January 2019

Effects Of A Confidence-Based, Individualized Remediation Strategy On Student Learning And Final Grades In A Multi-Campus Human Anatomy Curriculum

Ethan Snow

EFFECTS OF A CONFIDENCE-BASED, INDIVIDUALIZED REMEDIATION STRATEGY ON STUDENT LEARNING AND FINAL GRADES IN A MULTI-CAMPUS HUMAN ANATOMY CURRICULUM

by

Ethan Lyle Snow
Bachelor of Science, South Dakota State University, 2014

A Dissertation

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Grand Forks, North Dakota

August 2019

This dissertation, submitted by Ethan Snow in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

Kenneth Ruit, PhD

Mandy Meyer, PhD

Patrick Carr, PhD

Richard Van Eck, PhD

Jamie Foster

James Foster, PhD

Anne Kelsch

Anne Kelsch, PhD

This dissertation is being submitted by the appointed advisory committee as having met all of the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

Chris Nelson, PhD
Associate Dean, School of Graduate Studies

7/8/19
Date

PERMISSION

Title: Effects of a Confidence-Based, Individualized Remediation Strategy on Student Learning and Final Grades in a Multi-Campus Human Anatomy Curriculum

Department: Biomedical Sciences

Degree: Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Ethan Snow
June 17, 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

ACKNOWLEDGEMENTS

I thank the members of my advisory committee for their support of me and this project throughout my doctoral education, especially Kenneth Ruit for chairing the committee, supporting me and this project from its conception, and for showing and teaching me how to be a professional and dedicated educator and scholar.

I thank Janet Jedlicka, Chair of the Occupational Therapy department, and Mandy Meyer, Course Director of OT 422, for their years of extensive support for me and this project in OT 422. Implementing this project in OT 422 made the entire project especially important regarding multi-campus learning environments.

I thank Karyssa Kimery, Jena Kobus, Michael Lehnerz, Michael Black, and Ashley Cambronne, my 2016/2017 Teaching Assistants, for their help with data collection, analysis, and creation of the individualized portions of the examinations. Without their help, this project would not have been feasible.

I thank Scott Pedersen, my anatomy professor at South Dakota State University, for his positive impact on me and my undergraduate education that led me to pursue such a rewarding and exciting dissertation project and become an anatomist and educator-scholar.

Lastly, I thank Kim and Dee Snow, my parents, for teaching me to earn victories but not for self-righteousness (Colossians 3:23-24) and accept failures without losing hope (Jerimiah 29:11) – qualities have undoubtedly contributed to my successes, especially for this dissertation project, and will continue to do so throughout my professional career.

I dedicate this work to my wife, Alyssa Ann Snow, and sons, Ezra Benton Snow and Noah Everett Snow. Thank you for your never-ending support and encouragement, for being the motivation for everything I do, and for making me a better person each day.

ABSTRACT

Reliable measurement of student learning and delivery of comparable education across distributed campus sites are two significant challenges facing institutions across the country. Evidence-based practices for learning objective (LO) development and use can help overcome comparability challenges, but widely-used correctness-only assessment methods contribute to these challenges since they are only able to interpret correct answers as displays of *complete* knowledge and incorrect answers as displays of *absent* knowledge. Assessment instruments that measure correctness alone are not able to distinguish *guesswork* (*i.e.,* when a student lacks knowledge but randomly chooses the correct answer), *partial* knowledge (*i.e.,* when a student has learned some correct information but does not display complete knowledge), or *flawed* knowledge (*i.e.* when a student learned incorrect information) – all of which are significantly different performances from *complete* or *absent* knowledge yet occur undetected when examining correctness alone. Confidence-based assessments (CBAs) use a multi-dimensional method of assessing knowledge that includes measuring student confidence levels in each of their answer choices in conjunction with answer correctness. As a result, CBAs can detect *complete*, *partial*, *absent*, and *flawed* knowledge levels and distinguish *guesswork* and from other correct responses.

This dissertation presents a novel use of CBA principles in an individualized remediation strategy implemented in high-stakes examinations for three cohorts of professional-level students in an OT 422 *(Anatomy for Occupational Therapists)* course taught simultaneously across two University of North Dakota campus sites. The variables in this study included individualized (*i.e.,* different for each student) vs. standardized (*i.e.,* same for all students) remediation interventions, self-assessment vs. instructor-derived feedback, and general motivation and learning strategies. These variables are hypothesized to influence learning via remediation and final grades between individual students and the two site populations. The following hypotheses were tested:

1. A confidence-based, individualized remediation strategy increases student learning.
2. Self-assessment of confidence-based academic performances increases student learning via remediation.
3. Student motivations, learning strategies, and academic performances are comparable across distributed campus site populations.

Student learning, measured by difference in confidence-based performance levels (PLs) through remediation, was shown to increase one knowledge level (1-2 PLs) following the individualized remediation intervention ($p < 0.001$) and resulted in achievement-level performances for 47 (65.3%) of the 72 LOs retested by each student ($p < 0.001$). As a result of the intervention's ability to detect flawed knowledge and guesswork, regular positive remediation of these performances to *better but incorrect* confidence-based PLs caused student grades to decrease by an average of 1.2% ($p < 0.001$) and resulted in a lower final letter grade for 17.4% of students ($p < 0.001$). No significant differences in learning were found to result from self-assessment vs. instructor-derived feedback. Despite differences in two motivations (Self-Efficacy for Learning and Performance, and Test Anxiety) and three learning strategies (Rehearsal, Metacognitive Self-Regulation, and Peer Learning) across distributed campus site populations ($p < 0.01$), comparable final percentage and letter grades suggest effectiveness of the evidence-based practices used to develop the course as well as implement individualized assessments across distributed campus site populations.

In summary, the confidence-based, individualized remediation strategy we employed increases student learning by using CBA principles to more reliably assesses student knowledge, and using evidence-based assessment practices to evaluate student learning helps ensure the delivery of comparable education among distributed campus sites. Outcomes from this study support educators' ongoing efforts to overcome challenges associated with reliable measurement of student learning and providing comparable yet individualized education to distributed populations.

CHAPTER I

INTRODUCTION

Assessment, Evaluation, and Grading

While many factors influence learning, the final determiner of whether learning occurs ultimately lies in *assessment*. Since the mid-1980s when the American Association for Higher Education began more closely examining the roles of assessment at their conferences and in institutions, the term *assessment* has taken on many different, even contradictory meanings in academia.[1] Its most accepted definition, as coined by Theodore Marchese in 1987[2] and later published by other educator scholars[1,3] and supported by the National Institute for Learning Outcomes Assessment[4] is, "the systematic collection, review, and use of information about educational programs undertaken for the purpose of improving student learning and development." Quizzes, high-stakes examinations, assignments, surveys, polls, projects, *etc.* (collectively called "assessments") are used to carry out assessment and range in type and purpose.

There are two main types of assessment. *Formative* assessments provide feedback about the progress of learning, and *summative* assessments produce an official record of learning outcomes and overall sufficiency. The forthcoming discussion about these assessments and others is largely based on the organization and work of Mhairi McAlpine in *Principles of Assessment* (2002)[5] and Craig Scanlan in *Assessment, Evaluation, Testing, and Grading*[6] and expanded on from other references.

Formative assessments are used to provide feedback to the learner during the instructional process about struggles, misconceptions, and learning gaps for the purpose of assisting their learning and thus improving future performances. In this regard, examples of formative assessments are not defined by the instrument or task[7] but instead by the *intent* of the instrument or task.[8] While typical examples include reflection journals, informal presentations, peer discussions and homework assignments, even unit tests (which are not normally good formative assessments because there is typically nothing students can do to correct their misunderstandings once administered) can be used as formative assessments if proper feedback is provided and strategically used to assist learning.[9] However, using one assessment for multiple purposes can lessen the effectiveness of the separate desired outcomes, so educator-scholars such as James Pellegrino, Naomi Chudowsky, and Robert Glaser caution against doing so.[10] In general, any process that allows instructors to judge students' current states of learning and make instructional changes and/or provide feedback for the purpose of improving learning can be considered a formative process.[11]

Formative assessments can be executed at any time throughout a course and in few or ample supply. Because they can provide students meaningful feedback on their performances (as described by steps 6 and 7 of Gagne's nine events of instruction[12]), formative assessments can help students improve their learning. In this manner, they are especially beneficial for tracking personal growth and development of learning over time.[13] In further support, behavioral psychologist Lev Vygotsky describes how providing optimal scaffolding and support to students keeps them in a zone of maximum learning potential (which Vygotsky terms the *Zone of Proximal Development*).[14] Aside from the learning benefits they offer to students, formative assessments can also inform instructors about their students' strengths and weaknesses in learning the material and contribute to rationale for making future changes in their course.

A type of formative assessment known as *diagnostic assessment* is specifically designed to assess students' current level of knowledge and detect any weaknesses or misconceptions they may have. Therefore, the timing for administering diagnostic assessments is critical; implementing them too early and students may not be able to demonstrate enough learning to receive meaningful feedback, and implementing them too late will make the feedback meaningless for preparing and remediating in time. Many diagnostic assessments are two-part, where the first part highlights areas that need improvement and the second part shows progress made on the weaknesses/misconceptions. Typical examples of diagnostic assessments include pre- and post-tests, self-assessment exercises, interviews, and polls. In addition to providing crucial feedback to students, diagnostic assessments can inform teachers about the readiness of their students, influence their decisions for choosing best-fit pedagogical and assessment methods, and measure effects from their decisions. Because diagnostic assessments are meant to provide feedback as a means for improvement, they typically are not graded (or if they are, only for completion at optimal completion times throughout the learning process).

Summative assessments evaluate performances at the end of a unit of instruction or at the end of the entire course (or both, sometimes being one in the same) so that the conclusion reflects the learner's overall performance (*e.g.*, final letter grades). This is commonly done by means of high-stakes examinations, portfolios, completed projects, and/or other like assessments. While formative assessment feedback and results are typically kept between the student and the teacher, summative assessment results are often shared externally to communicate students' academic readiness and abilities to next-level educational programs/institutions, employers, agencies, *etc*. McAlpine and Scanlan expand on *how* specific types of assessments are used.

Convergent assessment strategies have one pre-established, correct result each student is aiming for. These assessments are easy to use when evaluating factual knowledge as they can be used to create "can do" lists for students relative to curriculum requirements to come to a final appraisal of academic ability.[1] Convergent assessments are common practice in courses in most U.S. institutions with a focus on students achieving specific, pre-established learning objectives (*i.e.,* the "can do" statements). If designed appropriately, convergent assessments allow reliable and valid comparisons to be made between student performances, and the resulting data offers evidence of both student learning and teaching effectiveness. Although significant work is required to construct learning objectives, map assessment items to learning objectives, collect and manage data, and analyze data for outcome achievement, convergent assessments allow a grading process that is ordered and timely because of clear, pre-established correct answers. The fast and easy grading system (made even faster and easier if computerized grading is used) explains the widespread use of convergent assessments. Unfortunately, while many instructors choose this assessment strategy, they struggle to employ best-practice principles in assessment and, as a consequence, have faith in a *fallacy of false quantification* (*i.e.,* a tendency to test only what is easiest to measure – the idea of "when all you have is a hammer, you only hit nails") and/or the *law of the instrument* (*i.e.,* altering the actual problem to fit the assessment tool – the idea of [Maslow's hammer] "when all you have is a hammer, you treat everything as if it were a nail").[15]

Contrary to convergent assessments, *divergent* assessments describe assessment strategies that allow for a range of correct answers. While this strategy requires less advance planning, the grading process typically more time-consuming and involves more scrutiny for determining whether responses are deemed correct or incorrect based on the learner's perspective of the instruction and curriculum. In this regard, divergent assessments produce a more descriptive evaluation rather than a binary "right or wrong" judgement and, therefore, offer a more authentic assessment of learning.

Assessments can also be *formal* or *informal*. While *formal* assessments accentuate objectivity for awarding grades and making decisions, *informal* assessments emphasize high subjectivity for providing better feedback to each student during the process. Furthermore, assessments can be *continuous* (*i.e*., occurring intermittently throughout a learning experience) or *final* (*i.e.,* occurring at the end of a learning experience), and they can be conducted on the *process* of learning (*i.e*., focusing on the development and practice underlying a particular skill or ability) or the *product* of learning (*i.e.,* focusing on the outcome of the learning process). In designing and employing pedagogical approaches, most instructors use a variety of assessment methods to accomplish their goals.

Formative, summative, and other types of assessments can also be used in program evaluation and determining teaching efficacy. Data from well-designed assessments can be important for informing needs and quality improvement decision-making.[1] For example, poor assessment outcomes may suggest an instructor has paired the wrong assessment strategy to their current curriculum design and pedagogy (or *vice versa*), or they may suggest certain student needs are not being met by the course's curriculum or the instructors pedagogy.

Regardless, the information resulting from assessments can help instructors make necessary changes for the purpose of improving student learning.

An equally important and significant purpose of assessment is to *drive student learning*. While this idea is especially supported by educator scholars such as Gina Brissenden and Tim Slater,[16] John Heywood, author of *Assessment in Higher Education: Student Learning, Teaching, Programmes, and Institutions*, support this concept but with the following caveat:

> "…it is evident that some parents, politicians and teachers have accepted the axiom that since assessment drives learning the curriculum ought to be assessment-led. Unfortunately, excessive zeal can lead to excessive assessment and impede rather than enhance the learning-teaching process. Notwithstanding the good intentions of those involved in such politics such excesses display illiteracy in assessment and its functions and limitations to the possible and potential. In contrast, teachers' anxieties about assessment are too often focused on the limitations of assessment and its perceived potential to harm some children and students rather than on the beneficial effects it could have on learning. Effective assessment depends on assessors having a substantial knowledge of human development and learning…"

So, while assessments of all types can and should drive learning, excessive assessment and lack of assessment understanding and justification can actually hinder learning. If they are well-understood and not excessive, assessments can demand higher-order thinking skills and drive students to achieve more authentic learning, encouraging them to become self-directed lifelong learners.[17] Ultimately, as instructors choose and develop assessments, it is important their decisions be guided by evidence in the literature and what will be most beneficial to student learning.

The terms *assessment* and *evaluation* are often mistakenly interchanged despite their similar but different meanings. In their *Process Education Teaching Institute Handbook,* Daniel Apple and Karl Krumsieg suggest the clear distinctions between the terms. In agreement with the definition presented above, they describe how *assessment* is an ongoing, formative, process-oriented method of measuring individual student learning for the purpose of using diagnostic outcomes to encourage improvement, whereas *evaluation* is an intermittent, summative, product-oriented method of measuring individual student learning for the purpose of comparing judgement outcomes to known standards.[18] To put these terms into context, consider a student, John, who has just completed an examination. A formative assessment of John's learning may be "John needs to work more on achieving learning objective "*x*" in order to improve his understanding of the course content," whereas a summative evaluation of John's learning may be "John answered 85% of the questions correctly, earning him a B letter grade and placing him in the upper 25th percentile of his class." Other educator-scholars, such as Jean Rea and Anne Lundquist, agree with the distinctions between assessment and evaluation presented by Apple and Krumsieg.[19,20] Ultimately, the similarities and differences between assessment and evaluation boil down to "what, when, why, and how" characteristics – what is being measured, when it is being measured, why it is being measured (*i.e.,* the intent), and how the measurements are used.

Furthermore, assessment is more than just *grading* – another term similar but different to and also often mistakenly interchanged with *assessment*.[21–23] *Grading* (in terms of measuring knowledge in academic education) is the codified process of measuring a student's level of knowledge (*i.e.,* amount of learning achieved) by *evaluating* their performance of that knowledge.[24] Accordingly then, *grading* is the mechanism by which *evaluation* is carried out. These terms are extremely alike though – so much so that reasonable interchangeability between them tends to not alter the meaning for their use if the meaning is made clear by the user.[25]

Two common types of grading/testing, as originally coined by Robert Glaser in 1963,[26] are *criterion-referenced* testing and *norm-referenced* testing.[5,27] A criterion-referenced testing system compares each student's total-earned points/percentage against an absolute scale (*i.e.,* the "criteria," *e.g.* 90-100 = A, 80-90 = B, *etc.*) that may even be pass/fail. Regardless the scale, students simply earn a grade based solely on their performance in this system. Criterion-referenced testing systems are commonly used in institutional courses designed for teaching and student learning. On the other hand, a norm-referenced testing system assigns grades to students by comparing their performances to other's (*i.e.,* the "norm," *e.g.,* top 10% = A, next 10% = B, *etc.*). This testing/grading system is commonly used for nationally-based tests (*e.g.,* the SAT, GRE, IQ tests, *etc.*) that show each participant's placement, based on their performance, compared to everyone else's. This is referred to as *grading on a curve*[28] since the resulting grades distribution is guaranteed to represent a bell-shaped curve. Despite the phrase taking on other meanings over time,[29] it should not be confused with *curving grades* (*i.e.,* adding points to grades).[30] Norm-referenced testing may be easy to use, but it promotes unhealthy competition between students rather than cooperation if used in learning environments since each student's grade is determined by the success of others in addition to their own[27,31] and it can do a serious injustice to teachers' professional skills as well.[28] For these reasons, many educators such as

Carol Tomlinson and Jay McTighe strongly discourage the use of norm-referenced testing and advocate for the use of criterion-referenced testing instead.[31]

Regardless of the grading system, behaviors, attendance, participation, timeliness, neatness, improvement, and effort can be subject to grading as well. These conditions are not measures of learning objective achievement (*i.e.,* learning outcomes) though,[22,25] so including them in the grading process causes resulting grades to be further from accurately representing student learning – the original purpose for establishing a grading system.[32] Even without these effects, the value of grades is already limited. Grades tell instructors only that a student has learned something – not what they have learned.[33] Although grades can be directly tied to learning objectives, they are often only loosely correlated with them, making *grading* by itself only able to *supplement* assessment but *not suffice for* it.

Increasingly common factors have been shown to cause students to focus more on grades than learning.[34,35] Studies have verified this by showing how not presenting students with grades enhances their future performances.[32,36,37] A system known as *standards-based* grading (or sometimes also called *mastery learning* or *competency-based education*) has even been developed to overcome the limitation and undesirable effects of grades by only reporting achievement levels (*i.e.,* pass or fail) for learning objectives to represent how much was learned and also what learning was mastered.[5,38] Many assessment-scholars have supported a great separation of assessment and grading as a safeguard for ensuring objective measurement of student learning. However, according to W. Allen Richman and Laura Ariovich, this "firewall" has more recently been challenged for reasons of both efficiency and pedagogy.[39] Regarding efficiency, Mark Salisbury believes assessing and grading are redundant efforts and suggests combining them to save time and resources.[40] Karen

McClendon and Eileen Eckert support Salisbury's idea on the basis of pedagogy and believe basing grades on learning objective achievement (a system which they call *outcomes based grading*) will motivate students to direct their effort away from worrying about grades and [back] toward better learning achievement.[41] While few institutions have experimented with this "all-in-one" assessment/grading method, it in theory would return assessment back to its primary purpose – to improve student learning and development.[39]

Evidence-Based Assessments

Assessment data is only as good as its development. Every instructor should ask themselves the following questions:

1) Do I know what an "*x*%" on an assessment means in terms of learning outcomes?
2) Have I strategically aligned each assessment item with a learning objective?
3) Are my assessment items comparable for each learning objective?
4) Do I test learning objectives proportionately to the course content?
5) Are my learning objectives properly written and measurable?
6) Can I provide individual learning outcome data for each student?
7) Do I provide and differentiate learning objectives, course objectives, and course goals for my students?
8) Do my learning objectives support my course objectives? Do my course objectives support my course goals? Do my course goals support the mission of my department/institution?

If an instructor answers 'no' to any of these assessment-related questions, then the assessment data they are collecting in their classrooms may not likely be meaningful.

Meaningful, evidence-based assessments support institutional and program/department missions.[42] Understanding mission is critical since it guides the development of all learning experiences, especially academic courses. Just as each program/department mission should support the institutional mission, the purpose for each course or learning experience should support its respective program/department mission.

Despite being commonly used in academia, the terms *course goal*, *course objective*, *learning objective*, and *learning outcome* are often misinterpreted terms despite their distinctly different meanings. One explanation for the general misuse of these terms in academia is that educators publish such widely-varying, sometimes even contradictory, definitions for these terms making them fairly difficult to understand and clearly differentiate.[43] Having considered their many published definitions and analyzed their meanings, the definitions and usage of these terms we present in this document are what we believe to be correct, actual meanings and usages of the terms.

