "General" category, a user profile, and a timestamp. The threads contain questions and answers related to the document content.

FIGURE 7.11  
(A) This probability map is for the hydrogen electron in its lowest-energy state. The electron is more likely to be found in the darker regions. (B) The size of this orbital is defined by an enclosed region where the likelihood of finding an electron is 95%.

In the modern model of the atom, orbitals of similar size are considered to be in the same principal energy level. The first six principal energy levels for the electron in a hydrogen atom are shown in Figure 7.12. In the hydrogen atom, the energies of the orbitals are restricted to the energies of the principal energy levels.

[Image #2]

Energy-level diagram

Bohr diagram

FIGURE 7.12  
In the modern model of the atom, principal energy levels describe the allowed energies of the electrons in orbitals. In the hydrogen atom, the allowed energies for the electron are the same as the energies of the Bohr orbits.

Question: Why are there multiple possible energy levels in a hydrogen atom? H just has one electron!

Orbitals come in different shapes and sizes. There are four types of occupied orbitals in atoms in their lowest energy states, which we label with the letters s, p, d, and f. The general shapes of

Start a discussion

In the modern model of the atom, orbitals of similar size are considered to be in the same principal energy level.

Important

Enter your comment...

Cancel Save as draft Post Post as anonymous

General over a month ago | Updated over a month ago

The location of where the electrons is very important, because they try to do a role in what and how they interact with other atoms.

see context | reply upvote (0) | bookmark | ☆

General over a month ago | Updated over a month ago

Could someone better explain what a principal energy level is?

see context | reply upvote (0) | bookmark | ☆

General over a month ago | Updated over a month ago

Where the electrons are located determines what they will be able to bond with.

see context | reply upvote (0) | bookmark | ☆

Comment Box for comments or questions

Figure 2. CLS screenshot: Student view in the Individual Mode for student annotation. The highlighted text (left) has been tagged "Important". In the "Comment on your annotation" box (right), the student may comment on this annotation to ask questions or start an online discussion.



classroomsalon.com/annotations/navigate.aspx?document=19875#

[Instructor] Chemistry 100 Class, Spring 2014  
Reading Assignment #6 The Mole

### Moles and Particles

Consider the molecular-level view of hydrogen sulfide gas shown in Figure 4.6. We can identify individual molecules of hydrogen sulfide, with the white spheres representing hydrogen atoms and the yellow spheres representing sulfur atoms. Each molecule is made up of two hydrogen atoms and one sulfur atom. In a collection of molecules, such as the four molecules pictured in Figure 4.6, we can count eight hydrogen atoms and four sulfur atoms, giving a ratio of two hydrogen atoms per sulfur atom, the same ratio as in each molecule.

**FIGURE 4.6**

Hydrogen sulfide, H<sub>2</sub>S, is given off when a sulfide ore reacts with acid.

Unfortunately, we cannot see atoms and molecules in this way because they are too small. Although we can use a scanning tunneling microscope to “see” atoms and molecules on the surface of a solid, as shown in Figure 4.7, this technique reveals only atoms on a surface, not those beneath. Furthermore, because atoms are so small, a dust-sized piece of a solid contains over 10<sup>16</sup> of each type of atom. We certainly wouldn’t want to count that many atoms individually, even if we could. Fortunately, we don’t have to see atoms to count them. Instead we can find the relative masses of the elements in a compound. From this, we can determine the 2:1 ratio of atoms and deduce the formula H<sub>2</sub>S.

**DISCUSSIONS**

[Student A]  
over a month ago  
Important  
What is a scanning telescope?  
REPLY see context

VIEW AS TREE ↑ parent ← previous / → next sibling

[Student B]  
over a month ago  
General  
is there significance to the colors chosen to represent each atom? For example, does the element exist in nature as a similar color or is there a chemical reaction that occurs that produces that color?... or anything?  
REPLY see context

VIEW AS TREE ↑ parent ← previous / → next sibling

[Student C]  
over a month ago  
General  
Yes.. Avogadro's number (6.022x10<sup>23</sup>)  
REPLY see context

VIEW AS TREE ↑ parent ← previous / → next sibling

REPLIES (1)

[Student D]

Reply button

Figure 3. CLS screenshot of the Discussion Mode: Student B commented on the circled text passage on the left. Any Salon member can reply to this comment by simply clicking the “Reply” button which opens up a new comment box.

## Classroom Salon Analytics

All Salon members, including the course instructor, can employ the features described above for the individual and discussion mode and also post documents.

In addition, CLS provides an entire suite of tools to author, manage, discuss, and refine content (6). Instructors or authors can upload and manage course material. Analytical tools, such as participation analysis, are visible on the program “Dashboard” (not shown) and also directly in the text. CLS color-codes text passages according to the frequency of reference. Passages marked in red are the ones most annotated; less frequently referenced passages appear in orange or yellow. Clicking on annotated text passages immediately opens student annotations and tags. Another way to quickly access analytics is given on the course Salon homepage. Next to each document name the total number of comments and the total number of participants per document (Figure 4) are shown.

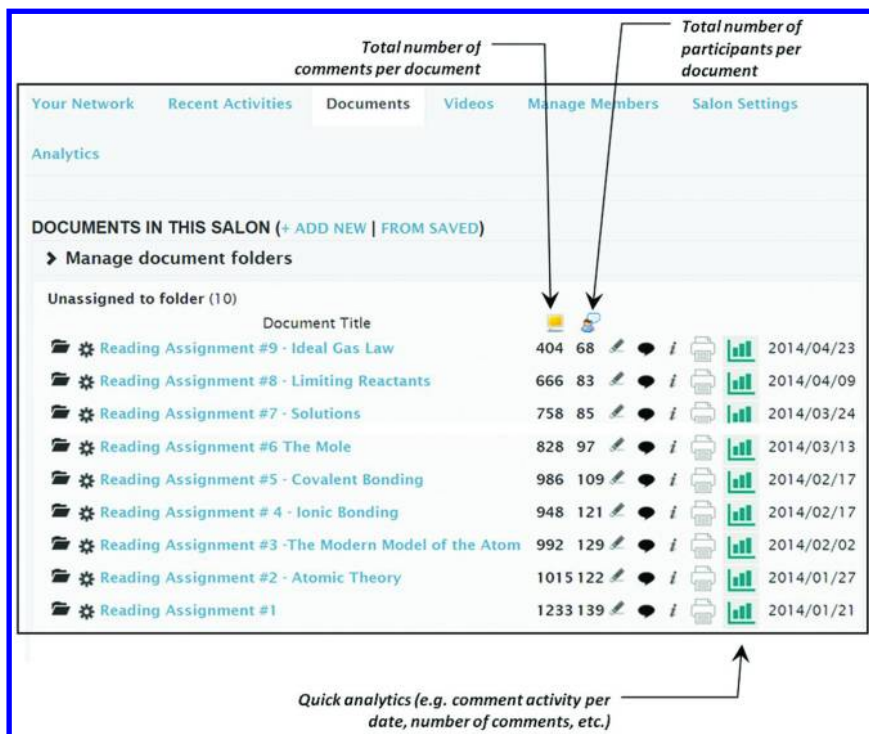


Figure 4. CLS screenshot: Quick access to Analytics.

In addition, clicking the bar graph icon (Figure 4) reveals more specific information regarding comment activity such as comment activity by document, date, or student (Quick analytics).

Filtering features (not shown) allows sorting annotations by tags so that student questions or discussion requests can be quickly accessed without scanning through hundreds of annotations.

For more comprehensive data access and analysis, comments and responses can be compiled into Excel or CSV format and downloaded (Figure 5).

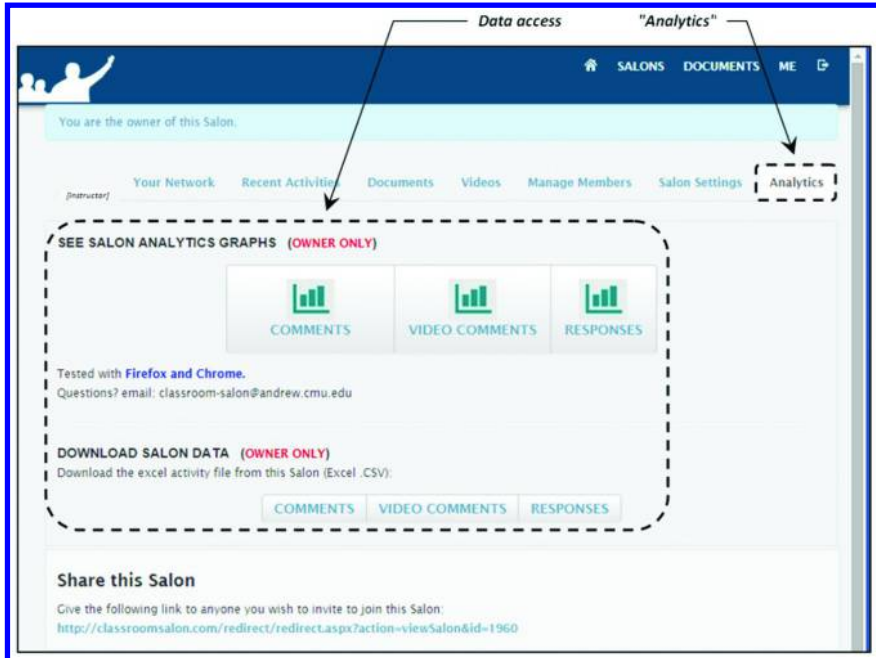


Figure 5. CLS screenshot (partial view): Clicking “Analytics” on the header bar allows complete data access.

### Challenge: Student Engagement in Large Classes

As mentioned before, CLS has initially been developed for the use in English writing classes, which traditionally are not large enrollment classes. Adapting CLS to large enrollment courses with up to 220 students opened up critical questions. How can a large number of students possibly annotate text, add comments, or pose and answer questions visible to all classmates and still be original and bring their personal views to the table? Will students embrace the idea of sharing their thoughts about science content? Will they engage in real discussions or just do the bare minimum of what they have been asked to do? Will the technology be user friendly? Was CLS another piece of technology not suitable for the challenges of a large class?

Online reading assignments outside the classroom do not fit the classic definition of active learning as being “any activity students do in a classroom other than just listening to the instructor’s lecture” ((7), p.189; (8), p.4). However, actively reading, annotating, evaluating, or discussing text with classmates gives students the opportunity to become more connected with the subject matter and promotes critical thinking. It also allows the instructor to learn about student questions and misconceptions and include these into instructional

feedback. Realistically, it is not possible for an instructor to read and evaluate each annotation in a large class, but even a quick look at students' comments will reveal a variety of questions and misconceptions that can be addressed immediately. It can be very challenging to incorporate any technology into a large classroom that will do more than just deliver homework questions or clicker answers, and there is evidence in literature ((8), p.20) that the use of active learning techniques can be quite challenging and "the instructor's ability to monitor student understanding seems to be inversely proportional to class size." Fortunately, some scholars "propose reframing lecture from a focus on the challenges of effectively teaching a large number of students to considering the lectures affording unique opportunities to promote active learning" (9). In this sense, CLS offers a unique platform for an open class-wide content discussion and provides the unique opportunity to collect an abundance of information about student learning and reflection. The intention of data analysis of students' comments, questions, responses, and discussion threads has to be determined by the course instructor or researcher. In case of the preparatory chemistry class described in this chapter, the course instructor utilized the program mostly to evaluate incoming student knowledge and misconceptions. The following will discuss the annotation analysis, instructional changes, and formative feedback based on the analysis results on student motivation and self-evaluation.

## **Large Classes**

It is not uncommon that mostly first-year students feel rather uncomfortable asking questions during lectures and therefore leave instructional periods with unanswered questions and misconceptions (10). This fact poses a great concern in large courses especially in natural and social sciences (11). CLS addresses this issue by creating a line of communication between students and instructor, which is essential for successful teaching and student understanding (12). In any classroom, instructors often follow their own concept of what they believe is successful teaching based on their knowledge and experiences. As we know, this does not necessarily mean that his or her teaching will also result in student learning. Chances are, most instructors have tried to recreate their lectures, changed textbooks, or incorporated new technology into their teaching in an effort to affect student learning positively. Some of these measures may or may not have shown some success. Effective instructional changes depend on both the teacher and learner. Unfortunately, one side of this equation is mostly unknown: instructors of large courses often do not know enough about the individual student's incoming knowledge to be able to tailor instructions to their needs.

## **Misconceptions – Existence and Identification**

Misconceptions or alternative beliefs about science concepts have been found to develop in learners intuitively long before they receive any type of formal science instruction (13, 14). Nieswandt found that learners develop ideas about science through interaction with their environment. Unfortunately, everyday experiences can provide evidence that supports incorrect assumptions (15). This

said, it is not unusual for 200 students in a college preparatory chemistry class to have a variety of different ideas about chemistry concepts. Chemistry concepts are often perceived as very complicated and difficult to understand. Most students entering college chemistry courses carry some misconceptions, some of which are even transferred from middle and high-school teachers (16). Misconceptions about general chemistry and other basic physical science concepts have been extensively researched (5, 17–21). Alarming, Halloun & Hestenes (22) found that when physics students were presented with experiences that challenged their beliefs, students were more likely to argue that outside laws or principles were interfering with the results rather than change their conceptions.

Active reading as part of the active learning strategy “What I Know – What I want to Know – What I Learned” (K-W-L) has been utilized to have students identify previous knowledge and consciously evaluate what they would like to learn and if learning has been successful. This strategy promotes active learning through reading, writing, discussions, and problem solving and engages students in higher-order thinking (23).

More traditional assessment of incoming student content knowledge includes placement exams, questionnaires, and quizzes whose results merely provide information about how students learn and what kind of misconceptions the learner may hold. Assessment instruments targeting misconceptions as described by Stein, Larrabee & Barmann (15), often require more time-consuming analysis. Pedagogical strategies like Just-in-Time-Teaching (JiTT), that deliver information about student learning more quickly, rely on what is called “Warm-up assignments” to collect student responses. Warm-up assignments have to be completed prior to lecture, and consist of a reading and writing assignment. The collected responses are evaluated to alter instruction according to the findings. According to Novak (24), this procedure creates a teaching-learning team of teachers and students, “making the lecture time as relevant as possible”. The purpose of this exercise is to design meaningful instruction tailored to the knowledge level of the student and has shown to improve student learning gains significantly (24).

Another important aspect of textual annotation is the interaction among a group of learners and the promotion of discussion and self reflection. Readers benefit from insight and perspectives of classmates (25). Shared annotation can promote collaborative learning and has been shown to improve reading comprehension compared to readers who annotate individually (26).

## **Feedback**

Feedback plays a very important role in knowledge development and skill acquisition (27, 28). In this context, it is important to realize that not all feedback has a positive effect on student learning. Studies, for example, have shown that feedback can also have either no effect or even debilitating effect on learning. Feedback can be negative, not specific, discouraging, or learning inhibitive (28). Therefore the specific method on how to provide feedback using the findings in CLS has been carefully implemented into the preparatory chemistry lecture and will be described in more detail below.