Course goals are very broad statements of student achievement that should be reached at the end of the course; they are statements of *why* the course exists (*i.e.* an overview of its purpose) and are used by both the teacher and student to determine, in general terms, what a course has to offer.[42–45] Although a few course goals are acceptable, a single all-encompassing goal for a course is often sufficient. Course goals should be realistic and achievable,[46] but they normally aren't directly measurable since they do not specify what students will learn or how they will learn it.[44,45] Instead, course goals are considered met or unmet depending on fulfillment of *course objectives*.[42,44]

After course goals have been established, the next step is to establish *course objectives* – statements that describe *how* the course goal will be accomplished (*e.g.* what the students will be expected to do, instructor's teaching responsibilities, types of assessments that will be used in the course, *etc*.).[42–44] Course objectives are more specific and typically more numerous than course goals yet still generally-stated with regard to curriculum content.[43] Although they pertain to the student, they are especially important for the teacher to guide delivery of instruction and administration of assessment. Once the instructor has established what the teacher and students must to do accomplish the course goal, *learning objectives* can then be established to help guide learning and assessment development.

Even though they are not the same, educators often use the term *course objective* in place of *learning objective. Learning objectives* are specific (and therefore often numerous) statements that describe *what* a learner should be able to do at established times and/or at the end of a learning experience; they are explicit expectations of student performance that faithfully reflect the course's curriculum content, and, as such, should guide student learning and assessment.[43,45,47] Learning objectives are arguably the most important component of meaningful, evidence-based assessments. In this regard, they typically receive the most scrutiny.

Properly-written learning objectives include the following components: the intended audience (*i.e.*, for whom the assessment is intended), a measurable behavior (i.e. what the learner is to do), any conditions the learner will encounter (*i.e.*, what the learner will use, have access to, or not be allowed to use), and the degree to which they are expected to perform (*i.e.,* measurement criteria of acceptable performance). This is commonly referred to as the ABCD method (Audience, Behavior, Condition, Degree) for writing learning objectives.[42,47–50] Robert Mager, who is credited with first outlining and explaining the necessary components of proper learning objectives, described these components in his 1962 book *Preparing Instructional Objectives*.[51] While each component is significant in its own regard, Mager and other educator-scholars agree that the *most* important component of a learning objective is the *behavior* component since that is what "…describes the kind of performance that will be accepted as evidence that the learner has mastered the objective."[52]

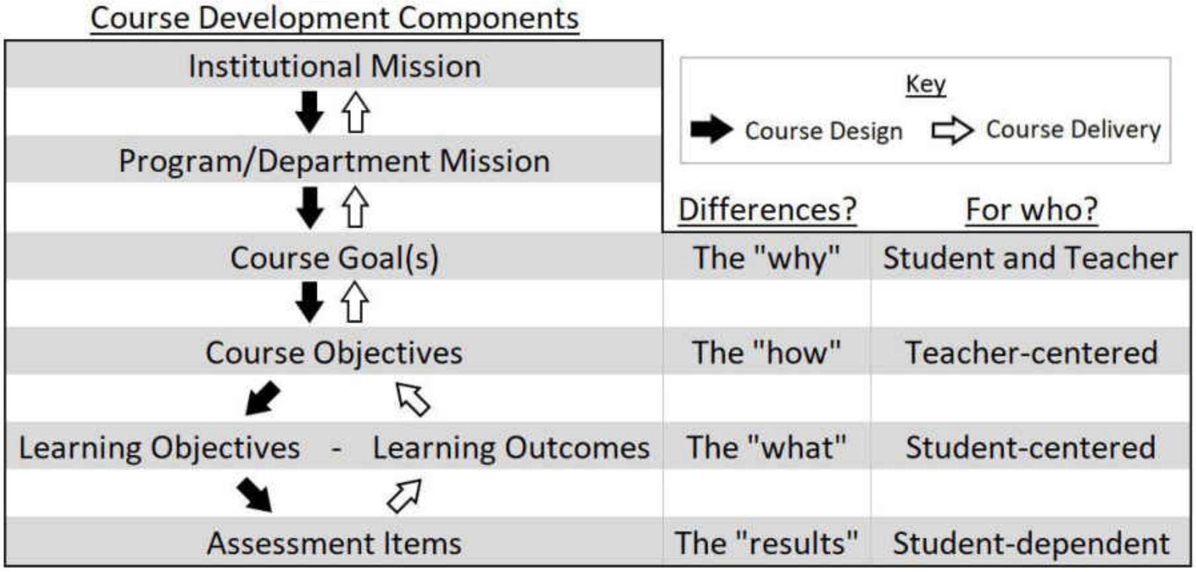
Learning objectives that are missing any of these components may not measure student learning as intended or even have anything to do with student learning at all. That said, the *audience* component is often omitted from written objectives with the assumption that the learner or student is to whom the learning objective pertains. Additionally, Mager admits that "it is not always necessary to include [*conditions*], and not always practical to include [*criteria*]," but he emphasizes the more components included in learning objectives the more clearly they will communicate their intended purpose.[51] If there is excessive repetitiveness in listing the audience, conditions, or criteria, an appropriate middle ground can be achieved by listing these components once at the beginning of the list of learning objectives.

In addition to including each of the above components, learning objectives must meet specific criteria in order to be properly constructed and inform assessment of learning. Learning objectives must be: 1) specific to a subject area; 2) observable and measurable (as guided by the six cognitive levels of Bloom's Taxonomy and their respective action verbs[53–55]); 3) attainable and not unrealistic; 4) relevant to course materials and available resources; and 5) time-bound such that the expectation for when they should be accomplished is clear.[50,56] George Doran first coined this criteria as SMART (<u>S</u>pecific, <u>M</u>easurable, <u>A</u>ttainable, <u>R</u>elevant, and <u>T</u>ime-bound) criteria in 1981[57], although Peter Drucker described this criteria in his book *The Practice of Management* as being necessary and useful for managing objectives in business in nearly three decades earlier.[58] The advantages of utilizing SMART criteria have made it quite popular. SMART criteria have been used by the Department of Education,[59] Center for Disease Control,[60] W.K. Kellogg Foundation,[61] United Way,[62] and other major organizations in different applications for writing objectives.[63] Toyin Tofade et. al. even used SMART criteria as a method of research in studying pharmacy students' ability to write SMART learning objectives.[64] Some sources swap certain SMART criteria words for synonyms (*e.g., achievable* for *attainable*), including the use of *realistic* instead of *relevant* to emphasize necessary resource availability over relevance (although both arguably go hand-in-hand). To make learning objectives more specific, formatting some to include "sub-learning objectives" (*e.g*., Learning Objective 7A vs. 7B, *etc*.) could direct students to specific content areas within one overall learning objective if they were not demonstrating sufficient knowledge in one area over another. This step would raise the total number of learning objectives, but the increased specificity is usually better able to guide student learning. Ultimately, if learning objectives are written with the ABCD method and to meet SMART criteria, they should be proper, capable of withstanding scrutiny, and able to guide curriculum development, instruction, and assessment practices to yield meaningful, evidence-based student learning and performance data.

After learning objectives are established, it is necessary to create an assessment strategy for each one that includes how and to what degree each learning objective is needed to be assessed in order to determine sufficient achievement. This necessary step is often forgotten, but if remembered it can give clear guidance to both instruction and assessment development, resulting in proportionate focus and necessary time spent on each leaning objective. Because learning objectives are student learning-centered, it is necessary to provide the learning objectives and the instructor's intended focus on each of them to the students to best guide their learning. After a student achieves a learning objective, it is then considered a *learning outcome* since a record of certain acquired knowledge or ability now exists for that student. Learning objective achievement (*i.e.*, *learning outcomes*) can be used to determine whether or not the course objectives were effective and course goals were met.

Achieving a strategic, top-down model of course design and delivery can show alignment of learning objectives to course objectives and course goals in support of program/department and institutional mission (Table I-1). Persistently using this model – and all the specifics therein – for course design/delivery will should put instructions in a position to answer "yes" to all of the difficult questions presented earlier. The effort to ensure meaningful, evidence-based assessment, far under weighs its positive impact on student learning, resulting course efficacy, and production of useful information for instructors, programs/departments, and institutions to further improve learning quality.

Table I-1. Top-Down Model of Course Design/Delivery for Evidence-Based Assessment. If a course is properly designed from broad- to narrow- spectrum components, each designed from the previous, then its delivery and assessment of learning can be "evidence-based" in terms of showing empirical data for achievement of each component.



Key

➡ Course Design ⇨ Course Delivery

Course Development Components	Differences?	For who?
Institutional Mission		
⬇ ⇧		
Program/Department Mission		
⬇ ⇧		
Course Goal(s)	The "why"	Student and Teacher
⬇ ⇧		
Course Objectives	The "how"	Teacher-centered
⬋ ⇖		
Learning Objectives - Learning Outcomes	The "what"	Student-centered
⬊ ⇗		
Assessment Items	The "results"	Student-dependent

Learning, Knowledge, and Guesswork

Learning is “a process of acquiring knowledge.”[65] Over time, many educators have elaborated on this basic definition by specifying how learning happens *over a certain period of time,* or *through study, instruction*, *experience*, *etc.* to *gain skills, values, abilities, etc.*,[66] but since any “process” entails carrying out some procedure (which takes time) and “acquisition of [even minimal] knowledge” is essential for demonstrating basic-level skills, values, abilities, *etc.*, using the basic, all-encompassing definition of *learning* is less limiting and better for its use in establishing a theoretical framework for measuring it. Because knowledge acquisition is the basic product of the learning process, the efficacy of learning can be quantified by measuring the amount and quality of resulting knowledge.

Knowledge, by basic definition, is commonly defined as “a belief that is true and justified”[67] It should not be confused with *intelligence*, which is “the ability to acquire and apply knowledge.”[68] The multi-dimensionality of knowledge creates four levels: *complete* (or *full*) knowledge, *partial* knowledge, *absent* knowledge, and *flawed* knowledge.[69] To measure knowledge, a person must compare how strongly they initially believe in information being true (*i.e.,* correct) to the *actual* trueness of that information (*i.e.,* compare what they *think* they know to what they *actually* know) in order to justify their belief and determine the amount and quality of knowledge that has been gained. Understanding the distinct differences in how beliefs and information coincide to determine knowledge is important, but these distinctions are not explicitly clear in the commonly used definition. For example, there is no such thing as trueness of a belief – only trueness of information based on facts. Alone, beliefs cannot be deemed correct or incorrect since they only exist (or don’t). Likewise, levels of belief intensity can be used to justify amounts of acquired information but not justify trueness of information. In simpler terms, the two independent components needed to measure knowledge are *belief justification* (used to determine amount of information) and *information trueness* (used to determine quality/correctness of information). For these reasons, a clearer and more accurate definition of *knowledge* would be "a belief that is *justified as true*" instead of "a belief that is *true and justified*." Based on these principles, educator-scholars agree that no single factor can reliably and accurately measure knowledge.

The concept of knowledge and the principles for empirically measuring it have been discussed for centuries, dating back as early as c. 300BC and c. 500BC with ideas from philosophers Aristotle and Confucius and popularized more recently by educator-scholars such as Darwin Hunt and James Bruno. While each of these individuals have their own philosophies on the concept of knowledge and how to measure it, all agree having knowledge requires more than solely demonstrating correctness of information.[67,70,71] Without belief justification, poor but otherwise accepted performances (*i.e.*, reward for *guesswork* despite absent knowledge), desirable but otherwise unrecognized performances (*i.e.*, no reward despite partial knowledge), and misinformation (*i.e.,* flawed knowledge) cannot be detected. For this reason, assessments that only measure correctness of student responses are simply insufficient in their ability to accurately and reliably measure knowledge.

Guesswork is the act of randomly constructing or choosing an answer to a question. While most guesswork justifiably results in no reward, educator-scholars have always been concerned about the opportunity for and frequency with which students guess correct answers by chance and how this affects the accuracy and reliability for *evaluating knowledge*. Regarding this concern, it is important to note that guesswork is not necessarily a misconception of knowledge or performance. Instead, guesswork is only a performance that may or may not be detectable depending on limitations of the chosen assessment method.

In the 1920s, multiple-choice examinations (MCQs) became widely used due to their easy grading and ability to feasibly assess large numbers of students, evaluate a wide range of objectives, and assess higher-order cognitive ability in accordance with Bloom's taxonomy with considerable reliability.[72,73] More recently, MCQs have become the main assessment tool worldwide, especially since the introduction of computer automation for scoring them.[72] However, MCQs face the greatest susceptibility for guesswork since they present *options* (*i.e.,* answer choices), including the correct answer, to each *stem* (*e.g,* a question or incomplete statement)[74] and typically demand a "forced response" from students even if they do not know the answer.[75] Questions that require students to construct responses do not result in correct guesswork nearly as often as MCQs since responses are constructed from students' own knowledge and are not chosen from a list of options. As MCQs began to dominate the assessment platform, educator-scholars' growing concern about detecting correct guesswork from MCQ's conventional correctness-only (also known as "number correct"[69,72] or "number right"[76,77]) "all-or-none" dichotomous scoring system – a rather crude method of assessment despite its remarkably common use in all levels of education[69,78] – began motivating them to begin investigating methods for addressing guesswork.[73]

Formula scoring (also called *correction for guessing*[79]) was developed as a popular mathematical method of correcting raw MCQ scores for correct guesswork. It is based on logic taking all assessment item performances into consideration to estimate the number of points gained from correct guesswork and then subtracting those points from the examinee's original score. *Formula scoring* uses the following equation:

$$x_c = R - \frac{w}{k-1}$$

where *R* is the number of correct answers, *w* is the number of incorrect answers, *k* is the number of alternatives per item, and x_c is the corrected score for guessing.[72,80,81] As best explained by James Diamond and William Evans, certain critical assumptions are considered for using this equation:

> "The derivations of this equation is based upon the assumptions that all wrong answers are guessed wrong and that all correct answers are obtained either by knowledge or guessing. The presence of misinformation and partial information is not considered. Theoretically, a student either knows the answer to an item and marks it correctly with probability 1.0 or he does not know the answer and guesses among k equally attractive alternatives."[82]

Accordingly then, all individual correct answers receive a weight of 1, and all incorrect answers receive a weight of -1/(*k*-1), giving formula scoring its alternative name, *right minus wrongs correction*.[80] This model also allows examinees to omit items without penalty if they are certain that their answer choice would be completely random.

Another model, known as the *random-guessing model*, was developed on the same principles and assumptions as *formula scoring* except it takes omitted items (items that examinees choose not to answer based on their belief that they would be making a completely wild guess) into account. The *random-guessing model* equation is as follows:

$$x_c = R + \frac{O}{k}$$

where *R* is the number of correct answers, *O* is the number of items omitted, *k* is the number of alternatives per item, and x_c is the corrected score for guessing.[80] According to Linda Crocker and James Algina, compared to the formula scoring model, "…this correction increases an examinee's observed score by awarding additional points for omitted items on the assumption that if the examinee had attempted the omitted item, the probability of selecting the correct response is $1/k$. Thus it is assumed that all guesses at omitted items would be made at random."[80]

Despite the differences in the *formula scoring* and *random-guessing* equations, Crocker and Algina show how despite their numerically different yields, the two equations for correcting for guesswork produce identical rank orders and perfect correlations if applied to the same set of item responses since the equation for formula scoring is a linear transformation of the random-guessing equation.[80] Despite their mathematical equivalency, Ross Traub and Ronald Hambleton believe the methods of implementing and utilizing these models for correcting for guesswork may impose different psychological factors on examinees' test-taking behaviors.[83]

Both guesswork-correction equations shows how guesswork, if left undetected and uncorrected for, significant inflates grades. While these scoring methods are based in unarguable logic, its accompanying assumption about not considering misinformation (*i.e.,* a strong belief of trueness in incorrect information) or partial information (*e.g.,* cuing, educated guesswork from eliminating thought-to-be incorrect answers, *etc.*) makes its reliability and validity vulnerable to criticism.[84] In response to these assumptions, Frederic Lord notes:

> "The asserted assumption is, of course, indefensible. Typically, examinees have some partial information about an item. For most multiple-choice items, they very likely can rule out one or more of the alternative responses with greater or lesser assurance. It is very difficult to be content with any kind of scoring based on an assumption of random selection."[77]

While Lord points out that formula scoring only assumes correct guesswork without being able to definitively detect it, another (perhaps greater) weakness is that it leaves students with little to no formative feedback for self-remediation. Notwithstanding these criticisms, educator-scholars such as Robert Frary describe scenarios in which formula scoring *should* be used – scenarios such as highly speedy tests or difficult tests with low score requirements where the benefits of formula scoring emphasized by the examinee's limitations of not having time or being able to eliminate even one incorrect choice on most items.[79]

To overcome limitations of formula scoring and random-guessing models, educator-scholars shifted their research efforts toward differentiating *wild* guesswork (*i.e.,* the random choosing of an answer from all possible choices; also called *pure* guesswork[78]) from *educated* guesswork (*i.e.,* the random choosing of an answer only after narrowing out one or more answers known to be incorrect and/or choosing an answer based on a cue, memory association, partial knowledge information, *etc.*; also called *informed guesswork*[74]). In 1953, Paul Dressel and John Schmid developed a *subset selection technique*[76] which they called the *free choice method*[85] (also called *partial knowledge award method*[72] or *liberal testing*[86]) that distinguishes wild guesswork from levels of educated guesswork and corrects item scores accordingly. Normally, students are only allowed to choose one MCQ answer option. In this method, Dressel and Schmid explain:

> "[Students are] informed that each item [has] one correct answer [option], but that they should mark as many choices as needed in order to be sure that they [have] not omitted the correct answer. Furthermore, they [are] informed that it would be to their advantage to mark as few answers as possible, inasmuch as the scoring formula [involves] a correction factor of [1/T_i] the number of incorrectly marked answers."

Their scoring formula is as follows:

$$s_c = (T_i * R) - W$$

where T_i is the total number of incorrect answers listed for the item, R is the number of marked answers that are correct (*i.e.,* 1 if the correct answer is contained within the selected responses or 0 if it is not), W is the number marked answers that are incorrect, and s_c is the resulting score for the item corrected for all levels of guesswork (wild through all levels of partial knowledge).[85] As a result, choosing a subset of answer choices with the correct answer contained within it awards partial credit for partial knowledge, but choosing any single answer or subset of answers that does not include the correct answer (*i.e.,* the correct answer was believed to be incorrect and thus omitted) will result in a penalty (*i.e.,* negative scoring) for demonstrating misinformed or flawed knowledge. The only way to earn full credit is to demonstrate full knowledge and choose the one correct answer.