## Implementation into a Large-Enrollment Course

### Trial and Error Phase

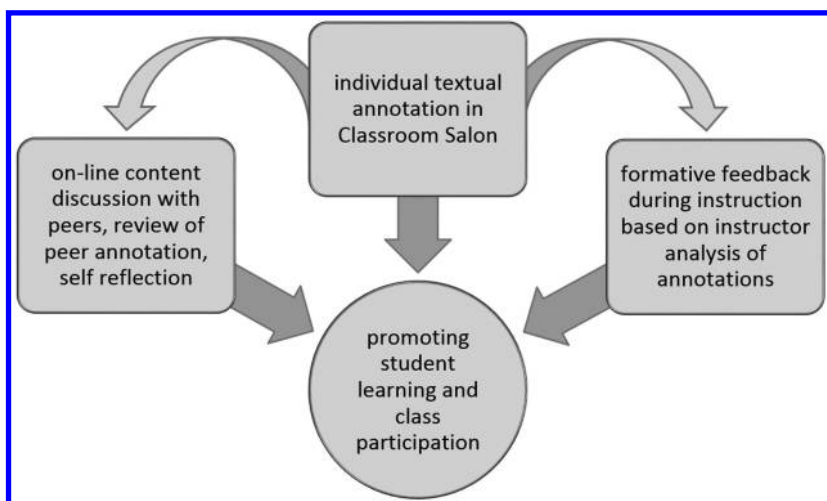
CLS's implementation into a chemistry classroom with more than 200 students has unintentionally gone through a semester-long trial phase. In retrospect, this time has been shown to be very informative, useful, and allowed for a smooth implementation in the following semesters. Student questions about CLS use have translated into support documents for students. These documents are now being handed out at the beginning of each semester and always include the course specific Salon link. Because participation in CLS is part of their course grade, students inquired about the number of comments the instructor expected from each student. This information is now being routinely posted on the course Salon homepage. During the trial stage it has also been found that the number of members per Salon plays a role when it comes to student participation. The initial idea to set up Salons according to each of the 10 or 11 discussion sections with 18-20 members has been deserted in favor of a large, class-wide Salon with up to 220 members due to the fact that the percentage of participating students in large Salons has shown to be higher compared to small Salons. The reason for this has not yet been clearly identified. Possible reasons could be that either students prefer the anonymity of a larger class when annotating or—even simpler—that a larger number of comments lead to more diverse and interesting discussions. It can be argued that the prior option is a phenomenon more prevalent in classes perceived as more difficult, such as science or math classes, but in regards to CLS, we have not yet studied this effect any further.

The trial semester was also used to select the best way to place content questions in CLS. The program allows the instructor or researcher to pose questions about the text in a separate window that can be accessed by clicking the tool bar on top of the screen. This method required students to look for the questions first under the “Question” tab, then type the response. This process seemed difficult for inexperienced CLS users, and we were able to increase the number of responses by embedding the questions directly into the documents. This way Salon members just highlighted the question and typed the answer in the comment box like they would do when annotating the text.

### Methodology – Final Implementation (Figure 6)

CLS has been utilized since spring 2012 in selected preparatory chemistry courses at UWM. Student enrollment typically ranges from 180 to 220 in these classes, and UWM offers as many as four sections per semester. The prerequisite for this course is intermediate algebra. Each course consists of three 50-minute lectures and one 50-minute discussion per week. Depending on class enrollment, each lecture section can have up to eleven discussion sections with a maximum of 20 students per section. The discussion sections are led by teaching assistants. Student class performance is measured through exams (three midterm exams

and the final exam), lecture attendance, on-line homework, and discussion. The discussion grade represented about 10% of the available total points and included both, participation during the actual face-to-face discussion and participation in CLS, in equal parts. The discussion participation was evaluated by teaching assistants; the course instructor accessed CLS analytics and downloaded the number of annotations per students. Students were graded entirely on participation in CLS, not on accuracy of answers or quality of annotations. All students in the targeted courses have been given the option to participate in the research study and only those who consent to participate via an approved consent form have been included in the study per IRB protocol.



*Figure 6. Implementation of CLS in Preparatory Chemistry course with student learning as mutual goal.*

### *Student Task in Classroom Salon*

Throughout the semester, students were asked to read and annotate nine 3-5 page long textbook excerpts in CLS (Reading Assignments 1 through 9) (in agreement with the publisher) as preparation for upcoming lecture content. The reading assignments were not meant to replace the entire course textbook. The course instructor still incorporated many other parts of the textbook into lecture and discussion. Once the Salon was created, students received detailed instructions for log in and CLS use. The course instructor then posted the reading assignments (RA) consecutively in the order of lecture coverage. The time between postings was on average 7-10 days. The editing features in CLS

allowed deletion of unnecessary content from the textbook selection, such as practice problems and addition of questions and tags. The instructor created three tags: “General” for general comments, “Important”, and “Discuss in Class”. Students were asked to either choose from this selection or create their own tags to label their comments. In regards to the expected number of annotations, each student was asked to annotate five text passages (including posted questions) and reply three times to other comments per reading assignment. Text-embedded questions were worded to target and encourage critical thinking and could not be answered by rote repetition of text passages or phrases. The timeframe of 7-10 days gave students enough time to annotate and also engage in CLS discussion with their classmates. While many student questions were answered by their classmates through replies and CLS discussions, answers to the posted questions and or unanswered student questions and requests were addressed by the course instructor during lecture.

It was found that some students did not feel comfortable to comment on their classmates’ annotations or join in lecture discussions in the beginning of the semester. This behavior however changed once students realized that their annotations and replies were not graded or evaluated in class.

The selected course content areas discussed in CLS included:

1. *Physical Properties*
2. *Atomic Theory*
3. *Modern Model of the Atom*
4. *Ionic Bonding*
5. *Covalent Bonding*
6. *The Mole*
7. *Solutions*
8. *Limiting Reactants*
9. *Ideal Gas Law*

Following are some examples of questions from different Reading Assignments (RA):

**RA 1 (Physical Properties):** *“Why do different solids have different densities?”*

**RA 2 (Atomic Theory):** *“Why did Ernest Rutherford choose gold for his experiment (Gold-Foil experiment leading to the discovery of the atomic nucleus). Could he have used a different material?”*

**RA 8 (Limiting Reactants,** in reference to images showing iron metal emerged in copper(II) chloride solution before and after reaction has occurred): *“Which compound is the limiting reactant in this reaction? Do you have enough information to determine the limiting reactant through calculation? Does the image provide a visual clue?”*



## Findings

The average number of annotations per reading assignment show that participating students on average annotated more frequently than they were instructed. Students also took the opportunity to create their own tags, such as “Conversion”, “Must know”, “Key”, or “Question”. The course instructor checked the annotations of each assignment for questions, comments, and misconceptions prior to lecture in which the content of the assignments was discussed. The instructor’s evaluation did not just concentrate on content, but also included aspects such as use of scientific terminology or evidence of student sentiments or concerns. The following selected annotations are taken from these categories.

1. *Some student comments showed the existence of common misconceptions, for instance:*

### **Misconceptions regarding the properties of water:**

- “The reason why ice floats (on water) is because water becomes less dense, is it?”
- “Ice is denser than water, because under cold conditions, H<sub>2</sub>O molecules expand. This expansion naturally increases the volume, and thus decreasing the density. That is why ice is able to float on water.”
- “Ice floats on water and it must have a lesser density. Air fills the space between water molecules.”
- “Air, perhaps? Which would make the substance “lighter” than the surrounding liquid form.”
- “A water molecule does not have the same density as a drop of water because the mass and volumes are different measurements.”

### **Discussing properties of the element gold:**

- “Gold has a large atom size which makes it easier to conduct electricity.”

### **Discussing atomic theory:**

- “The location of an electron is important because if it exists in a higher energy level this means that’s it is in an excited state rather than a relaxed state. It does play a role in reactivity because the farther the electron from the nucleus the less energy it takes to separate it. This creates isotopes and compounds.”

### **Discussing ionic and covalent bonding:**

- “Nonmetals are less electronegative therefore they gain electrons in order to become noble gases while metals lose electrons in order to become noble gases.”
- “Is this right? - Electronegativity: An atoms tendency to gain electrons. Effective Nuclear Charge: The total charge of an atom.”
- “The more valence electrons an element has, the greater the conductivity.”
- “There is a type of covalent bond, called polar covalent bonds, which are important because these kinds of bonds allow the formation of another kind of weak bond, a hydrogen bond. Water is an example of a molecule that has polar covalent bonds and engages in hydrogen bonding.”