The subset selection technique was much more accepted than formula scoring since it was able to more accurately assess wild guesswork and partial knowledge, but the negative penalty marking sparked criticism. Michael Akeroyd later adopted a similar method to this one, called the *dual response system*, but it did not use negative marking and its subset selections were limiting.[87] Lucia Otoyo and Martin Bush later used Akeroyd's framework with a novel marking scheme, calling it *subset selection without mark deductions*, to achieve full subset selection without negative marking.[78]

Despite the terms being previously accepted as one and the same and used interchangeably, Clyde Coombs *et. al.* began to emphasize need to differentiate *educated guesswork* from demonstrating *partial knowledge*, seeing that appropriate amount of credit should be awarded for *partial knowledge* but not, or at least not as often, for *educated* *guesswork*. They based the design of their study on the probability that students would more likely guess a correct answer from among remaining answer options if they were to instructed to choose the correct answer or include the correct answer in a subset selection. The rationale for this approach was that grade inflation from credit awarded for unanswered questions was statistically indifferent to the inflation of grades that would have been caused by the probability of guessing answers correctly. In light of this framework, Coombs *et. al.* established an inverse of the subset selection technique to more accurately differentiate partial knowledge from educated guesswork and other performances. Contrary to the *subset selection technique* which supported *educated guesswork* through a process of *explicit inclusion*, the *elimination procedure* (also called *elimination testing*[69] or *distractor selection method*[74]), as they called it, supports *partial knowledge* through a process of *explicit exclusion* by awarding credit to students for their ability to select any/all answer option(s) *except* the correct answer;[88] the only way students can earn full credit on an examination item is to select all of the *distractors* (*i.e.,* incorrect answer options), leaving only the correct answer unselected.[89] This way, if students omitted some but not all distractors and left the correct answer among the remaining answer options (simply inverse to the subset selection technique), partial credit could be awarded for correct partial knowledge. However, if some but not all distractors are omitted and the correct answer omitted as well, leaving a subset of distractors thought by the student to include the correct answer option shows a level of incorrect partial knowledge, or as Coombs *et. al.* identify it, *partial misinformation*. Ultimately, results from this procedure can be interpreted similarly as those of the *subset selection technique* to detect where and correct for the extent for which guessing occurs with perhaps a greater scrutiny for distinguishing partial misinformation (flawed knowledge) from partial knowledge.[69]

Because students are choosing incorrect answer options to receive credit in the elimination procedure, and there are more incorrect answer options than the sole correct one, eliminating the correct answer results in a hefty *n*-1 (*n* being the total number of answer options per question) negative scoring penalty – one that Akeroyd even describes as "savage."[87] Martin Bush points out this same criticism and suggest that his own -1 penalty for incorrect answer selection is "surely more psychologically acceptable," and he also criticizes the *elimination procedure* for encouraging examinees to think negatively instead of positively.[86] Despite these common criticisms, however, the ability of the elimination procedure to accurately and reliably detect *complete knowledge* from eliminating all incorrect answers, *partial knowledge* from eliminating a subset of (*i.e.,* one or more) distractors, *absent knowledge* from omitting the question or eliminating all options, and *flawed knowledge* from eliminating the correct answer option has continued to make it a popular assessment method among educator scholars.[69] Its inverse methodology to subset selection may also make the elimination method more effective at reducing the frequency at which wild guesswork occurs.[90]

All of these methods, from *formula scoring* to *elimination procedure*, made great improvements to accurately and reliably measuring knowledge, overcoming the limitations of "number correct/right" assessment methods to detect guesswork, and establishing the logical framework for understanding both. *Knowledge* can be *full*, *partial*, *absent,* or *flawed*,[69,89] and detecting *wild guesswork* vs. *educated guesswork*[91] within those constructs is not simple or easy. In fact, all of these methods mentioned still lack the ability to empirically detect every isolated occurrence; albeit extremely rare, students *could* act independently of the assumed learning behaviors without detection since only correctness of the chosen answer options is recorded. However, the underlying principles of these methods are key as they support the concept that self-perceived *belief justifications* of student performance can offer a fuller understanding of student *knowledge* when compared to *information trueness*. Still, educator-scholars and institutions recognize the limitations and ineffectiveness of correctness-only knowledge assessments yet still commonly use them because better solutions have been difficult to identify and feasibly implement.

Misinformation and Confidence-Based Assessments

While educator-scholars are concerned about wild guesswork (from absent knowledge) and educated guesswork (from partial knowledge), they are also concerned about performances indicating misinformation (*i.e.,* flawed knowledge). Misinformation is information that has been learned incorrectly but is believed to be correct. While educator scholars' concerns about guesswork primarily regard accuracy and reliability for *evaluation*, their concerns about misinformation also regard future implications and consequences from students acting on flawed knowledge. Therefore, one of the greatest motivations for detecting misinformation is to provide meaningful, formative feedback to students so that they are able to correct misinformation before acting on it.

Students can demonstrate flawed knowledge by exhibiting *full misinformation* or *partial misinformation*.[69,89] *Full misinformation* is much easier to detect as it is a result of recording a single incorrect answer option with complete belief that it is the sole correct answer. *Partial misinformation* involves displaying similar flawed knowledge but with less confidence, making it more challenging to detect. Partial misinformation stems from a flaw in partial information and requires knowing about the process by which students narrow out and/or arrive at answers (*e.g.,* the involvement of cuing, making educated guesses from eliminating thought-to-be incorrect answers, *etc.*). This makes the process of distinguishing *partial misinformation* from *partial knowledge* rather specific.

Since partial knowledge is commonly characterized by the ability to eliminate one or more distractors, knowing which answer options students identify as definitively incorrect is essential for distinguishing *partial misinformation* from *partial knowledge*. If educated guesswork results in a correct guess from a narrowed-down subset of answer options, then partial information is empirical and signifies partial knowledge. However, educated guesswork that results in an incorrect guess from a subset of narrowed-down answer options can mean one of two outcomes:

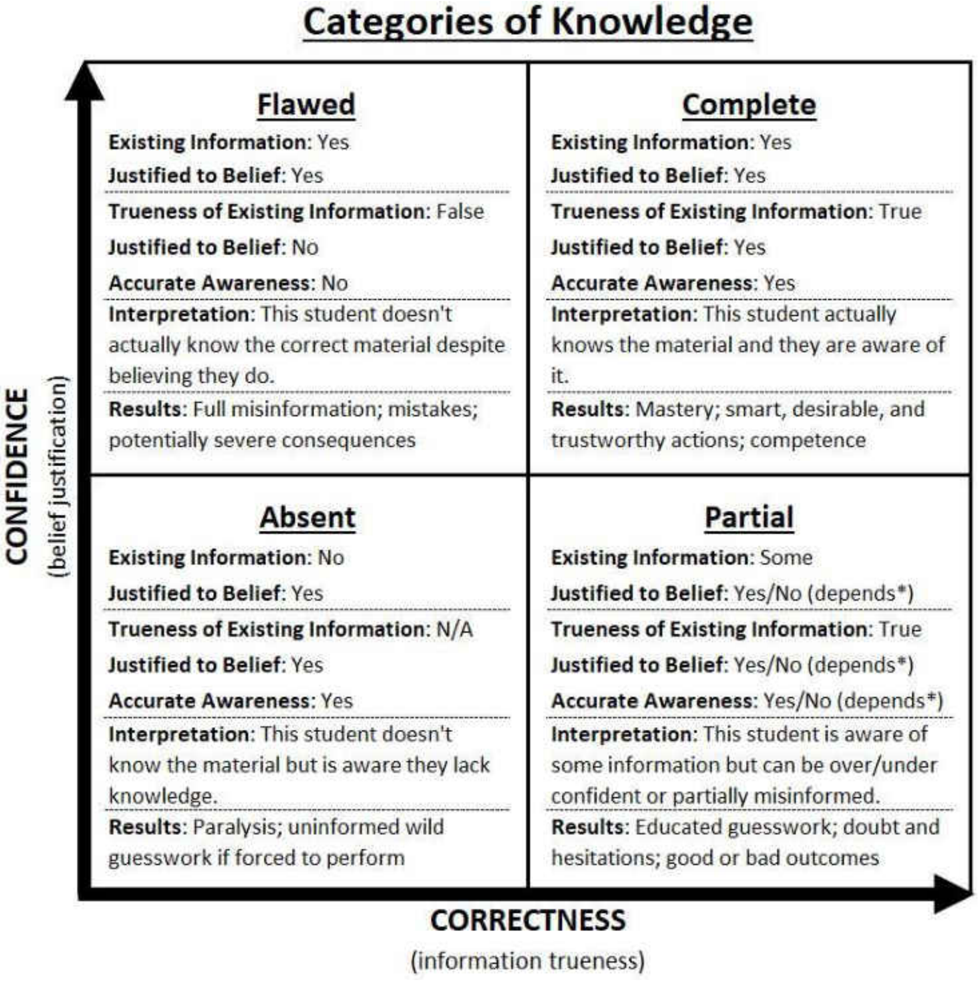
1) the incorrect guess was made from a narrowed-down subset of answer options that contained the correct answer, thus *partial information* is empirical and signifies *partial knowledge*, or
2) the incorrect guess was made from the narrowed-down subset of answer options that *did not contain the correct answer* (meaning the student had eliminated the correct answer option in believing it was definitively incorrect), thus *partial misinformation* is empirical and signifies *flawed knowledge*.

Given these strictures, if it is not known whether or not the correct answer was eliminated by a student with medium-level belief in their incorrect answer, then partial misinformation is undetectable and will always be interpreted as correct partial information instead and categorized as partial knowledge instead of flawed knowledge.

The previously-discussed methods for correcting for guesswork showed how using implied, indirectly-measured behaviors, such as implied belief justifications from eliminating answers or selecting a certain subset of answers, supports the essentiality of belief justifications for accurately and reliably understanding knowledge and examination performances. Most of the methods were able to detect, or at least account for, *full misinformation*, while methods like the *elimination procedure*[89] were better able to precisely differentiate *partial misinformation* from *partial knowledge*. Empirically determining when and how misinformation affects student performance, and an efficiently feasible way of doing so at that, still remains a challenge.

Assessments that strictly measure response correctness (*i.e.,* information trueness) alone are limited by their inability to *directly* detect students' self-perceptions (*i.e.,* belief justification) about how certain they are about the correctness of the answer option they choose. This "state of feeling certain about the truth of something" is known as *confidence.*[92] Because confidence is only dependent on *perceived* response correctness and not *actual* response correctness, it can be compared to response correctness and serve as the belief justification required to accurately assess knowledge. Thus, confidence is critical for *learning* and accurately and reliably measuring *knowledge* by helping detect all levels of knowledge, including guesswork and misinformation performances.

Four categories of knowledge can be distinguished by comparing confidence to correctness (Figure I-1): *complete knowledge*, *partial knowledge*, *absent knowledge* (*i.e., absent knowledge*), and *flawed knowledge*.[69,89] In accordance with how Coombs *et. al.* describe the levels of knowledge, Timothy Adams and Gary Ewen display how complete knowledge (referred to as "mastery") is characterized by correct information accompanied by high confidence and leads to smart actions. Partial knowledge is characterized by some correct information accompanied by medium confidence but often leads to doubt and hesitation when called to act. Absent knowledge is characterized by low levels of correctness and confidence characteristic of new learners who are uninformed, resulting in action paralysis. Finally, flawed knowledge is characterized by low levels of correctness accompanied by high confidence, indicating something has been learned incorrectly and could lead to mistakes being made.



Categories of Knowledge

Flawed	Complete
Existing Information: Yes **Justified to Belief**: Yes **Trueness of Existing Information**: False **Justified to Belief**: No **Accurate Awareness**: No **Interpretation**: This student doesn't actually know the correct material despite believing they do. **Results**: Full misinformation; mistakes; potentially severe consequences	**Existing Information**: Yes **Justified to Belief**: Yes **Trueness of Existing Information**: True **Justified to Belief**: Yes **Accurate Awareness**: Yes **Interpretation**: This student actually knows the material and they are aware of it. **Results**: Mastery; smart, desirable, and trustworthy actions; competence
Absent	**Partial**
Existing Information: No **Justified to Belief**: Yes **Trueness of Existing Information**: N/A **Justified to Belief**: Yes **Accurate Awareness**: Yes **Interpretation**: This student doesn't know the material but is aware they lack knowledge. **Results**: Paralysis; uninformed wild guesswork if forced to perform	**Existing Information**: Some **Justified to Belief**: Yes/No (depends*) **Trueness of Existing Information**: True **Justified to Belief**: Yes/No (depends*) **Accurate Awareness**: Yes/No (depends*) **Interpretation**: This student is aware of some information but can be over/under confident or partially misinformed. **Results**: Educated guesswork; doubt and hesitations; good or bad outcomes

Figure I-1. Categories of Knowledge Based on Correctness and Confidence. Evaluating confidence respective to correctness differentiates complete, partial, absent, and flawed levels of knowledge. Trueness of existing information is not applicable for absent knowledge because no information has been learned by students displaying this performance. *If over/under confident, then ‘No.’

Jon Warwick et. al. describe a similar relationship of confidence to knowledge, comparing “inclination to eliminate perceived distractions” to “proportion of answers eliminated that are actually distractors” as he describes the model of the liberal subset selection technique.[90] Consider the four levels of knowledge in the following circumstances:

1) A student chooses the correct answer, and he is highly confident has chosen the correct answer. This student displays complete (true and justified) knowledge, earns full credit, and would be trusted to make a smart action.
2) A student chooses an answer at random with medium confidence that the answer he chose is correct, indicating he arrived as his answer by partial knowledge and educated guesswork. If this student’s educated guess is correct, only partial credit (not full credit despite the correct answer choice) is deserved. If the student’s educated guess is incorrect and the correct answer is among the remaining options he guessed from, he also deserves partial credit. However, if the student’s educated guess is incorrect and the correct answer is not among the remaining options he was considering (meaning he identified the correction option as being certainly incorrect), then no credit is deserved as this represents a case of *partial misinformation*. Regardless the outcome, this situation would likely result in doubt and hesitation by the student to act.
3) A student chooses an answer at random with no confidence due to absent knowledge. If the randomly chosen answer is correct, correct wild guesswork is detected and can be corrected by rescinding credit. If the randomly chosen answer is incorrect, the student rightfully earns no credit for absent knowledge and

incorrect wild guesswork. Regardless the outcome, this student would not act due to being uninformed.

4) An incorrect answer is chosen with high confidence, indicating a level of full misinformation/flawed knowledge. No credit is rightfully awarded, but this student requires formative feedback to correct the misinformation since he would otherwise act on full misinformation and face consequences for his mistakes.

While confidence and correctness in the former circumstances are more aligned and in the latter are less aligned, these performances demonstrate how crucial a belief justification, such as confidence, is for distinguishing knowledge from simply "correctness".

Like correctness alone, confidence alone cannot accurately measure knowledge. High and low confidence levels, just as incorrect and correct answer choices, can each result in both desirable and undesirable performances; just as correctness alone presents an incomplete interpretation of knowledge in the case of guesswork, confidence alone presents an incomplete interpretation of knowledge in the case of misinformation. This illustrates the critical nature of accurately measuring both confidence and correctness and comparing them to one another to accurately and reliably measure knowledge. However, the objectivity of correctness makes it the better prime determinant for establishing into which category of knowledge is reflected by a performance on an examination item. This is why only correctness is accepted for using in widely popular "number right/correct" methods, as it is the better choice of the two to use alone despite its many acknowledged limitations.

Assessment methods that incorporate confidence would credit students appropriately for showing *complete*, *partial*, *absent*, and *flawed* knowledge levels and differentiate performances such as guesswork and misinformation within those categories that would have otherwise gone unrecognized and/or been categorized incorrectly. Comparing correctness and confidence follows order; only after correctness has been determined can any level of confidence then separate performances into each knowledge category. For example, if a student answers a question incorrectly, the level of confidence determines whether his/her knowledge is *absent* (from low confidence), *partial* (from medium confidence), or *flawed* (from high confidence). Similarly, if a student answers a question correctly, the level of confidence determines whether his/her knowledge is *absent* (from low confidence), *partial* (from medium confidence), or *complete* (from high confidence). This same methodology is capable of detecting of guesswork and misinformation.

While alone they have their limitations, together, correctness and confidence can provide an accurate interpretation of knowledge and performance. Together, they not only can distinguish between what students *think* they know and what they *actually* know, but they also separate student performances that would have otherwise gone unrecognized and/or been categorized incorrectly. For example, students who choose the correct answer by wild or educated guesswork are no longer considered to possess equivalent knowledge to students who choose the correct answer and are sure of their answer. These principles, and the idea to associate correctness with confidence to form a more accurate assessment method for measuring knowledge through examination performance, sparked the development and introduction of Confidence-Based Assessments, or CBAs.

The first use of a confidence-based assessment is credited to Kate Hevner, who began to pursue methods for correcting guesswork after the rise in popularity of multiple choice examinations in the early 20th century. Understanding that guesswork could be detected from a correct answer chosen with a low level of confidence, she designed a study in which students recorded their level of confidence (low, medium, or high) in the answer they chose as correct for each true-false examination question. She then compared their confidence levels to the levels of correctness in their answers and was able to detect guesswork and correct for it accordingly. Hevner published her findings in 1932 and showed the first empirical evidence for how this method can be used to detect guesswork and correct for those performances accordingly.[93]

Hevner's study was especially significant because true-false examination items have only have two answer choices, leaving a student with the highest probability of guessing the correct answer by pure chance if they do not know the correct answer or are unable to eliminate the only distractor. While this is true and of great concern, true-false type questions are easy to construct, quick to grade (especially with computer-aided grading systems), and often more closely related to real life situations than multiple choice questions with more than two answer options.[94,95] If more than two answer options are made available, such as in traditional MCQs, the probability of successful guesswork decreases due to the higher probability of choosing one of the many incorrect options over the one correct answer.

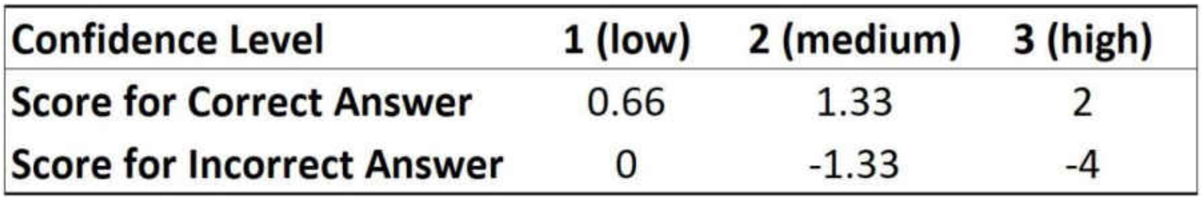
After Hevner introduced the idea of CBAs to reliably detect guesswork in true-false examinations, other educator-scholars began to further develop and study other aspects of CBAs, particularly those for assessing partial knowledge. Ghadermarzi *et. al.* began to use CBAs to estimate students' partial knowledge on MCQs. He hypothesized that when a student recorded a "medium" level of confidence in their answer choice, it indicated learned-yet-incomplete information was used to eliminate some incorrect options.[73] He concluded that this method for assessing partial knowledge fairly assesses knowledge and provides an authentic and effective method for examination.

While the use of confidence to detect partial knowledge began with Hevner in 1932, the idea of measuring the self-assessment of belief in information to better assess knowledge existed long before then. To examine just how long this principle has been considered, consider the following statement made by Aristotle circa 300BC: “He who thinks himself worthy of great things, being unworthy of them, is vain.” Even well over 2000 years ago, Aristotle and other philosophers knew that knowledge only existed if a person exhibited both trueness of information learned and justification through belief. Martin Bush’s study of the subset selection technique (described previously) allowed students to use their own intellect to choose multiple options as answers if they were not sure of the sole correct answer. Although he and others alike may not have declared to be examining confidence per se, there is no doubt he and other educator-scholars were exploring how self-assessment of belief sureness can be used to justify correctness and better assess knowledge. From these and other studies, it is important to remember that confidence-based assessments are ultimately “belief-based” assessments designed to assess knowledge through information correctness and belief justification and are often utilized by individuals during the learning process whether or not the behavior is being recorded and used for assessing knowledge in the academic setting.

While Hevner is credited with the introduction of using CBA principles, the term “confidence-based assessment” or “CBA” was not fully established until over 50 years later. In 1990, James Bruno established a method known as Information Resource Testing (IRT), which was later renamed to Confidence-Based Assessment (CBA)[73]. Years later, A. R. Gardner-Medwin become known for the development of CBAs as he standardized the method and principles for their use.[96,97] Although he later changed their name to *certainty-based assessments*,[98] the term *confidence-based assessment* is still more commonly used.

In the standard CBA method established by Gardner-Medwin at University College London (UCL), often referred to as the UCL or Gardner-Medwin method, students are instructed to choose/construct an answer to a question and then select/record their level of confidence in having answered the question correctly. While CBA studies have implemented confidence scales using 100,[99,100] five,[101] four,[85] and three[93,94] levels of confidence from which students can choose, Gardner-Medwin and other authors have demonstrated that using three levels of confidence is reasonable and produces valid and reliable results. The three confidence levels (Table I-2) are typically represented numerically with “1” indicating low confidence, “2” indicating medium confidence, and “3” indicating high confidence. In the Gardner-Medwin method, credit is awarded (or deducted) depending on the combination and comparison of confidence to correctness for each answered examination question.[102]

Table I-2. UCL (Gardner-Medwin) CBA Scoring Scheme. Scores listed are based on a question worth 2 points so that whole numbers can be used. Incorrect answers will receive -1.33 or -4 point penalties if accompanied by medium or high confidence levels, respectively.



Confidence Level	1 (low)	2 (medium)	3 (high)
Score for Correct Answer	0.66	1.33	2
Score for Incorrect Answer	0	-1.33	-4

Gardner-Medwin’s CBA principles support the assessment of knowledge through confidence and correctness, but his scoring scheme is based strictly on alignment of confidence and correctness which, as previously establish, does not always coincide with the knowledge level interpreted from each confidence-based performance. According to his scoring criteria, correct answers reported with high confidence (indicative of complete knowledge) merit full credit. Correct answers chosen with low confidence (indicative of correct guesswork and absent knowledge) demonstrate merit some credit – more than from choosing an incorrect answer instead. Incorrect answers reported with high confidence (indicative of *full misinformation* flawed knowledge) and merit the worst score, even a penalty. Lastly, two additional categories are considered for medium-level confidence in either correct or incorrect answers (indicative of partial knowledge) awarding correct answers but penalizing incorrect answers. Although Gardner-Medwin has shown how the CBA model of confidence and correctness alignment proves valid and reliable, inconsistencies between the order of desirable knowledge levels and the order of correctness/confidence alignments from his scoring criteria are concerning.

In order for a CBA method and related scoring scheme to be reliable, its variables must be consistent between performance combinations. The correctness variable is consistent since it is determined solely by either a computer-aided grading system or an instructor. Confidence, is recorded separately by each student, so establishing a standardized rationale for choosing each level of confidence and having students understand what each level of confidence means is crucial. With standard CBAs, a clear set of instructions is provided to students to ensure confidence responses are reliable and accurate and can be compared and used for fairly assessing knowledge. Typical qualitative instructions for choosing confidence levels after first choosing an answer are as follows:

- *If you do not know the answer and are unable to eliminate any incorrect answers, record a low confidence level ("1").*
- *If you are able to eliminate some incorrect answers but are unable to choose one final answer as the answer that you are certain is correct, record a medium confidence level ("2").*
- *If you believe with complete certainty that the answer you chose is the correct answer, record high confidence ("3").*

There is clearly a wider range of possible scenarios that could pertain to "medium confidence" if there are four or more answer options. Through his CBA studies Gardner-Medwin developed a quantitative, statistical rationale for choosing confidence levels and suggests using it to give more definitive guidance. Based on mathematical theory of how many answer options exist (five in this case) compared to how many answer options the student is able to eliminate as being incorrect, he instructed students to choose the their level of confidence based on the following:[103]

- *Choose low confidence (1) if you are less than 67% sure of your answer* (*i.e.,* if you must guess between three or more answers).
- *Choose moderate confidence if you are between 67-80% sure of your answer* (*i.e.,* if you must guess between two answers).
- *Choose high confidence if you are greater than 80% sure of your answer* (*i.e.,* if you are sure of one answer and not considering any others).

Each qualitative and quantitative approach promotes standardization of the reporting of confidence levels by helping students determine the appropriate confidence level to report that most accurately reflects their behavior at the time of examination. It also helps students understand how confidence is being taken into account in assessing what they have learned. However, this time "low confidence" could include more scenarios than the others and "high confidence" scenarios are more specific, reflecting Gardner-Medwin's significant scoring criteria associated with being highly confident.

A common concern about CBAs, expressed particularly by students, is one of objectivity. Students who express this concern typically consider themselves "not confident" people in general and worry their scores will suffer accordingly. This emphasizes the importance of teaching students about the type of confidence being measured in CBAs – how it is not one of general personality (*i.e.,* self-confidence in their abilities) but one confined to the direct perception of the factual information presented in the questions and chosen by pre-established guidelines (*i.e.,* confidence in given information and respective performance).

Another concern about CBAs regards the timing of confidence level collection. As most CBA methods prompt students to record a level of confidence secondary to choosing a final answer, one could conclude that the act of having chosen a final answer could raise confidence if the levels are not dictated mathematically like Gardner-Medwin outlines. Although no studies have addressed this concern directly, those that have collected confidence levels simultaneously with final answer choices demonstrate no significant differences in outcomes.[75] Collecting both answer and confidence level at once does simplify steps for answering questions for students though.

Up to this point, CBAs have only been discussed in the context of MCQs. This is due to how widely multiple choice type examination formats are utilized because of their ease of grading and support for feasibly collecting the large amounts of data that result from the administration of CBAs. Despite this, CBA methods are very versatile and can be incorporated into any assessment in which knowledge is being measured, including oral, constructed-response, and practical examinations.

Ultimately, one of the main ideas behind the use of CBAs is to examine and encourage the use of student metacognition. Metacognition is "higher-order thinking that enables understanding, analysis, and control of one's cognitive processes, especially when engaged in learning."[104] Examining this process of "thinking about thinking" or "knowing about knowing" can help instructors identify students who exhibit recurring behaviors of awareness or unawareness of their own knowledge level and quality. In cases of student unawareness of their own knowledge level and quality, especially in the case of exhibiting flawed knowledge (*i.e.,* the mistaken belief that displayed knowledge is more correct than it actually is), instructors can provide feedback to those individuals and help them establish more correct and efficient metacognition. The ability for instructors to do this is especially important for correcting student mistaken metacognition because, without being made aware of their deficit or error, those students are otherwise incapable of evaluating and understanding their incompetence. Social psychologists David Dunning and Justin Kruger studied how this absence of (or error in) metacognition leaves students unable to identify the misconception in their original self-awareness, often leaving them believing their cognitive ability is greater than it is. Dunning and Kruger called this phenomenon the Dunning-Kruger effect,[105] which encompasses educator-scholars' concerns about students who consistently demonstrate flawed knowledge (*i.e.,* high confidence in incorrect responses) as detected by CBAs.

Examining student metacognition is also meaningful beyond cases of misconception. When it is not misconceived, it can be especially useful for determining the progress of student learning. For example, a beginning learner may expectedly exhibit lower confidence in metacognition than an advanced learner. The beginning learner will need to continue working on studying more efficiently and developing better self-awareness of their own cognition, whereas an advanced learner displaying the same behavior may have failed to learn properly and require more rigorous interventions. In other words, performances (the "what") and metacognition (the "why") should be expected to change with learner development and experience (the "when"). In this manner, instructor and student use of metacognition should help drive student learning through formative feedback as well as provide a more accurate summative assessment of knowledge and learning.

Aside from straightforward assessment benefits, the metacognitive aspects of CBAs offer other significant benefits to learning. Hunt designed a study to test the hypothesis that retention of newly learned material is dependent on how confident students are in their answers. He found that students who had little to no confidence in learned information could only remember 25% of that information after only one week. By comparison, he found that students who were highly confident could remember 91% of the material after one week and 79% of material after an entire year[67,75]. Gardner-Medwin also recognizes how CBAs can stimulate a deeper, more reflective learning.[103]

In summary, the benefits of CBAs seem to far outweigh their concerns. CBAs offer considerable benefits over traditional correctness-only assessments by providing accurate and meaningful information about performance used to meaningfully assess all knowledge levels – complete, partial, absent, and flawed – and detect guesswork and misinformation. In addition to offering an alternative to overcoming the limitations of correctness-only assessments, CBAs stimulate student metacognition development, improve the learning process, and aid in long-term knowledge retention.

Distributed Education

Institutions of higher education strive to offer more flexible access to education in an effort to keep pace with changing socioeconomic forces, such as globalization and the advanced capability of electronic communication, that are increasing student mobility.[106] Often, the best way to achieve this is by establishing new physical teaching sites and/or offering online educational experiences. However, delivering a comparable education to different audiences, each with their own individual and group characteristics, poses significant challenges. Nevertheless, if those challenges can be overcome, the benefits of reaching additional audiences can be worthwhile to both the institution and respective students.

The core definition of *distributed education*, as analyzed by Lee Harvey, is [a delivery of education that] "… occurs when the teacher and student are situated in separate locations and learning occurs through the use of technologies (such as video and internet), which many be part of a wholly distance education program or supplementary to traditional instruction."[107] From this definition, it is clear that two important factors influence distributed education: a difference in physical location between teacher and student and the use of technology to bridge that distance. Expanded access is not the only advantage to distributed education though. Other advantages include alleviating capacity restraints, capitalizing on emerging market opportunities, and serving as a catalyst for institutional transformation.[108]

Opportunities for distributed education (also sometimes referred to as distributed learning) can be created between groups within a single institution location and anywhere online or between groups located between an institution's home and another campus site. An example of distributed education within institutions (*i.e.*, within the home site) can include individual courses that are simultaneously offered in online and face-to-face environments. In this case, two different groups of students take the same course but one group is interacting with the instructor face-to-face and utilizing the institution's physical resources and the other group is interacting with the instructor through internet and/or video communications typically without any physical institution resources. Regardless, each group, in theory, is being offered a comparable education through albeit disparate physical resources and delivery methods. This model of education, referred to as distance education, is one type of distributed education.

Although the terms *distance education* and *distributed education* are frequently used simultaneously, distributed education *includes* distance education and is not limited to strictly online instruction. While the above describes a distributed-distance education model for providing education to additional individual users through online courses, another model of distributed education provides additional groups of students a simultaneous learning experience and similar learning environment (and physical resources) in a different location. Many universities have satellite sites that allow them to deliver traditional face-to-face educational experiences remotely, understanding that the students who pursue this opportunity may have otherwise not been able to relocate to the home site for the experiences. Satellite sites are often established through partnerships with other institutions already established at the desired locations. While technological communication is essential in distance education, it is common but not necessarily essential for this type of distributed education model. While satellite sites often communicate with home sites through internet/video, they have their own teaching faculty or adjunct faculty from the host institution to supplement internet/video instruction.

While distributed education offers an advantage for delivering learning experiences to “anyone, anywhere,” complex challenges are associated with delivering these experiences. Challenges include IT support, costs for additional resources, student accessibility to support resources, and accreditation.[109–111] Providing a comparable distributed education presents many more challenges than simply offering a course in another location; while many home site resources can be accessed via internet, telephone, or video, many cannot. Aside from these and other related challenges, one of the most important challenges lies in delivering an education for which comparability across educational sites can be supported by evidence.

Regardless of the model for delivering distributed education, ensuring the delivery of a comparable education between sites or delivery platforms is critical. In addition to the fact that comparability of education across locations is a "best practice" in higher education, regional and programmatic accrediting bodies also enforce these principles. For example, the Liaison Committee of Medical Education (LCME) reviews and certifies (i.e., "accredits") the quality of medical programs in the United States and Canada, and within its 12 standards for accreditation, Standard 8.7 (Comparability of Education/Assessment) of the states, "A medical school ensures that the medical curriculum includes comparable educational experiences and equivalent methods of assessment across all locations within a given course and clerkship to ensure that all medical students achieve the same medical education program [learning objectives]."[112]

Other accrediting bodies dictate the needs for similar behaviors about distance education specifically. For example, Standard III (Program Outcomes, Curricula, and Materials) for the Distance Education Accrediting Commission (DEAC) states, "The effective design of program outcomes, curricula, and supplemental materials results in cohesive educational offerings and evaluation methods of student learning that are clearly connected to the stated [learning] [objectives]." Subheading H (Examinations and Other Assessments) of this standard goes on to state, "Examinations and other assessment techniques provide adequate evidence of the achievement of stated learning [objectives]. The institution implements grading criteria that it uses to evaluate and document student attainment of learning [objectives]."[113] Similarly, the Western Interstate Commission for

Higher Education (WICHE) guides regional accrediting bodies to review whether "the institution evaluates the educational effectiveness of its distance education programs (including assessments of student learning [objectives], student retention, and student satisfaction) to ensure comparability to campus-based programs,"[110] and the Council for Higher Education Accreditation (CHEA) and the National Center for Higher Education Management Systems (NCHEMS) have student outcomes and attainment standards that dictate, "All assessment methods and instruments used to determine student achievement strive toward being valid, reliable, and demonstrably linked to the learning [objectives] they purport to cover."[114] In reviewing each of the LCME, DEAC, WICHE, CHEA and NCHEMS accreditation standards for central commonalities, we conclude that the presentation and proper use of learning objectives is absolutely crucial for ensuring comparability in distributed education.

Evidence of comparability is not difficult to produce if well-written course learning objectives and a plan for assessing those learning objectives has been established and implemented into the course and are consistent across locations. Not only then can performances be compared between different populations in the same course but also between different populations in different courses.[115] This strategic and purposeful assessment of specific learning objectives is important in educational environments in which competencies must be demonstrated for certification and/or licensure.[116]

Motivations and Strategies for Learning

Meeting the needs of students from diverse backgrounds and broad demographic characteristics also makes delivering a comparable distributed education challenging. While assessing basic knowledge minimally requires measurements of information correctness and belief justification, the learning process and cognitive application in performances can be positively or negatively impacted by many motivations, learning strategies, and other intrinsic and extrinsic factors characteristic of individuals and populations.

Paul Pintrich and Wilbert McKeachie, among others, established a contextualized, social-cognitive paradigm of learning that examined the effect of student motivations and learning strategies on course-specific cognitive processes. While the concept that social factors influenced students' cognitive processes was not new, empirical links between the two were not yet clearly established and the claims that had been made were heavily criticized for lacking a basic theoretical framework. This led Pintrich to develop a tool for assessing students' motivations and learning strategies in order to help improve student learning. In 1986, he began the formal development of the 81-question self-report tool named the *Motivated Strategies for Learning Questionnaire* (MSLQ), which he published in 1991 after completing sufficient reliability and validation studies using the instrument.[117]

The MSLQ contains two sections: motivations and learning strategies. The two sections are further subdivided into 5 total constructs and 15 collective scales. *Value*, *expectancy*, and *affect* are the three general constructs that form the basis of MSLQ motivation scales. *Value components* refer to motivational scales that measure how much of a student's motivation to learn comes from their desire to learn and master material (intrinsic goal orientation), to earn good grades or the approval of others (extrinsic goal orientation), and to fulfill an importance or usefulness (task value). *Expectancy components* refer to motivational scales that measure how much a student believes that the learning outcomes that will result from a learning experience are contingent on their own efforts alone (control beliefs) and to what degree students expect that they will succeed at academic tasks (self-efficacy for learning and performance). Lastly, the *affective component* measures how much students worry about taking exams and how they believe that anxiety affects their academic performance (test anxiety).[117,118]

The learning strategies scales are categorized into two constructs: *cognitive and metacognitive strategies*, and *resource management strategies*. *Cognitive and metacognitive strategies* refer to the degree to which each student uses common strategies for learning. These strategies (scales) include reviewing information until they have memorized it for the short term (rehearsal), summarizing and making connections between material for the long term (elaboration), prioritizing and outlining information (organization), using existing knowledge to evaluate new information (critical thinking), and practicing control and awareness of one's own level of thoughts, knowledge, and reasoning (metacognitive self-regulation). *Resource management strategies* refer to how well each student regulates resources that are helpful for learning. These scales measure how well students regulate their study time and setting (time and study environment), persist through difficult or boring tasks (effort regulation), seek help from other students (peer learning), or pursue assistance from an instructor when needed (help seeking).[117,118]

The MSLQ collects a considerable amount of information on all 15 scales in only 81 total questions that take approximately 20-30 minutes to answer. The resulting information offers valuable feedback to both students and instructors. Pintrich provides suggestions for how to improve in each of the 15 scales in his manual for the use of the MSLQ.[118] These suggestions help students improve their performances and also help instructors mentor students accordingly.

For students, MSLQ results can validate the influence and effectiveness of their motivations and strategies for learning and offer rationale for making changes for improvement. However, just as the MSLQ reflects individual differences, each student accepts feedback differently; some openly welcome it while some despise it or do not know how to use it. Pintrich presumably knew this since he provided both individual feedback in addition to group feedback on these scales to the students who participated in his studies of the tool. Additionally, a "one shoe fits all" use and approach conflicts with the intent of the MSLQ. As needed for any form of feedback to be beneficial, Pintrich and other educator scholars note the requirement for students to be willing to accept the feedback and strive to use that feedback in order for it to improve their learning.[119] While individual scores are more important for student development and mentorship, whole-class group results can be of greater value to the instructor.

Whole-class average scores on the 15 MSLQ scales can tell an instructor a lot about their class – enough, in fact, to make predictions on how well the class will perform as a whole since motivations and learning strategies has been proven to be empirically linked to performance.[120] Therefore, if instructors can obtain MSLQ results early enough, they can adjust their course variables (pedagogical methods, assessment methods, *etc.*) to best suit the needs of the class. In that respect, many institutions have implemented the MSLQ as a 'needs assessment'.[117] Understanding the characteristics of a student population in a course can be crucial to interpreting and understanding performance outcomes.

Other similar socio-cognitive assessment tools have been created for similar purposes. For instance, the Learning and Study Strategies Inventory (LASSI) was developed in 1987 by Claire Weinstein to more generally assess students' awareness of their use of learning and study strategies than the MSLQ,[121] the Multidimensional Self-Concept Scale (MSCS) questionnaire was developed in 1992 by Bruce Bracken to evaluate six subscales of self-concept,[122] and the Academic Self-regulated Learning Scale (A-SRL-S) was established in 2010 by Carlo Mango to measure self-regulation in higher education.[123] Mango even used the MSLQ and LASSI to validate his A-SRL-S in 2011.[124] Despite the advantages of each of these tools, the extensively validated and reliable MSLQ is the most widely used. Despite scientific scrutiny, study after study, including a single review of its use in 56 empirical studies,[117] agrees that the MSLQ is a valid and reliable tool.[125–127]

While the MSLQ is found to be valid and reliable, a significant criticism it receives is that its "expected" results can vary greatly for students between courses. However, Pintrich acknowledges these variable results and explains that they in fact support the socio-cognitive paradigm he used to create the MSLQ. According to this framework, motivations and learning strategies are dynamic, contextually bound, and controlled by the student, and therefore the student should express their motivations and learning strategies differently between courses depending on their interest, relative self-efficacy, *etc.* in the course subject.[117] Another advantage to the MSLQ that likely contributes to its popularity is its modular organization which allows easy customization. Because each scale is independent to the others, an instructor can choose to eliminate scales in which they are less interested. Additionally, like the other similar tools, the MSLQ can be administered electronically, making data collection and results analysis feasible.

Since its conception, the MSLQ has been studied extensively. It offers empirical links between student's behavior and cognitive processes and supports the concept of evidence-based, socio-cognitive learning paradigms.

Purpose of the Current Study

The purpose of the current study is to investigate the effects on student learning and final grades of implementing a confidence-based, individualized strategy to remediate academic deficiency. The experimental setting is OT 422, a human anatomy lecture and laboratory course in the University of North Dakota's occupational therapy professional masters curriculum. The study's hypotheses are:

1. A confidence-based, individualized remediation strategy increases student learning.
2. Self-assessment of confidence-based academic performances increases student learning via remediation.
3. Student motivations, learning strategies and academic performances in a human anatomy curriculum are comparable across distributed campus sites.

References

1. Heywood, J. *Assessment in higher education : student learning, teaching, programmes, and institutions*. (J. Kingsley Publishers, 2000).

2. Marchese, T. J. Third Down, Ten Years to Go. *Am. Assoc. High. Educ. Bull.* **40**, 3–8 (1987).

3. Palomba, C. A. & Banta, T. W. *Assessment essentials : planning, implementing, and improving assessment in higher education*. (Jossey-Bass Publishers, 1999).

4. Jankowski, N. Assessing Student Learning Outcomes A: Best Practices in Assessment and Accredidation.

5. McAlpine, M. *Principles of assessment*. (CAA Centre, University of Luton, 2002).

6. Scanlan, C. Assessment, Evaluation, Testing and Grading. Available at: https://resource.mccneb.edu/edutut/online pets course/assessment_evaluation_grading.htm. (Accessed: 16th March 2019)

7. Moss, P. A., Pullin, D. C., Gee, James, P., Haertel, E. H. & Young, L. J. Sociocultural Implications for the Practice of Assessment I: Classroom Assessment. in *Assessment, Equity, and Opportunity to Learn* 222–258 (Cambridge University Press, 2008).

8. Shepard, L. A. Commentary: Evaluating the Validity of Formative and Interim Assessment. *Educ. Meas. Issues Pract.* **28**, 32–37 (2009).

9. Assessing Student Learning. *Technology, Academic Technology Division of Information, University of Wisconsin-Madison* Available at: https://at.doit.wisc.edu/wp-content/uploads/2016/05/R2C-Assessments.pdf.

10. Pellegrino, J. W., Chudowsky, N. & Glaser, R. *Knowing What Students Know: The Science and Design of Educational Assessment*. (National Academies Press, 2001).

doi:10.17226/10019

11. Trumbull, E. & Lash, A. Understanding Formative Assessment: Insights from Learning Theory and Measurement Theory. *West Ed* (2013).

12. Gagne, R. M. *Conditions of Learning*. (Holt, Rinehart, and Winston, 1985).

13. Heritage, M. Learning Progressions: Supporting Instruction and Formative Assessment. in *Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO)* 3 (CCSSO, 2008).

14. Vygotsky, L. S. *Mind in Society: The Development of Higher Psychological Processes*. (Harvard University Press, 1978).