### **Discussing reactions in solutions:**

- “The solution works as a catalyst.”
- “Reactions in solutions are often faster because the ionic atoms in a liquid are not stable and so they easily are able to break and break down any compounds that are put in the solution.”

### **Discussing gaseous reactions:**

- “Gaseous molecules are much more spread out and so when two gasses are reacted with each other they will get almost always get near 100% yields.”

### 2. *In addition, annotations showed incorrect or unsure use of scientific terminology:*

- “Fluorine atom needs 1 (electron) to fulfill itself under the octet rule | Nitrogen needs 3 to fulfill itself under the octet rule | In order for this to happen Nitrogen would need to bond to 3 Fluorine atoms in order to gain the 8 electrons it needs.”
- “The mole allows chemists to count atoms as accurately as possible.”
- “Is this is what is meant by conservation of matter? The two sides started with a specific number of atoms and the final outcome has to be that same # of atoms? (discussing limiting reactants).”
- “Electronegativity gives electrons mobility. Electronegativity is the ability of atoms to attract electrons because the nucleus has a positive charge. Elements can gain or lose electrons to be more like a noble gas.”
- “Hydrogen is an isotope and that is why it has multiple energy levels.”
- “The closer an electron orbits around the nucleus the less energy it uses.”

3. *Or questions, some of which were rather unexpected:*

- “Is color a chemical property?”
- “How does density change the volume of an object?”
- “What are the relationships between mass and volume? I thought they refer to the same thing, and one applies gravity to find weight.”
- “Can how fast or slow the molecules move have an effect on density?”
- “Does the density of water (in solid form) have anything to do with the polarity of H<sub>2</sub>O?”
- “Does density depend on the closeness of atoms?”
- “In order for atoms (and, consequently, molecules) to possess the chemical properties of a given element, there has to be at least two atoms bonded together, right? What I mean is, a single “carbon” atom would not have the properties of carbon, would it? Or does a single atom from any element have the same characteristics?”
- “How are atoms formed or created?”
- “Do all elements have more than one orbital?”
- “Do orbitals change shapes when they gain more energy?”
- “How are the orbital diagrams different for different elements?”

4. *Even student concerns or attitudes could be captured...*

*...Like the fear of science:*

- “I agree (with my classmate). I think it (the mole) may be too complicated for the common person to use, but for other scientific uses it could be helpful for large numbers of things.”
- “That is a good thought, but I am not sure myself.”

*...Or enthusiasm:*

- “It’s cool how we now know how atoms work!”

### *Instructional Changes Based on Classroom Salon Findings*

The revelation of student misconceptions, questions, and comments does not necessarily have to change classroom instruction entirely and is really determined by the intent to integrate CLS in the classroom. In case of the preparatory chemistry classroom, it was decided to address some issues differently. For instance, questions and misconceptions that were detected in annotations that seemed to be held by a larger number of students were regularly incorporated in lecture. This was done either in the form of short, purely instructional periods or in open class discussions. Class discussions were started by clicker questions in the beginning of class or by instructor comments.

Upon instructor evaluation, some isolated findings were chosen to be addressed in person, either by the instructor or a teaching assistant. Students, whose annotations revealed that they were frustrated with the CLS technology or the course material, were contacted to set up short CLS training sessions or encouraged to seek help from the instructor, teaching assistant or course tutor.

### *Class Instructor Feedback*

It was very important for the course instructor to create a positive relationship with the students, which is why findings from CLS were addressed in a supporting and non-evaluating manner. Student names were never revealed during in-class discussion and every question was addressed with the same level of importance. In order to give the pre-instructional reading assignments further meaning and to encourage student participation in CLS, students were made aware that a significant part of the lecture was based on their questions and annotations made in CLS. Knowing that their comments matter and their course instructor continuously strove to meet their needs, changed the class climate significantly in comparison to previous semesters. Students started to ask questions in lecture more often and contributed to spontaneous class discussions. As an additional result, students felt more comfortable using proper terminology in class and one instructor observed an increased occurrence of “smarter questions”, which was interpreted as evidence of the development of higher-order thinking skills.

### *How Did Students Like the Use of Classroom Salon?*

CLS has been utilized in four lecture sections of preparatory chemistry courses with a total of approximately 850 students. On average, about 90 percent of students in these classes participated in CLS. Student feedback in lecture revealed that students liked the accessibility of the reading assignment from their various electronic devices (laptop, iPad, or smart phone). Like every new technology in the classroom, detailed registration information and instructions on how to use the program were very important for users. Instructional videos available on the program site provided by the CLS team were very useful as well.

Responses given in a student survey described the integration of CLS into instruction as predominantly positive. Asked if they liked the use of CLS and seeing their classmates’ comments on the site, students responded:

- “Yes. Seeing my classmates’ responses made me aware of things I didn’t pick up on at first. In addition, it provided me with insight on how other people are thinking about the text.”
- “Yes, because it gave you a perspective from other people to why they thought a specific point was important and it allowed you to have a discussion with your classmates.”
- “It was nice because it was usually a good jumping off point and made me think further.”

## Discussion and Conclusions

The implementation of CLS in the preparatory chemistry course has been very easy because it did not require changes in course structure or curriculum and opened up more opportunities for instructor and students than initially anticipated. Increased communication and more tailored instruction that addressed actual student questions and needs changed the class climate significantly. Students seemed more open to participate in class discussions and asked content related questions in class more frequently than in previous semesters. CLS gave them a voice in the class, and their questions and comments were acknowledged. This realization has also positively influenced students' perception of the instructor role: departmental course evaluation surveys given at the end of the semester revealed that a higher percentage of students believed that the course instructor's intent is to help them succeed in the course. Given all the benefits, it is planned to incorporate CLS permanently into preparatory chemistry courses at UWM. CLS allowed collection of an abundance of student data through annotations and is currently being evaluated to investigate the effect of CLS on student course performance.

### Future Directions

Because CLS can have many different applications depending on instructor intent and course implementation, it can be incorporated into traditional classrooms as well as in flipped and blended courses.

CLS has shown to increase students' engagement with subject matter and also positively affect class communication. In the future, CLS is planned to be used for identifying at-risk students at UWM. Data collected in CLS will undergo linguistic analysis to identify what has been described in literature as "positive" or "negative" comments and also content-related performance through linguistic coding (29). Word clusters associated with emotions or specific content concepts can then be defined and identified with pattern and word matching software (DocuScope). The results of the linguistic analysis in combination with the already existing analytical features will be evaluated in the context of early identification of low-performing students.

### Acknowledgments

The author would like to thank the CLS team at Carnegie Mellon University, especially Dr. Ananda Gunawardena, for their support. Their continuous effort to provide instructors and students with an innovative, research-based educational social platform that serves students and instructors equally is magnificent.

### References

1. Davis, B. G. *Tools for Teaching*; Jossey-Bass: San Francisco, CA, 2001.
2. Forsyth, D. R. *The Professor's Guide to Teaching: Psychological Principles and Practices*; American Psychological Association: Washington, DC, 2003.

3. Isbell, L. M.; Cote, N. G. Connecting with struggling students to improve performance in large classes. *Teach. Psychol* **2009**, *36*, 185–188.
4. Piaget, J. *Six Psychological Studies*; Tenzer, A., Translator; Vintage Books: New York, 1968.
5. Taber, K. *Chemical Misconceptions, Prevention, Diagnosis and Cure: Theoretical Background*; Royal Society of Chemistry: London, 2002; Vol. 1, pp 53–66.
6. Kaufer, D.; Gunawardena, A.; Tan, A.; Cheek, A. *J. Bus. Tech. Commun.* **2011**, *25*, 299.
7. Fritz, M. Using a Reading Strategy to Foster Active Learning in Content Area Courses. *J. Coll. Reading* **2002**, *32*, 189–194.
8. Paulson, D. R.; Faust, J. L. Active Learning in the College Classroom. *J. Excellence Coll. Teach.* **1998**, *9*, 3–24; [www.calstatela.edu/dept/chem/chem2/Active/main.htm](http://www.calstatela.edu/dept/chem/chem2/Active/main.htm) (June 2014).
9. Winstone, N.; Millward, L. Reframing perceptions of the lecture from challenges to opportunities: Embedding active learning and formative assessment into the teaching of large classes. *Psychol. Teach. Rev.* **2012**, *8*, 31–41.
10. Stone, E. Students' attitudes to the size of teaching groups. *Educ. Rev.* **1970**, *21*, 98–108.
11. Bowers, J. Classroom communication apprehension. *Commun. Educ.* **1986**, *35*, 372–378.
12. Scott, P. H.; Mortimer, E. F.; Aguiar, O. G. The tension between authoritative and dialog discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Sci. Educ.* **2006**, *90*, 605–631.
13. Nieswandt, M. Von Alltagsvorstellungen zu wissenschaftlichen Konzepten: Lernwege von Schülerinnen und Schülern im einführenden Chemieunterricht. *ZfDN*, 2001, *7*, 3352; [ftp://ftp.rz.uni-kiel.de/pub/ipn/zfdn/2001/S.33-52\\_Nieswandt\\_2001.pdf](ftp://ftp.rz.uni-kiel.de/pub/ipn/zfdn/2001/S.33-52_Nieswandt_2001.pdf) (June 2014).
14. Mulford, D. R. An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Ed.* **2002**, *79*, 739–744.
15. Stein, M.; Larrabee, T.; Barman, C. A study of common beliefs and misconceptions in physical science. *J. Elem. Sci. Educ.* **2008**, *20*, 1–11.
16. Kruse, R. A.; Roehrig, G. H. A Comparison Study: Assessing Teachers' Conceptions with the Chemistry Concepts Inventory. *J. Chem. Ed.* **2005**, *82*, 1246–1250.
17. Nakhleh, M. B. Why some students don't learn chemistry? Chemical misconceptions. *J. Chem. Educ.* **1992**, *69*, 191–196.
18. Bowen, C. W.; Bunce, D. M. Testing for Conceptual Understanding in General Chemistry. *Chem. Educ.* **1997**, *2*, 1–17.
19. Stavy, R. Conceptual development of basic ideas in chemistry. In *Learning Science in the Schools: Research Informing Practice*; Glynn, S., Duit, R., Eds.; Lawrence Erlbaum: Hillsdale, NJ, 1995; pp 131–154.
20. Gabel, D. L.; Bunce, D. M. In *Handbook of Research on Science Teaching and Learning*; Gabel, D., Ed.; Macmillan: New York, 1994; pp 301–326.
21. Stavy, R. J. Using analogy to overcome misconceptions about conservation of matter. *J. Res. Sci. Teach.* **1991**, *28*, 305–313.

22. Halloun, I. A.; Hestenes, D. The initial knowledge state of college physics students. *Am. J. Phys.* **1985**, *53*, 1043–1048.
23. Bonwell, C. C.; Eison, J. A. Active learning: Creating excitement in the classroom. *ASHE-ERIC Higher Education Report No. 1*; The George Washington University: Washington, D.C., 1991; pp 1–104; <http://files.eric.ed.gov/fulltext/ED336049.pdf> (June 2014).
24. Novak, G. M. In *New Directions for Teaching and Learning*; Buskist, W., Groccia, J. E., Eds.; Wiley Periodicals: New York, 2011; Vol. 128, pp 63–73.
25. Wolfe, J. L.; Neuwirth, C. M. From the margins to the center: The future of annotation. *J. Bus. Tech. Commun.* **2001**, *15*, 333–371.
26. Johnson, T. E.; Archibald, T. N.; Tenenbaum, G. Individual and team annotation effects on students' reading comp G. Individual and team annotation effects on students' reading comprehension, critical thinking, and meta-cognitive skills. *Comput. Hum. Behav.* **2010**, *26*, 1496–1507.
27. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112.
28. Shute, V. Focus on Formative Feedback. *Rev. Educ. Res.* **2008**, *78*, 153–189.
29. Taguchi, N.; Kaufer, D.; Gomez Laich, M. P.; Zhao, H. *A corpus-linguistic analysis of online peer commentary*. Presentation at the 19th International Conference on Pragmatics and Language Learning, Indiana University, Bloomington, April 24–26, 2014.

## Chapter 11

# Developing the First Online General Chemistry Laboratory Exam

Jimmy H. Reeves\*,<sup>1</sup> and Deborah Exton<sup>2</sup>

<sup>1</sup>Department of Chemistry and Biochemistry, University of North Carolina  
Wilmington, 601 S. College Road, Wilmington, North Carolina 28403

<sup>2</sup>Department of Chemistry and Biochemistry, 1253 University of Oregon,  
Eugene, Oregon 97403-1253

\*E-mail: reeves@uncw.edu.

The explosion of resources made possible by the Internet and fueled by publishers and funding agencies has the potential to change chemistry instruction. Internet-based assessments may prove to be effective ways to evaluate learning in this brave new world. It is our belief that these new forms of assessment will eventually replace the black and white, paper and pencil multiple-choice exams so ubiquitous in large-enrollment chemistry courses today. In 2008, the American Chemical Society (ACS) Examinations Institute formed the first exam committee charged with developing a fully online exam, the General Chemistry Laboratory Exam. The exam employs digital video and question types such as “drag and drop” and “choose all”, and emphasizes understanding of experimental design as well as calculation, data interpretation, and lab technique. In this chapter we will describe the process of creating the exam and discuss the lessons learned about the current benefits and limitations of web-based testing.

### A Laboratory Exam? Why and Why Not

Though a required component of nearly every general chemistry course, the laboratory has suffered severe assessment neglect, especially in terms of nationally-normed standardized exams. For example, none of the seventeen exams previously offered by the ACS Examinations Institute in the “general



chemistry” category focus on assessment of the laboratory (1), and while a series of small scale laboratory activities were published by the Examinations Institute in 1994 (2), they are a “non norm-referenced exam product” and thus provide no information about student performance in comparison to other exam takers across the nation.

There are numerous reasons that nationally normed exams have historically been unavailable for the laboratory, chief among them the fact that laboratory skills are difficult to assess using paper and pencil exams, especially those in multiple-choice format. Nearly all national testing services rely heavily on these question formats for their exams, even when the exams are provided online. To assess the laboratory effectively, a more robust testing alternative has to be employed.

One possibility that merits consideration is the laboratory practical exam. Based on direct observation of students while they perform experiments, this approach can be an effective way to assess laboratory skills. However, laboratory practicals are difficult and time consuming to administer and grade, particularly in large enrollment classes. Moreover, assessment at a national level is complicated by the wide variation in the judgments of the instructors who grade them, even when detailed rubrics are provided. Thus, establishing a national “norm” and other descriptive statistics for a laboratory practical exam is problematic at best.

With these obstacles blocking its development, it’s easy to see why a nationally normed laboratory exam has been so long in coming. However, the need for such an exam goes well beyond the fact that the laboratory typically produces 25% of the course credit hours and should be appropriately assessed. There is ample evidence that chemists are emotionally tied to the notion that the laboratory experience is central to learning chemistry. In their recent paper presenting the results of a national survey of laboratory goals in undergraduate chemistry ((3), p 685), Buck and Towns write, “Nearly all faculty agree that laboratory is a vital component of the chemistry undergraduate curriculum; however, the explicit articulation of goals and aims within the literature is vague.” According to Reid and Shah ((4), pp 173-174), “Laboratories are one of the characteristic features of education in the sciences at all levels.... However, very little justification is normally given for their presence today.” Perhaps the most challenging language comes from Lagowski et al. ((5), p 145),

“Precious little direct evidence exists that [laboratory instruction] provides a useful function in the way students learn chemistry. In spite of this anomalous behavior of scientists in their apparent lack of interest in data supporting their behavior, most academic chemists, and indeed industrial chemists also, will swear fealty to some concept relating laboratory instruction to the formal teaching of chemistry, especially at the undergraduate level.”