15. Maslow, A. H. *The Psychology of Science*. (Harper & Row, 1966).

16. Brissenden, G. & Slater, T. Assessment Primer. *National Institute for Science Education (NISE), College Level One (CL-1) Team at University of Wisconsin-Madison. Field-tested Learning Assessment Guide.* Available at: http://archive.wceruw.org/cl1/flag/extra/contact.htm.

17. Angelo, T. A. & Cross, K. P. *Classroom assessment techniques : a handbook for college teachers*. (Jossey-Bass, 1993).

18. Apple, D. K. & Krumsieg, K. *Process Education Teaching Institute Handbook*. (Pacific Crest, 1998).

19. Rea, J. B. You Say Ee-ther and I Say Eyether: Clarifying Assessment and Evaluation. *Am. Soc. Train. Dev. Links Newsletter.* (2010).

20. Lundquist, A. E. Assessment, Evaluation and Research Relationships and Definitions in the Field of Student Affairs. *University of Texas at El Paso* Available at:

https://www.utep.edu/student-affairs/_Files/docs/Campus-Labs-Assessment-Evaluation-Research-Definitions-Handout.pdf.

21. Assessment and Grading. Available at: http://www.mathimp.org/publications/teacher/teach10.html.

22. Grading vs. Assessment of Learning Outcomes: What's the Difference? *Carnegie Mellon University* Available at: http://www.cmu.edu/teaching/assessment/howto/basics/grading-assessment.html.

23. Grading and Assessment. *Assessment Commons* Available at: http://assessmentcommons.org/grading-and-assessment/. (Accessed: 1st June 2019)

24. Grading definition and meaning. *Collins English Dictionary* Available at: https://www.collinsdictionary.com/us/dictionary/english/grading. (Accessed: 1st June 2019)

25. Assessing Learning Outcomes: Grading versus Assessing. *Champlain College* Available at: https://champlain.instructure.com/courses/200147/pages/grading-versus-assessing. (Accessed: 1st June 2019)

26. Glaser, R. Instructional technology and the measurement of learing outcomes: Some questions. *Am. Psychol.* **18**, 519–521 (1963).

27. Grading Systems. *University of North Carolina at Charlotte, Center for Teaching and Learning* Available at: https://teaching.uncc.edu/sites/teaching.uncc.edu/files/media/files/file/AssessmentAndGrading/GradingSystems.pdf.

28. Bresee, C. W. On "Grading on the Curve". *Clear. House A J. Educ. Strateg. Issues Ideas* **50**, 108–110 (1976).

29. Wall, C. R. Grading on the Curve. *InCider* **5**, 83–85 (1987).

30. Kulick, George|Wright, R. The Impact of Grading on the Curve: A Simulation Analysis. *Int. J. Scholarsh. Teach. Learn.* **2**, (2008).

31. Tomlinson, C. A. & McTighe, J. *Integrating Differentiated Instruction & Understanding by Design: Connecting Content and Kids*. (Association for Supervision and Curriculum Development (ASCD), 2006).

32. Schinske, J. & Tanner, K. Teaching More by Grading Less (or Differently). *CBE - Life Sci. Educ.* **13**, 159–166 (2014).

33. Grading vs. Assessment (Formative and Summative). *Mount Holyoke College* Available at: https://www.mtholyoke.edu/teachinglearninginitiatives/grading-vs-assessment. (Accessed: 1st June 2019)

34. Kuntz, B. Focus on Learning, Not Grades. *Association for Supervision and Curriculum Development (ASCD) Education Update* **54**, (2012).

35. Grades vs. Learning - Shifting Attention to What's Important. *The Graide Network* (2018). Available at: https://www.thegraidenetwork.com/blog-all/2018/8/1/retiring-the-red-pen-shifting-attention-from-grades-to-learning.

36. Butler, R. & Nisan, M. Effects of No Feedback, Task-Related Comments, and Grades on Intrinsic Motivation and Performance. *J. Educ. Psychol.* **78**, 210–216 (1986).

37. Chamberlin, K., Yasué, M. & Chiang, I.-C. A. The impact of grades on student motivation. *Act. Learn. High. Educ.* 146978741881972 (2018). doi:10.1177/1469787418819728

38. Townsley, M. What is the Difference between Standards-Based Grading (or Reporting) and Competency-Based Education? *Competency Works; Learning from the*

Cutting Edge (2014). Available at: https://www.competencyworks.org/analysis/what-is-the-difference-between-standards-based-grading/.

39. Richman, W. A. & Ariovich, L. All-in-One: Combining Grading, Course Program, and General Education Outcomes Assessment. *Natl. Inst. Learn. Outcomes Assess.* **Occasional**, (2013).

40. Salisbury, M. Grades and Assessing Learning: Can't We Get Along? *Insid. High. Ed.* (2012).

41. McClendon, K. & Eckert, E. Grades as Documentation of SLO Achievement: Constructing an Outcomes-Based Grading System. *[PowerPoint slides.] California Association for Institutional Research (CAIR)* (2007). Available at: https://cair.org/wp-content/uploads/sites/474/2015/07/McClendon-Eckert-1.pdf.

42. Kissel, H., Miller, B. J. & Young, H. *Writing Objectives*. (James Madison University).

43. What is the difference between course objectives and learning outcomes? *San Francisco State University* Available at: https://ueap.sfsu.edu/sites/default/files/assets/docs/student_learning_outcomes.pdf.

44. Goals vs. Objectives. *Weber State University* Available at: https://weber.instructure.com/courses/307280/pages/goals-vs-objectives.

45. Guidelines for Including Learning Outcomes on Course Syllabi. *American University* Available at: https://american.edu/ocl/volunteer/upload/Guidelines-for-Including-Learning-Outcomes-on-Course-Syllabi.pdf.

46. Steere, D. E. & Cavaiuolo, D. Connecting Outcomes, Goals, and Objectives in Transition Planning. *Teach. Except. Child.* **34**, 54–59 (2002).

47. University of Connecticut. Writing Learning Objectives. (2014). Available at:

https://kb.ecampus.uconn.edu/2014/07/31/writing-cognitive-objectives/.

48. Williams, B. *Writing Objectives*. (Penn State University).

49. Dalto, J. ABCD: The Four Parts of a Learning Objective. (2013). Available at: https://www.convergencetraining.com/blog/abcd-the-four-parts-of-a-learning-objective.

50. White, R. ABCD's of SMART Objectives. *Louisiana State University* Available at: https://www.slideshare.net/bwhitelsu/abcds-of-smart-objectives. (Accessed: 19th March 2017)

51. Mager, R. F. *Preparing Instructional Objectives*. (Fearon Publishers, 1962).

52. Dalto, J. Robert Mager's Performance-Based Learning Objectives. (2014). Available at: https://www.convergencetraining.com/blog/robert-magers-performance-based-learning-objectives.

53. Shabatura, J. Using Bloom's Taxonomy to Write Effective Learning Objectives. *Teaching Innovation & Pedagogical Support* (2018). Available at: https://tips.uark.edu/using-blooms-taxonomy/. (Accessed: 16th March 2019)

54. International Assembly for Collegiate Business Education (IACBE). Bloom's Taxonomy of Educational Objectives and Writing Intended Learning Outcomes Statements. (2016).

55. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. (David McKay Company, Inc., 1956).

56. Dalto, J. How to Write SMART Learning Objectives. (2013). Available at: https://www.convergencetraining.com/blog/how-to-write-smart-learning-objectives.

57. Doran, G. T. There's a S.M.A.R.T. way to write management's goals and objectives. *Management review* **70**, 35–36 (1981).

58. Drucker, P. *The Practice of Management*. (Revised edn., Butterworth-Heinemann, 2007, 1954).

59. Bryan, W., DiMartino, J. & Center for Secondary School Redesign. *Writing Goals and Objectives: A Guide for Grantees of the Smaller Learning Communities Program*. (United States Department of Education. Academy for Educational Development, 2010).

60. Center for Disease Control; Division for Heart Disease and Stroke Prevention; State Heart Disease and Stroke Prevention Program. *Evaluation Guide: Writing SMART Objectives*. (Department of Health and Human Services. Center for Disease Control and Prevention, 2013).

61. W.K. Kellog Foundation. Logic Model Development Guide. (2004).

62. Harris, E. & Harvard Family Research Project. Afterschool Evaluation 101: How to Evaluate an Expanded Learning Program. (2011).

63. Bjerke, M. B. & Renger, R. Being smart about writing SMART objectives. *Eval. Program Plann.* **61**, 125–127 (2017).

64. Tofade, T., Khandoobhai, A. & Leadon, K. Use of SMART Learning Objectives to Introduce Continuing Professional Development Into the Pharmacy Curriculum. *Am. J. Pharm. Educ.* **76**, 68 (2012).

65. Merriam-Webster Dictionary. Definition of Learning. *Merriam Webster* Available at: https://www.merriam-webster.com/dictionary/learning.

66. Malamed, C. 10 Definitions of Learning. Available at:

http://theelearningcoach.com/learning/10-definitions-learning/.

67. Hunt, D. P. The concept of knowledge and how to measure it. *J. Intellect. Cap.* **4**, 100–113 (2003).

68. Merriam-Webster Dictionary. Definition of Intelligence. *Merriam Webster* Available at: https://www.merriam-webster.com/dictionary/intelligence.

69. Ben-Simon, A., Budescu, D. V. & Nevo, B. A Comparative Study of Measures of Partial Knowledge in Multiple-Choice Tests. *Appl. Psychol. Meas.* **21**, 65–88 (1997).

70. Bruno, J. Information Reference Testing (IRT) in Corporate and Technical Training Programs. *UCLA* (1995).

71. Bruno, J. E. Using Testing to Provide Feedback to Support Instruction: A Reexamination of the Role of Assessment in Educational Organizations. in *Item Banking: Interactive Testing and Self-Assessment* 190–209 (Springer Berlin Heidelberg, 1993). doi:10.1007/978-3-642-58033-8_16

72. Alnabhan, M. An Empiracle Investigation of the Effects of Three Methods of Handling Guessing and Risk Taking on the Psychometric Indices of a Test. *Soc. Behav. Personal. an Int. J.* **30**, 645–652 (2002).

73. Ghadermarzi, M., Yazdani, S., Pooladi, A., Bahram-Rezaei, M. & Hosseini, F. A Comparative Study between the Conventional MCQ Scores and MCQ with the CBA Scores at the Standardized Clinical Knowledge Exam for Clinical Medical Students. *J. Med. Educ.* **14**, 31–37 (2015).

74. Bush, M. Reducing the need for guesswork in multiple-choice tests. *Assess. Eval. High. Educ.* **40**, 218–231 (2015).

75. Adams, T. M. & Ewen, G. W. The Importance of Confidence in Improving

Educational Outcomes. in *University of Wisconsin System's 25th Annual Conference on Distance Teaching and Learning* 1–5 (2009).

76. Jaradat, D. & Sawaged, S. The Subset Selection Technique for Multiple-Choice Tests: An Empirical Inquiry. *J. Educ. Meas.* **23**, 369–76 (1986).

77. Lord, F. M. Formula Scoring and Number-Right scoring. *J. Educ. Meas.* **12**, 7–11 (1975).

78. Otoyo, L. & Bush, M. Addressing the Shortcomings of Traditional Multiple-Choice Tests: Subset Selection Without Mark Deductions. *Pract. Assessment, Res. Eval.* **23**, (2018).

79. Frary, R. B. Formula Scoring of Multiple-Choice Tests (Correction for Guessing). *Educ. Meas. Issues Pract.* **7**, 33–38 (1988).

80. Crocker, L. & Algina, J. *Introduction to Classical and Modern Test Theory.* (Holt, Rinehart, and Winston, 1986).

81. Magnusson, D. *Test theory.* (John Addison-Wesley Publishing Company& Sons, Ltd, 1967).

82. Diamond, J. & Evans, W. The Correction for Guessing. *Rev. Educ. Res.* **43**, 181–191 (1973).

83. Traub, R. E. & Hambleton, R. K. The Effect of Scoring Instructions and Degree of Speededness on the Validity and Reliability of Multiple-Choice Tests. *Educ. Psychol. Meas.* **32**, 737–758 (1972).

84. Budescu, D. & Bar-Hillel, M. To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. *J. Educ. Meas.* **30**, 277–291 (1993).

85. Dressel, P. L. & Schmid, J. Some Modifications of the Multiple-Choice Item. *Educ.*

Psychol. Meas. **13**, 574–595 (1953).

86. Bush, M. A Multiple Choice Test that Rewards Partial Knowledge. *J. Furth. High. Educ.* **25**, 157–163 (2001).

87. Akeroyd, M. Progress in Multiple Choice Scoring Methods 1977-81. *J. Furth. High. Educ.* **6**, 86–90 (1982).

88. Yaniv, I. & Schul, Y. Elimination and Inclusion Procedures in Judgment. *J. Behav. Decis. Mak.* **10**, 211–220 (1997).

89. Coombs, C. H., Milholland, J. E. & Womer, F. B. The Assessment of Partial Knowledge. *Educ. Psychol. Meas.* **16**, 13–37 (1956).

90. Warwick, J., Bush, M. & Jennings, S. Analysis and Evaluation of Liberal (Free-Choice) Multiple-Choice Tests. *Innov. Teach. Learn. Inf. Comput. Sci.* **9**, 1–12 (2010).

91. Hammond, E. J., McIndoe, A. K., Sansome, A. J. & Spargo, P. M. Multiple-choice examinations: adopting an evidence-based approach to exam technique. *Anaesthesia* **53**, 1105–8 (1998).

92. Confidence | Definition of ‘Confidence’. *Oxford Dictionary* (2019).

93. Hevner, K. A Method of Correcting for Guessing in True-False Tests and Empirical Evidence in Support of IT. *J. Soc. Psychol.* **3**, 359–362 (1932).

94. Belal, A. & Ammar, D. Confidence-Based Assessment of Two-Alternative Format Tests. in *Paperpresented at the 5th Conference of Learning International Networks Consortium (LINC) MIT* (2010).

95. Weiss, B., Weiss, B., Gridling, G. & Trödh, C. Embedded systems exams with true/false questions: A case study. in *VIENNA UNIVERSITY OF TECHNOLOGY* (2006).

96. Gardner-Medwin, A. R. Confidence assessment in the teaching of basic science. *ALT-J* **3**, 80–85 (1995).

97. Gardner-Medwin, A. R. Confidence-Based Marking: Encouraging rigour through assessment. in *University of Bristol. Proceedings of The Physiological Society* **J Physiol 567P**, **WA10**, (Physiological Society, 2005).

98. Gardner-Medwin, A. & Curtin, N. Certainty-Based Marking (CBM) for reflective learning and proper knowledge assessment. in *Re-Engineering Assessment Practices (REAP) Int. Online Conference on Assessment Design for Learner Responsibility, Proceedings for Raising students' meta-cognition (self-assessment) abilities* (2007).

99. Farrell, G. A comparison of an innovative web-based assessment tool utilizing confidence measurement to the traditional multiple choice, short answer and problem solving questions. in *10th CAA : International Computer Assisted Assessment conference : proceedings of the conference on 4th &amp; 5th July 2006 at Loughborough University* (Professional Development, Loughborough University, 2006).

100. Farrell, G. & Leung, Y. Convergence of validity for the results of a summative assessment with confidence measurement and traditional assessment. in *12th CAA International Computer Assisted Assessment Conference : Proceedings of the Conference on 8th and 9th July 2008 at Loughborough University.* 195–204 (Loughborough University, 2008).

101. Gritten, F. & Johnson, D. M. Individual differences in judging multiple-choice questions. *J. Educ. Psychol.* **32**, 423–430 (1941).

102. Gardner-Medwin, A. R. & Gahan, M. Formative and Summative Confidence-Based

Assessment. 147–155 (2003).

103. Gardner-Medwin, A. Confidence-Based Marking - towards deeper learning and better exams. In: Bryan, C and Clegg, K, (eds.) Innovative Assessment in Higher Education. in 141–149 (Routledge, Taylor and Group; Francis, 2006).

104. Dictionary.com. Definition of Metacognition. Available at: https://www.dictionary.com/browse/metacognition.

105. Kruger, J. & Dunning, D. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *J. Pers. Soc. Psychol.* **77**, 1121–1134 (1999).

106. Matheos, K. & Archer, W. From distance education to distributed learning surviving and thriving. (2004).

107. Harvey, L. Analytic Quality Glossary. Available at: http://www.qualityresearchinternational.com/glossary/.

108. Oblinger, D. & Kidwell, J. Distance Learning: Are We Being Realistic? *Educ. Rev.* **35**, 30–34 (2000).

109. McGee, P., Carmean, C. & Jafari, A. Distributed Learning: Making Systems that Work. in *EdMedia + Innovate Learning - World Conference on Educational Multimedia, Hypermedia & Telecommunications* (eds. C. Montgomerie & J. Seale) **2007**, 1360–1364 (Association for the Advancement of Computing in Education (AACE), 2007).

110. Oblinger, D., Barone, C. & Hawkins, B. *Distributed Education and Its Challanges: An Overview*. (American Council on Education, 2001).

111. Brian Hawkins. Distributed Learning and Institutional Restructuring. *Educom Rev.* **34**,

(1999).

112. Liaison Committee on Medical Education. Functions and structure of a medical school. (2019). Available at: http://lcme.org/publications/#Standards.

113. Commission, D. E. A. Accredidation Handbook. Available at: https://www.deac.org/UploadedDocuments/Handbook/DEAC_Accreditation_Handbook.pdf.

114. National Center for Higher Education Management Systems (NCHEMS). *The Competency Standards Project: Another Approach to Accreditation Review. (CHEA Occasional Paper)*. (Council for Higher Education Accreditation (CHEA), 2000).

115. Lovato, C. & Murphy, C. Comparability of student performance and experiences in UBC's distributed MD undergraduate program: The first 2 years of implementation. *BC Med. J.* **50**, 380–383 (2008).

116. Fletcher, J. D., Tobias, S. & Wisher, R. A. Learning Anytime, Anywhere: Advanced Distributed Learning and the Changing Face of Education. *Educ. Res.* **36**, 96–102 (2007).

117. Duncan, T. G. & McKeachie, W. J. The Making of the Motivated Strategies for Learning Questionnaire. *Educ. Psychol.* **40**, 117–128 (2005).

118. Pintrich, P. R., Smith, D. A., Garcia, T. & McKeachie, W. J. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). (1991).

119. Pintrich, P. R. & de Groot, E. V. Motivational and self-regulated learning components of classroom academic performance. *J. Educ. Psychol.* **82**, 33–40 (1990).

120. Robbins, S. B. *et al.* Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychol. Bull.* **130**, 261–288 (2004).

121. Weinstein, C. E., Palmer, D. & Schulte, A. C. Learning and Study Strategies Inventory (LASSI). *Clear. FL H H Publ.* (1987).

122. Bracken, B. A. & Pro-Ed (Firm). *MSCS : Multidimensional Self Concept Scale*. (Pro-Ed, 1992).

123. Mango, C. Assessing and Developing Self-regulated Learning. in *The Assessment Handbook* 26–42 (Philippine Educational Measurement and Evaluation Association, 2009).

124. Mango, C. Validating the Academic Self-Regulated Learning Scale with the Motivated Strategies For Learning Questionnaire (MSLQ) and Learning and Study Strategies Inventory (LASSI). *Int. J. Educ. Psychol. Assess.* **7**, (2011).

125. ERTURAN İLKER, G., ARSLAN, Y. & DEMİRHAN, G. A Validity and Reliability Study of the Motivated Strategies for Learning Questionnaire. *Educ. Sci. Theory Pract.* **14**, (2014).

126. Credé, M. & Phillips, L. A. A meta-analytic review of the Motivated Strategies for Learning Questionnaire. *Learn. Individ. Differ.* **21**, 337–346 (2011).

127. Pintrich, P. R., Smith, D. A. F., Garcia, T. & Mckeachie, W. J. Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educ. Psychol. Meas.* **53**, 801–813 (1993).

CHAPTER II

EFFECTS OF A CONFIDENCE-BASED, INDIVIDUALIZED REMEDIATION STRATEGY ON STUDENT LEARNING AND FINAL GRADES

Abstract

Background

Totaling the number of correct answers on high-stakes course examinations represents a one-dimensional approach to measuring student knowledge and learning. While one easy and generally accepted method of assigning final grades is to compare students' total points accumulated from formal assessments to a predetermined scale, confidence-based assessments (CBAs) measure knowledge and learning multidimensionally and include measures of both correctness and a belief justification (*e.g.*, confidence) to differentiate performances indicating complete, partial, absent, and flawed knowledge. CBAs can also detect student awareness of absent knowledge (low confidence for an incorrect answer), misinformation (high confidence for an incorrect answer), guesswork (low confidence for a correct answer), as well as content mastery (high confidence for a correct answer). The present study investigated whether employing a method of individualized re-examination in which student confidence was taken into account to remediate poor academic performance impacted student learning (in direct relation to stated learning objectives) and final grades.