Despite these findings, there continues to be significant support for the benefits of the laboratory in chemistry instruction. In response to the concern that virtual chemistry laboratory simulations widely available on the Internet (6) might replace hands on laboratory activities in some chemistry courses, the American Chemical Society issued a Public Policy Statement (2011-2014) (7) which begins:

“Hands-on activities enhance learning significantly at all levels of science education. (See 1, 2, and 3 in reference). These activities are usually the basis for a laboratory class or laboratory portion of a class. In a hands-on chemistry course, students directly experience laboratory chemicals and their properties, chemical reactions, chemical laboratory apparatus, and chemical laboratory instruments. These activities are essential for learning chemistry.”

So who is right? Asserting that “Hands-on activities enhance learning significantly at all levels of science” (7) begs the question of how those enhancements are to be measured on a case by case basis. It’s highly unlikely that all laboratory experiments regardless of design “enhance learning significantly.” If laboratories are indeed “essential for learning chemistry,” shouldn’t assessing them to determine which designs are effective be an essential component of an overall course evaluation?

Evidence has been published demonstrating improvements in critical thinking skills when active learning strategies that promote inquiry based discovery (8), cooperative-based lab instruction (9) or critical observation and reflection (10) were employed. Nevertheless, a more robust assessment approach that applies across all teaching strategies and allows a comparison of student performance to a national sample of test takers is needed. As scientists we must put an end to our “anomalous behavior” and replace “sworn fealty” with meaningful evidence of the benefits (or lack thereof) of the instructional approaches currently utilized in our laboratories.

But effective assessment is not possible without well-constructed learning goals and learning goals for laboratories “are often criticized as being poorly articulated or nonexistent” (11). Nevertheless, a recent analysis of interviews with forty faculty responsible for laboratory programs at a variety of educational institutions revealed common cognitive and psychomotor goals such as critical analysis, conceptual understanding, learning laboratory techniques and using laboratory equipment (12). These goals fit well with question categories that were proposed by attendees of a presentation by Tom Holme, Director of the ACS Examinations Institute, at the 2007 ACS National meeting in Boston (13). The prioritized list is presented in Table 1.

**Table 1. Consensus of Question Categories for a Laboratory Exam**

calculation and data interpretation

lab technique

lab content

experimental design

safety

For example, critical analysis is required for appropriate data interpretation and experimental design requires conceptual understanding of both the problem to be solved and the appropriate experimental techniques needed to solve it. That the participants were able to agree on these categories was evidence that an exam built around them could be an effective assessment tool. Buoyed by this result, the authors volunteered to co-chair a nine member exploratory committee to examine the feasibility of developing a general chemistry laboratory exam. The members of this committee are designated with asterisks in Appendix 1.

## Exploring the Possibilities

The goals of the exploratory committee were to identify the experiments most commonly used in general chemistry courses; determine the skills and techniques the students are expected to master, and consider the resources required to exploit the online environment. If sufficient commonality could be demonstrated among the experiments offered and the techniques emphasized by different laboratory programs, a national exam would be feasible. To address the first two issues, a total of thirty seven laboratory manuals, used by over 1000 schools from around the country, were collected and reviewed. They ranged from manuals developed and distributed by major publishers to those authored in-house and custom published, shown in Figure 1.



*Figure 1. A sampling of the laboratory manuals reviewed to identify common experiments.*

Sixteen experimental topics and thirteen techniques appeared (in various forms) in most of the manuals, and these were used to develop a web-based survey. In this survey, participants indicated their preferences for the topics and experimental techniques to be included in an ACS laboratory exam by responding *definitely yes*, *probably yes*, *probably no*, or *definitely no* to each of the topics and techniques listed. Similar input was solicited from individuals who participated in a workshop that was conducted at the Biennial Conference in Chemical Education (BCCE) held at Indiana University in 2008. A total of 125 instructors participated in the survey, and based on their responses the topics and techniques were rank ordered in terms of their perceived importance in the general chemistry laboratory curriculum. The resulting lists are presented in Tables 2 and 3. Survey results also indicated that faculty were interested in using the lab exam for program, course and student assessment, that a clear majority (74%) favored the inclusion of laboratory practicals with the exam and that two thirds favored “an online format that may include multiple question types, videos, animations and simulations” over “traditional multiple choice” in either paper and pencil or web-based format.

**Table 2. Priority of Topics to Be Included in a General Chemistry Laboratory Exam (Ranked Highest to Lowest)**

Volumetric analysis (titrations)  
Stoichiometry  
Kinetics (determination of the rate law)  
Spectrophotometry/Beers Law  
Properties of Acids and Bases  
Calorimetry  
Gas Laws  
Le Chateliers Principle  
Density  
Acids and Bases Determination of  $K_a$   
Determination of Equilibrium Constant  
Molecular Models (not computer modeling)  
Qualitative Analysis (general, not qual schemes)  
Synthesis  
Qualitative Analysis (qual schemes)  
Spectroscopy/Atomic Emission

**Table 3. Priority of Techniques To Be Included in a General Chemistry Laboratory Exam (Ranked Highest to Lowest)**

appropriate use of glassware  
solution preparation  
use of buret  
graphical analysis/calibration curves  
use of pipet  
use of analytical balance  
weighing by difference  
identification of standard glassware  
filtration  
dilutions  
gravimetric analysis  
gas collection (displacement of water)  
thin layer/paper chromatography

After considering all of the information gathered by the committee, the following conclusions were reported to the Director of the Examinations Institute:

- Data from surveys and the workshop confirmed substantial similarity among lab programs
- Feedback indicated a significant demand for the exam
- Lab practical exams are considered to be a valuable means of assessment, but variability in instructor grading makes any statistical analysis problematic
- Computer technology that includes digital video has reached a point where it is now feasible to provide an alternate form of laboratory practical assessment.

In August 2008, the decision was made to develop the first nationally-normed ACS General Chemistry Laboratory Exam designed to be delivered on-line and to also include a hands-on lab practical component.

### **The Work of the Exam Committee**

The committee members that developed the exam are listed in the Appendix. They included faculty from Research 1 universities (n = 6), comprehensive universities (n = 6), private colleges and universities (n = 4) and community

colleges ( $n = 3$ ). At the first meeting of the committee, held at the Salt Lake City ACS meeting in spring, 2009, it was decided that the exam would include a computer-graded virtual component for which national statistics would be computed, and a laboratory practical component for which no national data would be collected.

At this initial meeting, the committee also decided that the virtual exam would be web-based, consisting of a series of scenarios that utilize digital media and a variety of question types to create a simulated laboratory practical that also probed conceptual understanding. The challenges and expenses associated with designing this first of its kind on-line exam imposed limitations on the number of scenarios that could be developed. In the end, the committee chose the first six experimental topics listed in Table 2 as scenario candidates. The committee also decided that all of the categories listed in Table 1 should be addressed in each scenario because the choice of scenarios to be included in any given exam would be left to the instructor. Over the next year, committee members worked in teams of two or three to develop storyboards and compile lists of required videos and images. Storyboards and questions were continually discussed and refined by the entire committee at no less than eight Exam Committee meetings. In May 2010, the videos and images were created in a laboratory at the University of Oregon by Will Doolittle of *Moving Image Productions*. Once the scenarios and media were developed, programmers from *Metior Inc.*, the company hired to produce the online versions of ACS exams, worked with committee members to create the online exam.