Methods

On each of six high-stakes human anatomy lecture (selected-response) and cadaver laboratory (constructed-response) examinations, three cohorts of occupational therapy

students reported their level of confidence in their answers to each examination question using a 3-point Likert-type scale. Each examination question was specifically linked to a student learning objective (LO). After each examination, each student received written feedback (2016 cohort) or completed a self-assessment exercise (2017 and 2018 cohorts) regarding the LOs on which they performed poorest based on their confidence in and correctness of their answers. On the subsequent examination, each student in the 2016 and 2017 cohorts was administered six Individualized Remediation Questions (IRQs) retesting each of six identified LOs identified by the instructor by using a pre-determined scale of six possible confidence-based performances ordered by representative levels of knowledge. Students in the 2018 cohort were given Standardized Remediation Questions (SRQs) based on the previous two years' IRQs to compare the impact of an individualized remediation strategy with a standardized group remediation strategy. Examinations were scored and remediation performance was analyzed.

Results

For each cohort, students performed with higher confidence and greater correctness on their remediation questions. Upon remediating their poorest performances, students typically achieved mean post-remediation (post-R) performances between 1 and 2 performance levels (PLs) above their mean pre-R PL. Although no significant differences were found to exist in the number of desirable remediation performances (RP) earned by students between any of the years, greater mean remediation scores resulted from IRQs in 2016 and 2017 vs. SRQs in 2018 ($p < 0.001$). IRQs also better influenced learning for students with lower grade percentages, while IRQs and SRQs tended to equally help students with higher grade percentages. As a direct result of increased learning, IRQs and SRQs positively influenced LO achievement; the LOs retested by IRQs resulted in a mean remediation of over 2 PLs, whereas those that were retested by SRQs only resulted in a mean remediation of just under 1.5PLs ($p<0.001$). This positive remediation of LOs led to 68% of them reaching new achievement (defined as reaching a mean PL of partial knowledge) in 2016, 54.5% in 2017, and 43.1% in 2018. Despite these desirable results, students perceive the remediation strategy negatively affects their learning and LO achievement. Lastly, the interventions decreased final percentage and letter grades ($p<0.001$) as a result remediating poor correct performances to better incorrect performances.

Conclusions

This study demonstrates how a novel use of confidence-based performances can be used to create an individualized remediation intervention that positively influences student learning and LO achievement. Fundamentally, the positive remediation did not cause the decrease in grade; the decrease in grade was an effect of more accurately assessing students' actual levels of knowledge and learning after remediation efforts. Better understanding student confidence and attitudes toward learning may inform other strategies to successfully remediate poor academic performance.

Introduction

The typical and basic requirement for academic assessments is the ability to measure student knowledge. However, assessments that only measure correctness of responses can be deficient in their ability to accurately measure knowledge. As a result, final percentages and letter grades from widely-used "number right" scoring methods do not accurately or reliably convey student knowledge. Knowledge is "a justified belief in true information,"[1] so one must measure and compare information trueness (or *correctness*) in association with belief justification (*e.g., confidence*) to fully ascertain knowledge. Because knowledge is multidimensional, assessments cannot accurately quantify knowledge unless both components are measured.

Educators have long studied confidence as one of many variables that influences learning and assessment. Confidence-Based Assessments (CBAs) are assessments that accurately measure knowledge by evaluating both student confidence in the correctness and actual correctness of their answers. In doing this, CBAs compare what students *think* they know to what they *actually* know. CBAs can detect the four levels of knowledge (*flawed*, *absent*, *partial*, and *complete*) and provide valuable interpretations of student performances, such as student awareness of lack of knowledge (low confidence for an incorrect answer), unawareness of misinformation (high confidence for an incorrect answer), guesswork (low confidence for a correct answer), as well as content mastery (high confidence for a correct answer). While correctness-only assessments are only able to interpret performances as *absent* or *complete* knowledge, CBAs can also detect *partial* and *flawed* knowledge, even from what would may have otherwise been considered *absent* or *complete*.

While the principles of CBAs have been used to more accurately assess knowledge by centuries-old philosophers, the first published case of the principles being used to assess academic knowledge was by Kate Hevner in 1932. Nearly six decades later, James Bruno established a method known as Information Resource Testing (IRT), which was later renamed to Confidence-Based Assessment (CBA).[2] Years later, A. R. Gardner-Medwin further developed CBAs as he popularized the method and principles for using them after showing years of compelling data.[3–5] Largely due to being able to detect important issues such as misinformation and guesswork, other studies have gone on to show how CBAs can ensure valid, more accurate assessments,[6] facilitate reflection for deeper learning among students,[2] and improve educational outcomes,[7] all of which are principles that contribute to the present study.

The objective of this study was to employ the principles of evidence-based assessments and CBAs to an individualized remediation strategy for students in an Occupational Therapy (OT) anatomy course (OT 422: Anatomy for Occupational Therapists). While reported confidence levels traditionally weigh into credit awarded for examination answers in CBAs,[3] we propose the use of unweighted confidence levels is necessary for using our suggested re-ordering of the CBA performance levels (PLs) to detect poor performances on LOs. Students would then be given feedback on their poorest LO performances in the form of an email from the instructor or a self-assessment exercise and given a chance to show remediation on them during subsequent examination. The purpose of our study is to determine the effects of this confidence-based, individualized remediation strategy on student learning and final grades. We hypothesize this strategy will increase student learning via remediation, and student learning will be especially increased from self-assessing confidence-based academic performances.

Materials and Methods

Research Design

An experimental research design was implemented for the 2016, 2017, and 2018 student cohorts of University of North Dakota's OT 422 course (Anatomy for Occupational Therapists). The independent variables included the type of feedback (instructor email in 2016 and self-assessment exercise in 2017/2018) and type of remediation questions (individualized in 2016/2017 and standardized for the group in 2018). The dependent variable for all years was student learning (directly tied to LO achievement) and resulting final grades. While effects from the interventions are found by comparing data between cohorts, each intervention had effects on its own cohort and can be examined alone after being factored out of the performance reports and grades post hoc. This allows collection and analysis of two data sets (one with the intervention and one without) from the same group of students without having a separate control group. This study was approved via the procedures of the University of North Dakota (UND) Institutional Review Board (IRB) and ruled as level 4 exempt (IRB-201809-062).

Subjects and Setting

Three cohorts of OT 422 students at the UND School of Medicine and Health Sciences served as subjects. OT 422 is a 5-credit, twelve-week course that meets for a lecture and laboratory five days per week (Monday through Friday) for an approximated 15 contact hours/week. The students are enrolled in OT 422 during the same summer they begin the professional program. Typically, the majority of students are female: 58 of 64 (90.6%) in 2016, 61 of 68 (89.7%) in 2017, and 59 of 65 (90.8%) in 2018. While most of the students begin the OT program directly from their undergraduate career either through UND's early-admission system or by entering with an already-earned bachelor's degree, it is also common to have non-traditional students within the cohorts.

The OT 422 course is delivered through a hybrid of traditional lecturing and active learning. Each class day begins with a PowerPoint lecture presentation (approximately 1 hour in length) covering the anatomical topic of the day followed by a dissection-based, active learning cadaver laboratory (approximately 2 hours in length). While the laboratory component is self-directed, instructors and teaching assistants facilitate learning when needed and assure that progress expectations are being met. Formal assessment of student learning is established through six lecture examinations and six laboratory examinations that determine the majority of the students' final grades. The methods of examination align with the respective teaching methods; lecture examinations consist entirely of multiple choice questions (MCQs) whereas laboratory examinations consist entirely of constructed-response questions (CRQs). Each examination items was strategically tied to a single LO.

Preparation of Learning Objectives and Examination Items

The OT 422 course LOs were developed according to the ABCD method, which specifies that LOs must list the intended Audience (*i.e.,* who the assessment is intended for), a measurable Behavior (*i.e.,* what the learner is to do), any Conditions the learner will encounter (*i.e.,* what the learner will use, have access to, or not be allowed to use), and the Degree at which they are expected to perform (*i.e.*, measurement criteria of acceptable performance).[8–12] Afterward, the 76 LOs were evaluated for SMART criteria in order to be proper for evidence-based assessments of learning. These criteria included Specificity (*i.e.*, they must be specific to a subject area), Measurability (*i.e.,* they must be observable and measurable as guided by the six cognitive levels of Bloom's Taxonomy and their respective action verbs[13–15]), Attainability (*i.e.,* they must be reasonable and not unrealistic), Relevance (*i.e.*, they must be pertinent to the course material and assessment), and Timeliness (*i.e.*, they must be time-bound so that the timeline expectation is clear on when they should be accomplished).[16,17] As a result of evaluating the LOs for SMART criteria, we found that many of the LOs referred to general regions of the body (*e.g.*, upper limb). LOs were revised into more specific "sub"-learning objectives denoted by letters directing students to more specific areas of the body (*e.g.*, pectoral girdle, arm, forearm, or hand) if they were not demonstrating sufficient knowledge in one area over another. This step of specificity brought the total number of LOs for this course from 76 to 125, all of which now properly constructed, and measurable to the degree that evidence-based student performance and learning data can be collected. See Appendix A for the complete list of the LOs used in OT 422.

After establishing the LOs, a clear and purposeful strategy for assessing each LO was designed. While often not undertaken by instructors, this important step guides the development of curricula, creation of assessment items, and choice of content delivery methods.[8,14,18] Determining content proportionality needs begins by establishing the amount and type of assessment the instructor feels is needed to be done in order to sufficiently assess student proficiency on that LO. For example, OT 422's LO27A reads "identify the muscles of the (A) pectoral girdle." In this case, the instructors determined that four assessment items distributed equally between two MCQ lecture examinations and two CRQ laboratory examinations would be able to sufficiently assess student proficiency on that LO. This process results in a meaningful and necessary number of questions needed per LO per assessment. Appendix B illustrates the breakdown of LO proportions and assessment strategy and how it was used to create proportions and an assessment strategy for each of the 125 OT 422 LOs.

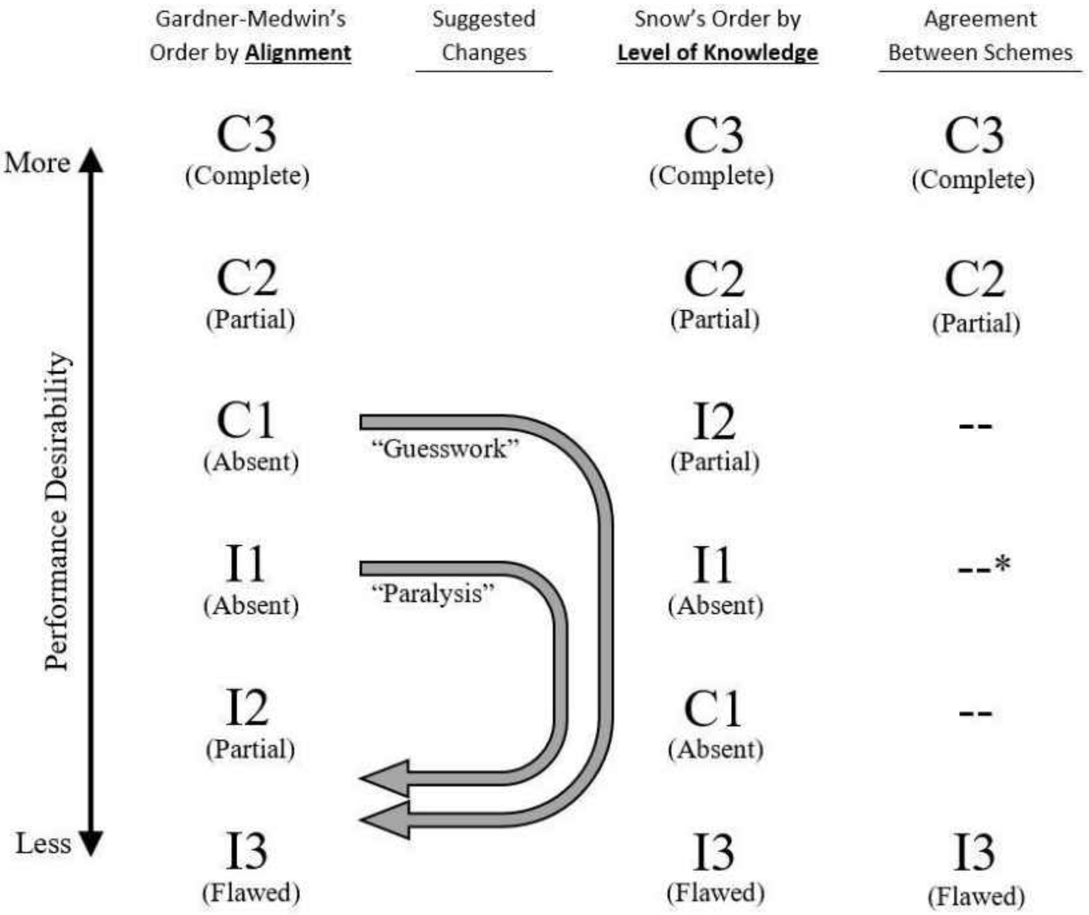
Ultimately, the results of these steps present a clear guide to developing assessment items that are strategically linked and appropriately proportional to each evidence-based LO. These are necessary and crucial steps for establishing purposeful, evidence-based assessments that reflect the course curriculum, instruction delivery, course objectives and course goals.

Performance Rank and Scoring Criteria

In order to accurately detect and measure all levels of knowledge, we employed a confidence-based assessment method to each of the six high-stakes lecture and laboratory examinations. The CBA method employed in this study is one modified from A. R. Gardner-Medwin's CBA scheme.[3] Instead of ordering performances by the alignment of correctness and confidence as he did, we ordered the performances by their interpreted level of knowledge. In doing so, the guesswork (correct (C) and low confidence (1)) and paralysis (incorrect (I) and low confidence (1)) performances indicating absent knowledge are moved below an incorrect performance with medium confidence indicating partial knowledge. Additionally, the guesswork (C1) performance is ordered below the paralysis (I1) performance because a correct answer was undeservedly awarded by chance. See Table II-1 for a comparison between Gardner-Medwin's and Snow's order of confidence-based performances.

Gardner-Medwin's sequence of performances by correctness and confidence alignment justifies his CBA scoring scheme. Performances with higher confidence levels, such as I3 and C3, are assigned higher-weighted (negative and positive) point values for being furthest from and closest to alignment, respectively. Performances with lower confidence levels, such as I1 and C1, are assigned lower-weighted point values. This scoring scheme prevents students from trying to "game the system" and record a confidence level that is inaccurate for the sake of attempting to gain more credit points. In this manner, students will score best if they learn the correct material and are confident in that material. If they do not know the material and/or record an inaccurate confidence level, their score will be negatively impacted.

Table II-1. Performance Rankings of Two Confidence-Based Assessment Schemes. According to Gardner-Medwin's scheme, any correct response is more desirable than an incorrect response, and alignment to confidence levels further ranks performances; higher confidence levels increase performance desirability for correct responses and decrease performance desirability for incorrect responses. According to Snow's scheme, levels of knowledge (complete, partial, absent, and flawed) interpreted by performances are the primary determinant for performance rank, followed secondarily by accuracy of performance alignment to level of knowledge (e.g., "paralysis" ranks over "guesswork" despite both indicating a lack of knowledge). Although significant differences result between the schemes, the most and least desirable performances are the same. Performance combinations (or performance levels, or PLs) are combinations are represented as combinations of correctness and confidence abbreviations. C = correct response; I = incorrect responses; 1 = low confidence; 2 = medium confidence; 3 = high confidence.

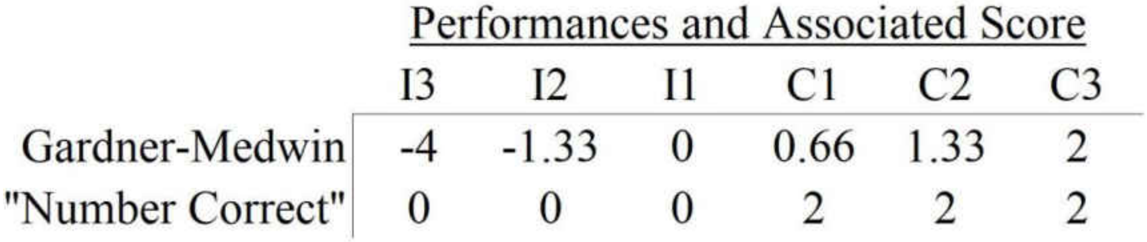


Gardner-Medwin's Order by **Alignment**	Suggested Changes	Snow's Order by **Level of Knowledge**	Agreement Between Schemes
C3 (Complete)		C3 (Complete)	C3 (Complete)
C2 (Partial)		C2 (Partial)	C2 (Partial)
C1 (Absent)	"Guesswork"	I2 (Partial)	--
I1 (Absent)	"Paralysis"	I1 (Absent)	--*
I2 (Partial)		C1 (Absent)	--
I3 (Flawed)		I3 (Flawed)	I3 (Flawed)

*Indirect agreement

As a result of proposing to re-order the performances from Gardner-Medwin's scale for detecting lowest knowledge level performances for remediation purposes, we also propose to adjust the scoring scheme. In theory, all performances representing each level of knowledge *should* be scored the same, regardless of correctness or confidence. However, Gardner-Medwin's scoring criteria rewards students more for correct guesswork over incorrect guesswork despite each performance equally indicating a lack of knowledge (see Table II-2). Nonetheless, if we *were* to award increasingly more points for performances indicating increasingly higher knowledge levels, (*e.g.,* 1 point for I3, 2 points for I1 and C1, 3 points for I2 and C2, and 4 points for C3, as according to Snow's order of CBA PLs), there would be no reason for any student to ever record low confidence (1) on a performance when they could always record medium confidence (2) and receive more credit. In this regard, an increasing point value for increasing knowledge levels will not work. However, Gardner-Medwin's scoring scheme would be inappropriately applied to our performance sequence for similar reasons described above, regardless of having re-ordered the performances. For these reasons, we propose to award full credit for correct answers and no credit for incorrect answers (a traditional scoring method known as "number correct" and commonly used in correctness-only assessments[19–21]) and use the benefits of CBAs to accurately detect poorest performances. This removes the ability for students to try to "game the system" and instead focuses their efforts on answering with correct answers and recording accurate confidence levels to facilitate learning.

Table II-2. Scoring Criteria for Two Confidence-Based Assessment Schemes. In Gardner-Medwin's scoring scheme, performances are awarded more credit points as they increase in rank, beginning with two penalty scores and increasing to full credit. In the "number correct" scoring scheme, credit is awarded based on correctness alone; regardless of respective confidence levels, no credit is earned for incorrect performances and full credit is earned for correct performances. Scores are reported below as being out of 2 credit points total.



	Performances and Associated Score					
	I3	I2	I1	C1	C2	C3
Gardner-Medwin	-4	-1.33	0	0.66	1.33	2
"Number Correct"	0	0	0	2	2	2

While the *number correct* scoring criteria we propose to use awards correct guesswork (*i.e.,* C1) and not incorrect guesswork (*i.e*., I1), despite both indicating a lack of knowledge, we predict that ranking the C1 performance to be second in line for being remediated will have a decreasing "correction" effect on percentage grades as students can positively remediate a "C1" performance with two incorrect performances (I1 and I2).

Routine Examination Procedure

For each cohort, six lecture and six laboratory examinations were administered in pairs as routine course assessments every 8 class days (approximately). Students were given a paper answer sheet with numbered answer lines to record their answers to the corresponding questions. An additional line next to each answer line was given as a place to write their level of confidence in their answer. The answer sheet directions read:

> *"For each question, please record your answer choice(s) to each question in the first blank and your level of confidence in that answer in the second blank. Record your level of confidence in each answer you choose using the following scale:*
>
> *1 = not confident; 2 = somewhat confident; 3 = very confident."*

Students were required to acknowledge their honesty and accuracy in their recorded confidence levels by signing a statement to that effect at the end of their answer sheet. Answer sheets were collected and scored against a pre-established key according to correctness-only confidence-based scoring criteria (see Table II-2), awarding up to two credit points per question. During the subsequent class day, students were returned their scored answer sheets for review and reflection of their performances. They were given a copy of the examination for lecture, and for laboratory they were shown the examination key that included the question, the structure tagged on the cadaver associated with the question, and the correct answer. During examination review, the 2017 and 2018 cohorts completed a self-assessment exercise (SAE) to identify the six LOs which they believed they performed poorest on in lecture and in laboratory. The answer sheets for all cohorts (with the SAEs in 2017 and 2018) were again collected, organized, and prepared for instructor analysis for the remediation intervention.

Intervention Procedure

For each LO tested in each examination, one additional, similar question testing the same LO was created as a remediation question (RQ) and archived in a Microsoft Word document. Laboratory remediation questions remained constructed-response-type questions and were created in Microsoft Word using tagged structures on anatomical images instead of real cadaver specimens). Most LOs were tested in both lecture and laboratory settings and therefore required the creation of two different-type RQs. Additionally, LOs that were tested in one examination and again in a later examination required the preparation of an additional RQ. Having a complete RQ question bank created clear organization during the remediation process and efficiency for creating the remediation portions of examinations. The same RQ question bank was used for all three cohorts.

For the 2016 and 2017 cohorts, all confidence-based examination question performances were analyzed separately for each student for the six LOs on which he/she performed poorest according to Snow's ordering of CBA PLs. These six LOs dictated which six individualized remediation questions (IRQs) each student was to be given. Each student's six IRQs were then transferred to their own Microsoft Word document, printed and administered to each student as an additional component of the subsequent examination. This remediation analysis process was completed twice (once for lecture examinations and once for laboratory examinations) between each set of exams creating a total of 12 IRQs per student per set of examinations and 72 total remediation questions per student throughout the entire course. It was determined to administer six IRQs per examination as this equates to approximately 20% of the approximately 31 LOs tested in each of four curriculum units (see Appendices A and B). It was determined that this amount would be both practicable and impactful without creating unmanageable work for the instructors or students.

While many questions on each examination could test the same LO, special attention was given to choosing six *different* LOs for all three cohorts; if one LO was associated with more than one of the six poorest performances, it was only counted as one LO and the next poorest performance's associated LO was added to the list. This was done to intentionally focus remediation efforts on showing mastery of LOs and tie more meaning to performance improvement. The pre-established LO rankings (see Appendix B) were used to "break ties" between multiple LOs associated with the last performance level when less were needed to reach six total LOs. This was also necessary for managing the number of RQs needed to have prepared for both lecture and laboratory reexamination. The LO rankings were given to the students in 2017 and 2018 to facilitate their self-assessment exercises (SAEs).

To test the effectiveness of an individualized remediation strategy as opposed to a standardized group strategy, standardized remediation questions (SRQs) were administered to the 2018 cohort instead of IRQs. The SRQ LOs were determined and created by the same methodologies as those used to determine IRQs for the years prior *except* that the 2016 and 2017 cohorts' most common IRQ pre-remediation (pre-R) performances were the performances used to select which SRQs would be administered.

Lastly, just as they were collected on the regular examination questions, confidence levels were collected on the remediation performances to determine the extent of remediation between pre-R performances and post-remediation (post-R) performances. Each examination answer sheet (except the first) included answer lines and confidence lines for the RQs. Additionally, as routine procedure an end-of-semester course survey was administered to the students to collect their perceptions of the course and their learning. Customized questions were included in this survey to collect their perceptions on the interventions and its effects.

Data Collection and Analysis

All performance data was transferred to Microsoft Excel for analysis. Student t-tests were used to compare data sets. The following variables were tested:

a) Mean changes in student confidence and correctness from IRQ intervention (2016 and 2017 cohorts) were calculated separately from pre-R and post-R PLs.

b) Individual and class mean remediation amounts were calculated from comparing pre-R and post-R PLs for all three years to determine the effect of IRQs and SRQs on student learning. In treating Snow's order of performances as an ordinal scale of equally-achievable performances, each performance was represented by a numerical value of 1 through 6 in increasing order of performance desirability. Successful remediation was determined to result from individual or mean post-R PLs ≥ 4.50. Class mean remediation scores were compared between cohorts to determine the mean remediation effects of individualized vs. standardized remediation interventions and instructor email vs. SAE feedback methods.

c) Remediation per individual and all LOs were calculated to determine and compare the interventions' effects on LO achievement. LO achievement was determined to result from individual or mean post-R PLs ≥ 4.50.

d) Correlations between pre-R PLs and post-R PLs, post-R PLs and grades, pre-R PLs and overall remediation, and overall remediation and grades were calculated using Pearson correlation tests.

e) Alignment between student and instructor perceptions of poorest performance was examined by comparing instructor-chosen LOs for remediation to student-selected LOs collected from their SAEs. Student responses from an end-of-semester course survey about how the interventions impacted their learning were also examined.

f) IRQ and SRQ impact on student grades was determined by comparing student grades, including after factoring out the intervention effects. Additional course survey data was used to compare how students thought the intervention affected their grades compared to how the intervention actually affect their grades.

Results

Effects on Confidence and Correctness (Separated)

Figure II-1 shows that both individualized and standardized remediation interventions on poor performances has a positive effect on raising student confidence levels; in all three years, post-R confidence levels were significantly higher than pre-R confidence levels. Although this may expectedly be a positive result, mean pre-R confidence levels are quite high – approaching a level of "medium confidence". This is due to the fact that the poorest performance, according to Snow's ordering of CBA PLs, is one that is incorrect but displays the highest level of confidence. From this, it is important to remember that although higher levels of confidence are generally desirable, a positive remediation can result from remediating a high-confidence performance with a low-confidence performance. For example, performances indicating flawed knowledge (*i.e.*, I3) can be positively remediated with performances indicating absent knowledge (*i.e.*, C1 or I1) or partial knowledge (*i.e.,* C2 or I2), all of which have lower confidence levels than demonstrated in the I3 performance. If performances were ranked on confidence alone, we would expect confidence changes between pre-R and post R performances to be much greater than shown in the data, but this would be at the cost of rewarding students for flawed knowledge.

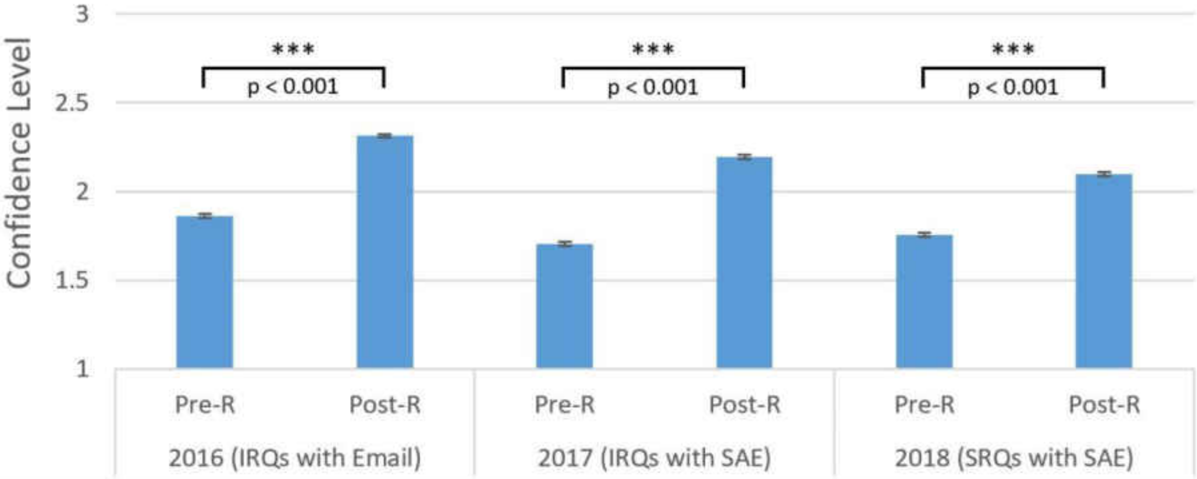


Figure II-1. Changes in Student Confidence from Remediation Strategy. Students in all three cohorts displayed higher confidence in answers to their remediation questions than they did in their pre-remediation (pre-R) performances (standard error also reported). Although pre-R and post-remediation (post-R) confidence levels appeared to change significantly in each IRQ and SRQ remediation strategy, changes appeared to be greater with IRQs. Confidence levels were reported as 1 = low confidence, 2 – medium confidence, and 3 = high confidence. Confidence levels associated with responses that earned partial credit were factored out of this analysis for reasons described earlier.

Despite I3 being considered the least desirable performance, mean pre-R confidence levels are not *that* high. I3 occurred less frequently as a pre-R PL than C1 or I1 (see Figure II-5), the next two poorest performances contributing most to the low mean pre-R confidence levels. This indicates that students do not know the material (and acknowledge so with a low confidence level) slightly more so than believing they know the material when in fact they have learned content incorrectly. From C1 or I1 pre-R performances, remediating with higher confidence (except in the case of I3) is desirable as it would indicate partial or complete knowledge. By this rationale, we can conclude that higher post-R confidence levels alone suggests positive effects on learning.

Alternatively, correctness – the counterpart to confidence necessary to properly measure knowledge – can also offer some value when examined alone. Generally speaking, instructors wish for students to perform as correctly as they can. Figure II-2 shows the effects of IRQ and SRQ remediation strategies on correctness alone. It shows that students are generally remediating from a lower-correctness level to a higher-correctness level with either strategy. While this is indeed a positive result and perhaps be expected, it is important to remember that a positive remediation does not necessarily mean moving from an incorrect performance to a correct performance; in fact, a student can show positive remediation by remediating a poor correct performance, such as C1 (*i.e.,* correct guesswork), with a better but incorrect performance, such as I2 (*i.e.*, partial knowledge).

The more common of the two worst pre-remediation performances was C1. If we were to primarily order the performances based on correctness alone (as done in Gardner-Medwin's order of CBA PLs) we would expect to see much larger differences between pre-R and post-R correctness levels with pre-R levels being much more incorrect and post-R levels being much more correct. This would be at the cost of rewarding students for guesswork (C1) over I1 and I2 performances. Despite ranking a correct answer as second-to-lowest in our order of performances, the present study shows that students are achieving higher levels of correctness upon remediation.

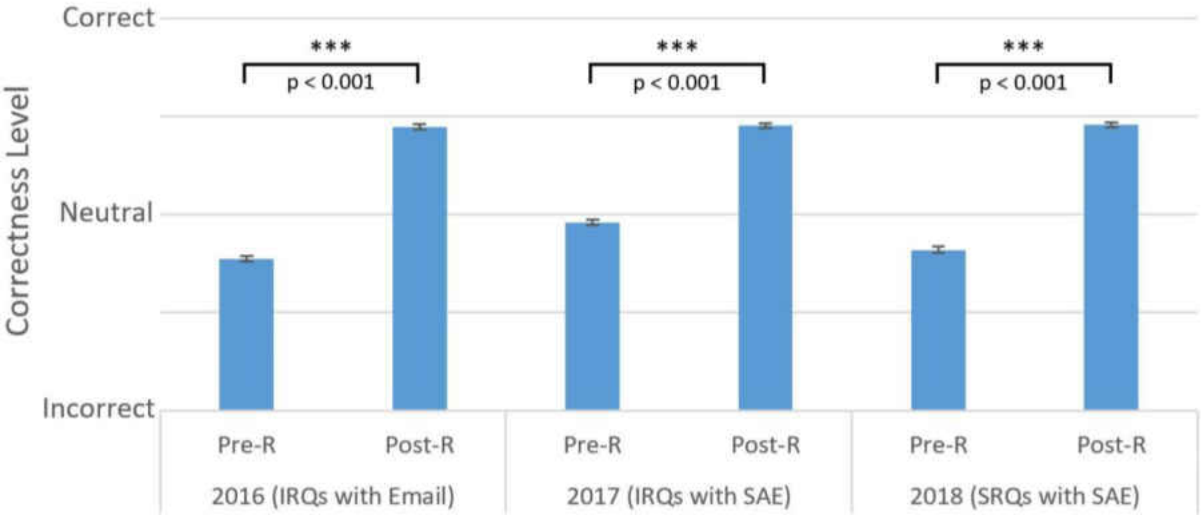


Figure II-2. Changes in Student Correctness from Remediation Strategy. Students in all three cohorts displayed higher correctness in answers to their remediation questions than they did in their pre-remediation (pre-R) performances (standard error also reported). Regardless of associated confidence, correctness of answers was scored as incorrect earning no credit (0/2 points), and correct earning full credit (2/2 points).

Despite the fact that our re-ordering of CBA PLs to reflect knowledge levels is not dictated by confidence or correctness alone, Figures II-1 and II-2 show how using our order of performances for remediation generates significantly higher post-R confidence and correctness levels. While the confidence-only and correctness-only analyses can each separately offer valuable information for assessing knowledge, knowledge is, by definition, multidimensional. Regardless of the value confidence and correctness can each offer to assessing knowledge, the most accurate assessment of knowledge can only be achieved through evaluating one in direct relation to the other.

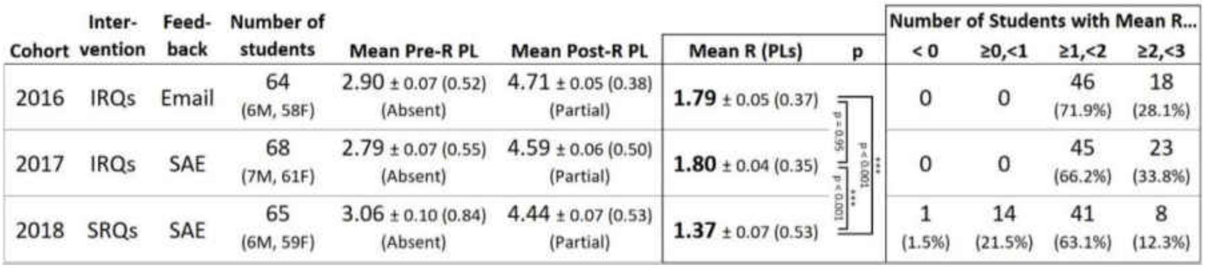
Effects on Student Learning

Three cohorts of OT 422 students were examined for effects of confidence-based individualized and standardized remediation strategies. Using confidence-based performances ranked by knowledge level (see Snow's CBA PL rank in Table II-1), the 2016 and 2017 cohorts' poorest performances were individually retested for six LOs for each of six high-stakes lecture and laboratory examinations. As a result, students showed a mean remediation of knowledge by nearly two PLs (see Table II-3). Because each level of knowledge consists of two or less performance levels, an increase in performance by two PLs means a mean remediation of at least one knowledge level was achieved. More specifically, the mean pre-R PL was closest to I1, indicating a lack of knowledge, and the mean Post-R PL was closest to C2, indicating a level of partial knowledge. These mean performance changes support the previously discussed data showing how, on average, confidence and correctness each increased separately as a result of the remediation strategy. Additionally, no students displayed a negative effect from the remediation strategy. In fact, every student in 2016 and 2017 showed a mean remediation greater than or equal to one PL as a result of the IRQ intervention.

The only independent variable between the 2016 and 2017 cohorts was the type of performance feedback provided to the students. In 2016, an individualized email from the instructor was sent to each student after every examination. It included the six LOs on which they would be retested on as well as their initial performances that resulted in those LOs most needing to be retested. In 2017, students completed an in-class self-assessment exercise (SAE) in place of receiving an instructor-derived email. This change was implemented with the expectation that it would cause students to be more mindful of their performances and thus aide their remediation efforts.

Each year, students were instructed about the concept of CBA PLs so they could use that understanding to direct their remediation efforts. After all, remediating each CBA performance requires a different approach. In 2017, both mean pre-R and post-R PLs were slightly lower than those in 2016. Because pre-R PLs are determined from regular unit examination questions, they are independent of any feedback intervention. However, lower post-R PLs in 2017 suggesting that SAEs did not have the effect we thought it would on student learning. Ultimately, these data suggests that the effectiveness of the IRQ strategy was independent of these two feedback variables for student learning.

Table II-3. Mean Student Remediation. The 2016 and 2017 individualized remediation question (IRQ) interventions resulted in similar mean student remediation (R) of nearly two performance levels (out of the 6 performance levels listed in Table II-1). No difference was found to result from email or self-assessment exercise (SAE) feedback. Students achieved significantly less remediation from standardized remediation questions (SRQs) in 2018 than from IRQs in either cohort. The majority of students in each cohort earned a mean remediation score between 1 and 2 performance levels. Individual student remediation is shown later in Figure II-4 and discussed in association with relative grade percentages. Means are presented as "MEAN ± SEM (StDev)."



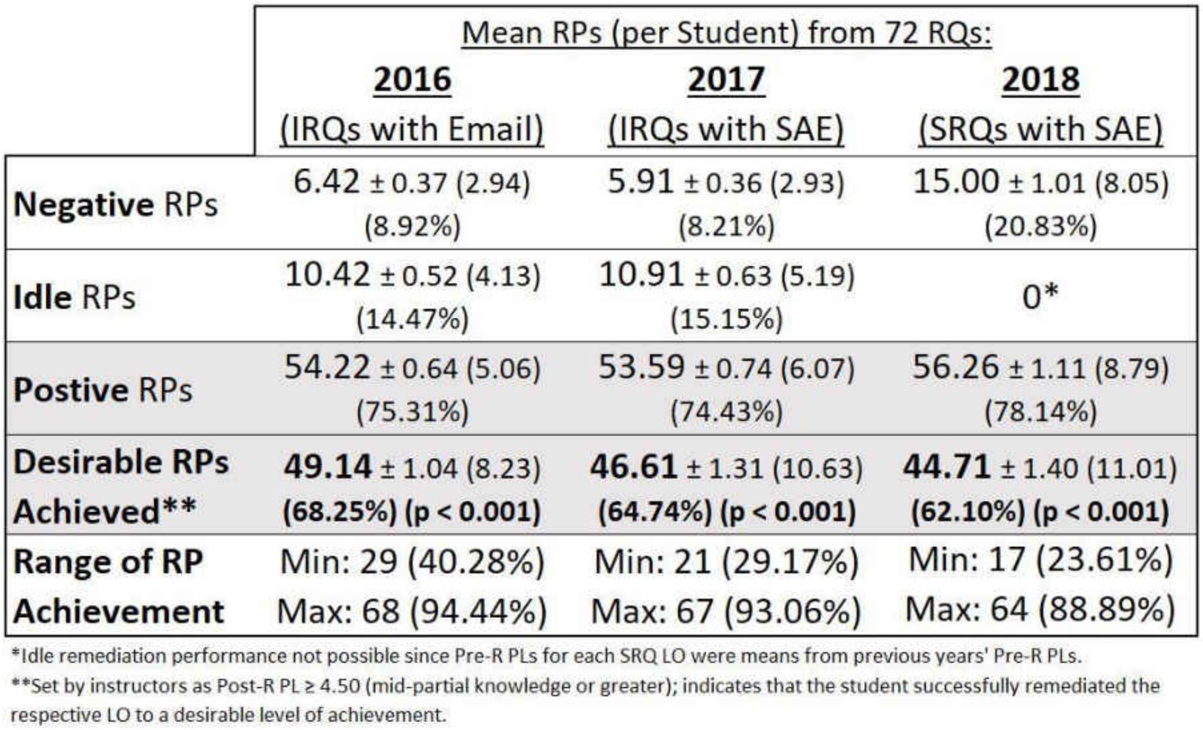
	Inter-	Feed-	Number of					Number of Students with Mean R...			
Cohort	vention	back	students	Mean Pre-R PL	Mean Post-R PL	Mean R (PLs)	p	< 0	≥0,<1	≥1,<2	≥2,<3
2016	IRQs	Email	64 (6M, 58F)	**2.90** ± 0.07 (0.52) (Absent)	**4.71** ± 0.05 (0.38) (Partial)	**1.79** ± 0.05 (0.37)	p = 0.95 p < 0.001 ***	0	0	46 (71.9%)	18 (28.1%)
2017	IRQs	SAE	68 (7M, 61F)	**2.79** ± 0.07 (0.55) (Absent)	**4.59** ± 0.06 (0.50) (Partial)	**1.80** ± 0.04 (0.35)	p < 0.001 ***	0	0	45 (66.2%)	23 (33.8%)
2018	SRQs	SAE	65 (6M, 59F)	**3.06** ± 0.10 (0.84) (Absent)	**4.44** ± 0.07 (0.53) (Partial)	**1.37** ± 0.07 (0.53)		1 (1.5%)	14 (21.5%)	41 (63.1%)	8 (12.3%)

To test the effectiveness of an individualized remediation intervention, a standardized remediation method was implemented for the 2018 cohort. The SRQ intervention followed all of the previously-used IRQ principles and methods except that all students' received the same SRQs; while 2016 and 2017 students were retested on different LOs, 2018 students were all retested on the same LOs. SRQ LOs were determined by the six most commonly retested LOs for each 2016 and 2017 lecture and laboratory examination. Additionally, pre-R PLs varied student to student and even performance to performance for each LO in 2016 and 2017. In order to evaluate individual student remediation in 2018 from LOs chosen from previous students' individual pre-R performances, the mean of all of the pre-R performances for each separate SRQ LO was established as a single pre-set pre-R performance level for each of the 2018 students.