In spring, 2011 two versions of each of the six scenarios were made available for testing. Versions differed either by the phrasing of a given question or the question type employed. Ten schools and over 1400 students participated in the trial testing, with each school choosing the scenarios they wished to test. As is the practice for all examinations developed by the Examinations Institute, results were tabulated based on question *difficulty* and *discrimination*. *Difficulty* is defined in this context as the fraction of correct responses, while *discrimination* is the difference between the number of correct responses of the top 25% of students (as measured by their overall exam score) and the bottom 25%, divided by the number of students in the top (or bottom) 25%. Thus, *discrimination* ranges from 1.0, indicating that all of the students in the top 25% got the question correct and all of those in the bottom 25% got it wrong, to -1.0, indicating the reverse. Using these measures as a guide, the scenario versions were merged and the final version of the exam was created at the BCCE at Penn State in summer, 2012. The average values for difficulty and discrimination for each of the scenarios are presented in Table 4.

These values fall within the guidelines recommended by the Examination Institute (0.3 – 0.85 for difficulty and greater than 0.25 for discrimination), although a small number of questions fell outside the ranges. Exceptions were made in these cases to ensure that questions assessing important topics such as experimental design (which consistently proved difficult for students) were represented or that key concepts associated with a given experiment were addressed. A practice scenario that provides representative media and question types was also created and is currently available (14). A screenshot of an experimental design question from this scenario is provided in Figure 2.

**Table 4. Average Difficulty and Discrimination Scores for Questions Chosen for the Final Version of the General Chemistry Laboratory Exam**

<i>Scenario</i>	<i>Number of Students</i>	<i>Average Difficulty</i>	<i>Average Discrimination</i>
Calorimetry	936	0.50	0.42
Stoichiometry	357	0.62	0.36
Acid Base (Qual Analysis)	337	0.57	0.51
Spectrophotometry	1275	0.61	0.43
Kinetics	1157	0.53	0.45
Volumetric Analysis	1464	0.54	0.43

*Figure 2. Screen shot of an experimental design question from the “Determining Density” practice scenario.*

The Laboratory Practicals portion of the exam was developed and tested by five committee members and is focused on getting the job done, (i.e. demonstrating mastery of experimental techniques and the ability to follow a detailed procedure, not probing conceptual understanding. Practical exams on

solution preparation, spectrophotometry, and titration, complete with student materials, instructor and preparation guidelines and grading rubrics, were produced and will be available from the Examinations Institute (1).

## The Final Product

In its final form, the online portion of the ACS General Chemistry Laboratory Exam consists of six scenarios, most with accompanying videos. *Calorimetry*, *Stoichiometry* and *Descriptive Acid Base* cover material typically taught in the first semester of a traditional two semester general chemistry course, while *Spectrophotometry/Beers' Law*, *Kinetics* and *Volumetric Analysis* are primarily focused on second semester material. The scenarios to be tested are chosen by the laboratory instructor and are designed to require a maximum of 25 minutes each to complete. The exam, which is currently available on a limited basis, must be administered in a secure environment and monitored using software provided by *Metior*. Based on a schedule provided by the instructor, exams are made available in a limited time window to computers with secure IP addresses and students are issued a code to access their individual exams. Proctors are able to monitor the progress of the students in real time and provide extra time as needed or reinstate students who have been incorrectly logged out. Feedback from instructors who have given the exam indicates that students experience little difficulty interacting with the exam interface, viewing the videos, and/or finishing in the allotted time. Because grading is automatic, a wealth of data is collected continuously, providing the possibility of in-depth analysis. Instructors receive a full report of their students' performance, along with nationally-normed data that can be presented to administrators and other officials interested in curriculum and programmatic assessment. Security issues associated with access to web-based exams require that exam questions only be available on-line during the designated, limited exam period. A summary version on paper of the exam may be made available to instructors who want more detailed information about their students' performance. The ACS Examinations Institute staff and the Board of Trustees are developing policies that provide the maximum access to information possible while adhering to critical security requirements.

## Lessons Learned

Although the ACS Examinations Institute is in the process of providing its traditional multiple choice exams in an online format, the General Chemistry Laboratory Exam is the first to be designed from its conception to be delivered online. Indeed, the exam is so novel that it initially encountered copyright difficulties because the US Copyright Office had never before issued one for an exam that contained digital video. And like any new endeavor, the originality of the exam was limited by the tools available to create it. The software platform provided by *Metior* was originally designed as an online homework system with independent questions offered in a multiple-choice format. Additional question formats such as "choose all" or "drag and drop" had to be developed as add-ons



to the original framework. Within these design limitations there was no way to implement questions that built on the response of a previous question or to create a qualitative analysis scenario that would allow the student to choose which experimental measurements (such as pH) to consider first, with more credit awarded for correctly identifying unknowns using fewer measurements. It is hoped that future iterations of fully online exams will push the envelope of creative methods to provide more effective assessment of students educated in the digital age.

Finally, it is important to emphasize the lesson that providing secure online exams involves significant legal and logistical challenges. Carving out the Examinations Institute's place in this brave new world of online testing will involve significant time and financial resources, but the benefits show promise to be enormous. To be able to assess the learning that occurs in the wired world of interactions, simulations and videos that is quickly becoming our students' learning environment, we must utilize the tools that only the Internet can provide. This exam represents a first step in that exciting endeavor.

### Appendix

#### *Members of the ACS General Chemistry Laboratory Exam Committee*

<b>Name</b>	<b>Institution</b>
Pia Albuquerque	Grambling State University
Margaret Asirvatham *	University of Colorado at Boulder
Katherine Bichler *	Concordia University Wisconsin
Mark Cannon	BYU Hawaii
Jen Civelli	Cape Fear Community College
Deborah Exton *	University of Oregon
John Gelder *	Oklahoma State University
Tom Greenbowe *	Iowa State University
Joe March	University of Alabama Birmingham
Anne-Marie Nickel *	Milwaukee School of Engineering
Jason Powell	Ferrum College
Bob Pribush	Butler University
Jimmy Reeves *	University of North Carolina Wilmington
Sherill Soman *	Grand Valley State University
Brooke Taylor *	Lane Community College
Bill Vining	SUNY – Oneonta
Scott Von Bramer	Widner University
Gabriela Weaver	Purdue University
Dave Wilson	Parkland Community College in Champaign

\* indicates a member of the original Exploratory Committee.

## References

1. ACS Examinations Institute website homepage, URL <http://chemexams.chem.iastate.edu/exams> (September 17, 2014).
2. Laboratory Assessment page at the ACS Examinations Institute website, URL <http://chemexams.chem.iastate.edu/laboratory-assessment> (September 17, 2014)
3. Bruck, L. B.; Towns, M. H. *J. Chem. Educ.* **2013**, *90*, 685–692.
4. Reid, N.; Shah, I. *Chem. Educ. Res. Prac.* **2007**, *8*, 172–185.
5. Elliott, M. J.; Stewart, K. K.; Lagowski, J. J. *J. Chem. Educ.* **2008**, *85*, 145–149.
6. The Online Labs in chemistry website provides a partial list of virtual laboratory offerings, URL <http://onlinelabs.in/chemistry> (September 17, 2014).
7. Committee on Professional Training website, URL <http://www.acs.org/content/acs/en/policy/publicpolicies/invest/computersimulations.html> (September 22, 2014)
8. Gupta, T.; Burke K. A.; Mehta, A.; Greenbowe, T. J. *J. Chem. Educ.* Web Publication URL <http://pubs.acs.org/action/doSearch?text1=Greenbowe&field1=Contrib&type=within&publication=346464552> (September 21, 2014)
9. Sandi-Urena, S; Cooper, M; Stevens, R. *J. Chem. Educ* **2012**, *89*, 700–706.
10. Rickey, D.; Tien, L. T. *Trajectories of Chemistry Education Innovation and Reform*; ACS Symposium Series 1145; Holme, T., Cooper, M. M., Varma-Nelson, P., Eds.; American Chemical Society: Washington, DC; 2013; pp 31–45.
11. Bruck, L. B.; Towns, M. H.; Bretz, S. L. *J. Chem. Educ.* **2010**, *87*, 1416–1424.
12. Bretz, S. L.; Fay, M; Bruck, L. B.; Towns, M. H. *J. Chem. Educ.* **2013**, *90*, 281–288.
13. *Assessing Chemistry Laboratory Courses*; 233rd ACS National Meeting and Exposition, Boston, MA, August 21, 2007.
14. Website of practice exam. Note that an access code available through the site must be used to access the practice exam. URL <https://etest.cesd.umass.edu/partners/acs/demo> (September 24, 2014)

# Subject Index

## A

- Assessment of chemistry learning, role of
  - non-content goals
  - arguing importance of non-content assessment in chemistry, 149
  - practical implications, 151
  - theories of learning and role, 150
- description of quantitative survey
  - participants, 153*t*
- methodology for study
  - sample, 153
  - survey development, 152
  - survey items, 152
- quantitative survey results and
  - discussion, 153
  - general chemistry instructors' self-reported incorporation, 154*f*
  - instructors' evaluation of student performance, 157*f*
  - instructors' self-reported methods of assessment, 156*f*
  - methods of assessment, 155
  - non-content goals and skills, 155
  - statistics, 155

## C

- Classroom salon
  - analytics, 167
  - challenge, student engagement in large classes, 168
  - discussion mode, 166*f*
  - feedback, 170
  - future directions, 178
  - implementation into large-enrollment course
    - class instructor feedback, 177
    - common misconceptions, 174
    - findings, 174
    - incorrect or unsure use of scientific terminology, 175
    - instructional changes based on classroom salon findings, 176
  - methodology, final implementation, 171
  - questions from different reading assignments (RA), 173
  - student task in classroom salon, 172
  - trial and error phase, 171

- individual mode for student annotation, 165*f*
- large classes, 169
- misconceptions, existence and identification, 169
- online platform, 162
  - discussion mode, 163
  - individual mode, 163
- preparatory chemistry course, salon course homepage, 164*f*
- quick access to analytics, 167*f*
- Cultivating assessment understanding to support teaching and learning
  - standard 1, choosing tests
    - assessment types by item format, 31
    - assessment types by test interpretation, 30
    - assessment types by test purpose, 30
    - other ways of differentiating tests, 32
  - standard 2, developing tests
    - content definition, 33
    - defining passing scores, 35
    - distractor analyses, 38
    - item banking, 35
    - item development, 34
    - item difficulty, 36
    - item discrimination, 37
    - other analyses of interest, 38
    - overall plan development, 33
    - reliability indices, 39*t*
    - reporting results, 35
    - scoring of responses, 35
    - statistical evaluation of item and test quality, 36
    - technical report, 35
    - test administration, 34
    - test design and assembly, 34
    - test production, 34
    - test specifications, 34
  - standard 3, giving and understanding tests, 39
  - standard 4, using tests, 42
  - standard 5, grading learners, 42
  - standard 6, communicating results, 43
  - standard 7, promoting fairness
    - fairness in test development, 43
    - fairness in test use, 44
  - teacher competence in educational assessment of students, 28
  - standards, 29*t*

## D

- Data analysis
  - matching type of analysis to research question
    - qualitative research, 17
    - quantitative research, 16
  - relating results of data analysis to research question
    - qualitative research, 18
    - quantitative research, 16
- Designing methodology and tools (evaluation plan)
  - qualitative research and tools, 15
  - quantitative research, 13
    - quantitative research tools, 14
- Detecting differential item functioning based on gender subgroups, 47
  - DIF items by direction of favor and classification
    - logistic regression and Mantel-Haenszel methods only, 59*t*
    - Mantel-Haenszel methods only, 59*t*
  - gender differences, 49
  - methods, 56
  - results, 57
  - statistical methods
    - 2 × 2 contingency table, 55*t*
    - comparison of statistical method to determine DIF, 56
    - item plots, 54*f*
    - item response theory, 52
    - logistic regression, 55
    - the Mantel-Haenszel procedure, 51
  - uniform and nonuniform DIF items, DIF detection methods, 57*t*
  - uniform and nonuniform DIF items by direction of favor and classification, 60*t*
  - uniform DIF items by direction of favor and classification, 58*t*
- Developing first online general chemistry laboratory exam
  - average difficulty and discrimination scores for questions chosen, 188*t*
  - consensus of question categories, 183*t*
  - experimental design question, 188*f*
  - exploring possibilities, 184
  - final product, 189
  - lessons learned, 189
  - priority of techniques to be included, 186*t*
  - priority of topics to be included, 185*t*
  - purpose, 181
  - sampling of laboratory manuals, 184*f*
  - work of exam committee, 186

## F

- Feedback in testing, the missing link
  - example of posted student report, 107*f*
  - example of student response sheet including prompt for confidence, 108*f*
  - multiple-choice testing feedback, timing, 99
  - schematic of testing phases of study, 106*f*
  - STEM testing studies, 105
  - student response, 103
  - testing affect, 95
  - testing design, 94
  - testing feedback, 97
  - types of multiple-choice testing feedback, 98

## I

- Innovative uses of assessments for teaching and research, 1
  - information assessments, 2
  - method to use, 3
  - organization, 4*t*
  - purpose, 3
- Investigate student learning in organic chemistry classes
  - comparing number of hours students studied, 140*t*
  - comparing students post-examination prediction to their actual score, 139*t*
  - comparing students pre-examination prediction to their actual score, 137*t*
  - comparing students' self-assessment, 141*t*
  - grading scale used, 138*t*
  - methodology, 135
  - post-exam perception, 138
  - pre-exam perception, 136
  - time spent studying, 140

## M

- Matching evaluation plan to question data analysis
  - defining question and selecting appropriate research approach, 8
  - mixed-methods research, 11
  - qualitative research, 10
  - quantitative research, 9
  - theoretical frameworks, 12

designing methodology and tools  
(evaluation plan)  
presentation of results

## O

Organic chemistry practice exam, 67  
cognitively complex question on exam,  
84*f*  
construction, 68  
evaluation  
ACS anchoring concept content map  
(ACCM), 75  
cognitive complexity, 74  
difficulty, 73  
discrimination, 73  
number and percent of question types,  
77*t*  
number of questions assigned to each  
ACCM category, 76*t*  
student mental effort, 75  
test reliability, 74  
types of artifacts on practice exam, 76  
example of test item followed by  
respective mental effort rating, 70*f*  
groupings of 50 exam items, 69*t*  
impact on preparation for final ACS  
exam, 86  
linear correlation of student performance  
averages  
with respect to expert cognitive  
complexity, 79*f*  
with respect to student-rated mental  
effort, 83*f*  
positive linear correlation of  
student-rated mental effort, 83*f*  
results, 77  
algorithmic rules about alkene  
stability, 80*f*  
instructor information, 78  
multi-step reaction series, 80*f*  
number of questions of varying  
difficulty, 78*t*  
percent of questions of varying  
discrimination, 81*t*  
student metacognitive information, 82  
student completed self-assessment, 72*f*  
summarizes 2004 ACS OR04 percentile  
performance, 86*f*  
summarizes 2004 ACS percentile  
performance, 87*f*  
summary and future uses, 88  
summary of average on 2004 OR04 ACS  
(2) exam and hour exams, 85*t*

## P

Presentation of results  
qualitative research, 23  
quantitative research  
ANOVA summary table, 21*t*  
bar graph of mean exam scores by  
clicker use, 20*f*, 22*f*  
comparisons, 19  
descriptive statistics for exam scores,  
19*t*  
relationships, 23

## R

Resurrection points from final exam  
performance, 115  
analysis methodology and summary  
statistics, 119  
conclusions and implications, 129  
defining concept of resurrection points,  
116  
each general chemistry course,  
resurrection points earned summary,  
124*t*  
final letter grade, 121*t*  
minimum and maximum percentage  
points needed, 122*t*  
general chemistry courses, summary,  
120*t*  
instances of under performance and over  
performance, 126*f*, 127*f*  
learning tool, 117  
motivation in learning, 118  
number of students under or over  
performing, 125*t*  
odds of not getting final letter grade, 128*t*  
results and discussion, 123  
student perception, 118

## U

Use of resurrection points, 117

## W

Work of exam committee, 186