Like IRQs, SRQs were determined to have a significant effect on their respective cohort alone, but they were determined to have less of an effect on student remediation than the 2016 and 2017 IRQs ($p<0.001$). Mean SRQ pre-R PLs were higher than those of previous years. This is because they were determined by previous years' pre-R performance data for the most commonly retested LOs as previously described, and not all 2016 and 2017 students' poorest performances contributed to the mean 2018 pre-R PLs. Relatedly, mean SRQ post-R PLs were lower than those of previous years. This is likely because not all students were retested on the LOs on which they most needed to be retested. As a result, retesting LOs about which students were already knowledgeable gave them less range for possible remediation, and retesting LOs that students weren't expecting (*i.e.*, did not come to their attention in their SAEs) gave them less of a chance to raise their individual performance level for having not prepared sufficiently.

Table II-4 shows the direction frequency of mean remediation for the 72 individual student remediation performances (RPs). Each of the three cohorts answered most of their 72 total remediation questions (RQs) with a higher post-R PL than corresponding pre-R PL, resulting in majority of performances showing positive remediation. Still, each student in 2016 and 2017 answered approximately 10 of those 72 questions with the same post-R PL as their respective pre-R PLs, showing no remediation progress (*i.e.*, idle remediation) for the corresponding LOs. In 2018, idle remediation was not possible for students to achieve since pre-R PLs for each SRQ LO were pre-determined from previous years' pre-R PLs; consequently, their would-be idle RPs resulted in either positive or negative RPs. This explains the higher mean positive and negative RPs that resulted in 2018. Additionally, in comparing the negative remediation results across all three cohorts in Table II-4, it appears the majority of the 2018 cohort's would-be idle RPs became negative RPs. This likely explains why SRQs led to lesser classwide mean remediation than IRQs.

Table II-4. Direction of Student Remediation and Resulting Achievement. The majority of the 72 student Remediation Performances (RPs) were positive for all three cohorts. Little difference in direction of RPs was found between the 2016 and 2017 cohorts. The 2018 cohort's would-be idle RPs seemed to mostly become negative RPs.This provides one likely explanation for the lesser remediation in 2018 shown in Table II-3. Means are presented as "MEAN ± SEM (StDev)."



	Mean RPs (per Student) from 72 RQs:		
	2016 (IRQs with Email)	**2017** (IRQs with SAE)	**2018** (SRQs with SAE)
Negative RPs	6.42 ± 0.37 (2.94) (8.92%)	5.91 ± 0.36 (2.93) (8.21%)	15.00 ± 1.01 (8.05) (20.83%)
Idle RPs	10.42 ± 0.52 (4.13) (14.47%)	10.91 ± 0.63 (5.19) (15.15%)	0*
Postive RPs	54.22 ± 0.64 (5.06) (75.31%)	53.59 ± 0.74 (6.07) (74.43%)	56.26 ± 1.11 (8.79) (78.14%)
Desirable RPs Achieved**	**49.14** ± 1.04 (8.23) **(68.25%) (p < 0.001)**	**46.61** ± 1.31 (10.63) **(64.74%) (p < 0.001)**	**44.71** ± 1.40 (11.01) **(62.10%) (p < 0.001)**
Range of RP Achievement	Min: 29 (40.28%) Max: 68 (94.44%)	Min: 21 (29.17%) Max: 67 (93.06%)	Min: 17 (23.61%) Max: 64 (88.89%)

*Idle remediation performance not possible since Pre-R PLs for each SRQ LO were means from previous years' Pre-R PLs.
**Set by instructors as Post-R PL ≥ 4.50 (mid-partial knowledge or greater); indicates that the student successfully remediated the respective LO to a desirable level of achievement.

While any positive RP is desirable, a positive RP does not necessarily imply a satisfactory PL was reached or the respective LO was achieved. In other words, a positive RP can happen anywhere along the scale of PLs, but achievement can only be determined based on the respective PL. This principle is critical for better understanding the student remediation data in Table II-3. The instructors decided a Post-R PL of 4.50 would be the minimum parameter for determining achievement as a mid-partial knowledge level (*i.e.,* a

4.50 PL) or greater would indicate a student had demonstrated adequate knowledge in the respective LO. As Table II-4 shows, students generally but not always reached LO achievement from positive RPs, and some students achieved many more and many less LOs than the classwide mean. Regardless, while differences in RP directionality and student remediation scores were evident in 2018, the number of desirable RPs achieved per student from positive remediation were found to be comparable between all three years, suggesting once again that the feedback interventions did not significantly influence student learning.

While classwide effects of the intervention is useful for informing big picture decision-making, looking at overall individual student performances is useful for academic advising during the learning process. Remediation performances for individual students can highlight general performance behaviors and identify unawareness of one's own knowledge level. Beyond its usefulness for advising, individual performance data is necessary for understanding how the remediation strategies impacted each student. For example, Figure II-3 shows an individualized remediation performance report (RPR) for four select students.

Figure II-3 shows individual student RPRs for four select students from the 2016 cohort: two with high (Figure II-3A) and low (Figure II-3B) remediation scores and two with high (Figure II-3C) and low (Figure II-3D) percentage grades. Principally, the lower the pre-R PL, the greater the range for possible remediation. This explains why students with higher remediation scores typically have RPRs that look like those of Student 50 in Figure II-3A; nearly all of his/her pre-R PLs are I3, and he or she remediated most of those to a post-R PL of C3. The student representing a low percentage grade (Figure II-3D) demonstrated very similar pre-R performances except he/she did not remediate as many of those I3 pre-R PLs to C3, causing him or her to have a slightly lower (but still relatively high) remediation score.

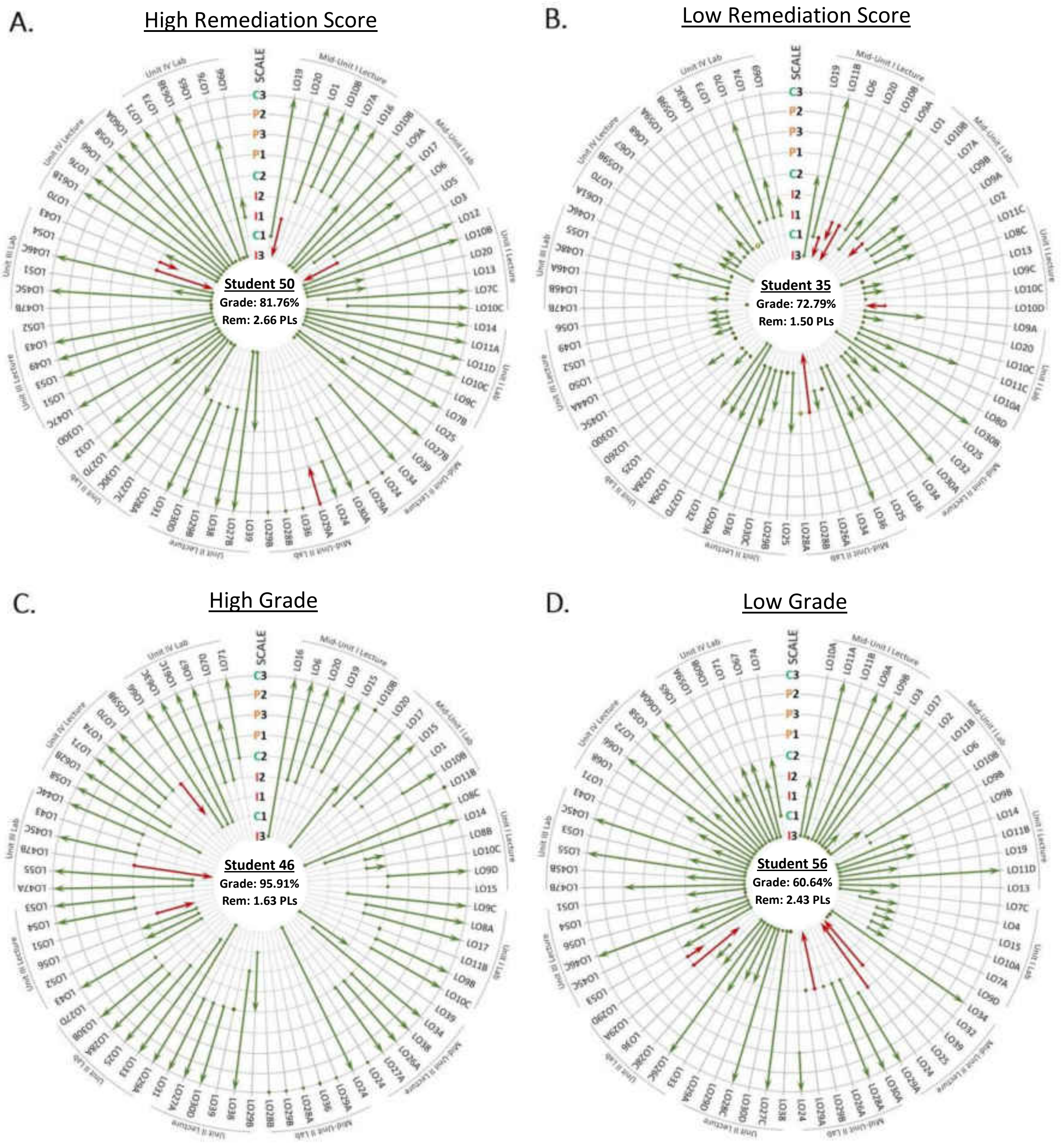
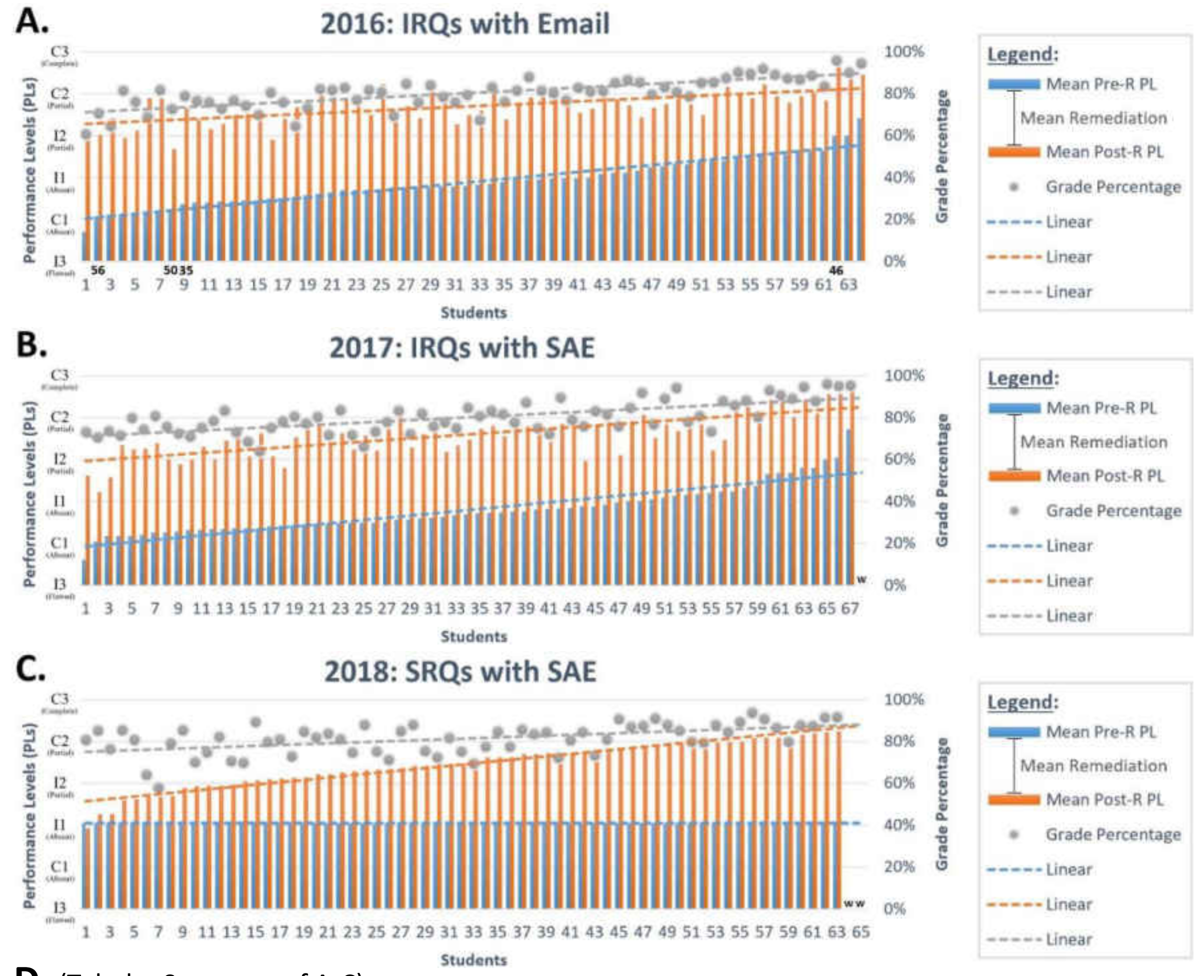


Figure II-3. Individual Student Remediation Performance Reports (RPRs). Individual RPRs are presented for four select students from the 2016 cohort who represent general high (A) and low (B) remediation scores and high (C) and low (D) grade percentages. Each student's 72 IRQ LOs are listed around the edge of their RPR, along with an indicator for which examination they pertained to. The scale of PLs is listed down the top center of each RPR. The "P" PLs presented in these RPRs, indicating incorrect responses that merited partial credit, have been factored out of the remediation scores presented. Each student's remediation progress for each LO is represented by the length of a green arrow (positive remediation) or red arrow (negative remediation), with the start of the arrow at the respective pre-R PL and the tip of the arrow at the respective post-R PL. Single dots indicate idle remediation where the post-R PL was the same as the pre-R PL.

Alternatively, low remediation scores can be caused by two factors: 1) little to no remediation was achieved or 2) little to no remediation was *possible* to be achieved. Student 35 (Figure II-3-B) did not achieve a high remediation score despite the significant possible range, but Student 46 (Figure II-3C) earned a low remediation score because his/her pre-R performances were already high, mostly C2, leaving only one PL possible for maximum remediation. This observation suggests a very important point – that low remediation scores are *not necessarily* undesirable if they are only low because the respective students are performing well to begin with (like Student 46); low remediation scores are only undesirable if they are caused by poor post-R performances given a greater range for remediation (as seen with Student 35). These considerations are why individual student RPRs, such as those in Figure II-3, can be valuable to understanding student performances.

Furthermore, as undesirable as lower remediation scores can be, they are not detrimental unless they become negative. *Any positive* remediation score, *even if it is low*, is at least *somewhat* desirable as it means the student is moving their knowledge level for the respective LOs in a desirable direction. Alternatively, high remediation scores are *always* desirable because they mean students are achieving significantly more and better knowledge.

While Figure II-3 displays the importance of individual student RPRs, Figure II-4 displays the classwide correlations between individual remediation scores and increasing grade percentages for each cohort. Results from the 2016 intervention (Figure II-4A) show a correlation between lower grades and higher individual remediation scores. This correlation is supported by what was seen in the RPRs of Figure II-3C and D and is likely due to the principles described when discussing Figure II-3. The four students referenced in Figure II-3 are identified in Figure II-4A for comparisons to their classmates.



D. (Tabular Summary of A-C)

		Correlations			
Condition 1	**Condition 2**	**2016**	**2017**	**2018**	**Explanation**
As mean Pre-R PLs increase,	mean Post-R PLs increase.	r = 0.69 **(strong)** p < 0.001	r = 0.77 **(strong)** p < 0.001	Not Applicable	Students who perform better overall (i.e., on all the original exam questions) should continue to do so upon remediaiton.
As mean Post-R PLs increase,	grades increase.	r = 0.62 **(strong)** p < 0.001	r = 0.73 **(strong)** p < 0.001	r = 0.46 **(medium)** p < 0.001	Better remediation of poor performances should result in higher grades since higher PLs are more associated with correct answers.
As mean Pre-R PLs increase,	mean R decreases.	r = 0.70 **(strong)** p < 0.001	r = 0.45 **(medium)** p < 0.001	Not Applicable	Higher mean Pre-R PLs make less possible range for remediation.
As mean R decreases,	grades increase.	r = 0.49 **(medium)** p < 0.001	r = 0.12 (small) p = 0.338	r = (0.46) **(medium)** p < 0.001	Varies significantly depending on intervention.

Figure II-4. Remediation Performances vs. Grade Percentages. In 2016 (A), students with higher percentage grades achieved less mean remediation of IRQs (r = 0.49, p < 0.001). Although much less obvious, a similar but lessened correlation was seen in 2017 (B) due to the introduction of SAEs (r = 0.12, p < 0.001). Oppositely, students in 2018 (C) tended to achieve greater remediation with increasing percentage grades due to the standardized nature of the remediation questions (r = -0.46, p < 0.001). A tabular summary of A-C (D) explains why other significant correlations are also seen. The four 2016 students represented in Figure II-3 are identified in (A).

The intervention in 2017 (Figure II-4B) produced a very slight but similar correlation; the only difference was the implementation of self-assessment exercises in place of instructor emails for performance feedback. Although the different feedback mechanisms were previously found to have no direct significant effect on students' mean remediation (see Tables II-3 and II-4), the comparison of A and B in Figure II-4 suggests that they *could* contribute to remediation performance differences for students *depending on the students overall percentage grade*. If this is true, the individualized email seemed to help more students with lower percentage grades achieve greater remediation in 2016, and the SAEs seemed to help more students with higher percentage grades achieve greater remediation in 2017. Regardless, it's important to note that these correlations contain highly variable performances, suggesting the importance of examining individual student RPRs over classwide correlation.

Lastly, the standardized group intervention in 2018 (Figure II-4C) produced a correlation of higher remediation scores to higher percentage grades and lower remediation scores to lower percentage grades, opposite to the effect of the individualized interventions, particularly for low-scoring students. Because 2018 pre-R PLs were pre-determined from 2016 and 2017 performances for all students to best represent a standardized evaluation of performance, they are identical for each student and thus visually produce a flat line across all 2018 students' performances. When compared to A and B in Figure II-4, Figure II-4C suggests that IRQs and SRQs have similar effect on high-scoring students, but low-scoring students are much more positively impacted by IRQs. Remediation scores were most variable in 2018, even including one negative mean remediation score and fourteen remediation scores below 1.0 PL (Table II-3).

Despite the decreased effect the standardized remediation intervention had on student learning, it served its purpose in providing a positive remediation strategy for students and demonstrating that an individualized approach yields better student learning outcomes. These results also suggest support our suggested ranking of confidence-based performances by knowledge level for detecting and remediating poor LO performances.

Effects on Learning Objective Achievement

While intervention effects on student learning were found to be impactful, an equally important aspect of these interventions is their influence on classwide LO achievement. A total of 72 pre-R and 72 post-R performances were examined for each student throughout the course. Figure II-5 displays the mean pre-R and post-R PL frequencies per LO for 2016 (A), 2017 (B), and 2018 (C). As shown, significant differences between pre-R and post-R frequencies for the less desirable and more desirable PLs, but generally not mid-range ones, was found to result. This alone suggests a significant shift in performance levels but requires examining the PL pre-R and post-R frequencies to determine the direction of the shift. Because less desirable PLs show high pre-R frequencies and low post-R frequencies, and more desirable PLs show low pre-R frequencies and high post-R frequencies, we can conclude that LO achievement was shifted in the positive direction.

The result is highlighted by the negatively-sloped pre-R and positively-sloped post-R best-fit linear trend lines included in each of A, B, and C graphs of Figure II-5. Higher pre-R frequencies were expected for poorer PLs since they were selected as pre-R PLs demonstrating poor performances. Therefore, only post-R performances determine the degree and direction (positive or negative) of remediation for all LOs. Individual LO pre-R and post-R PL frequencies can be viewed in Appendix C.

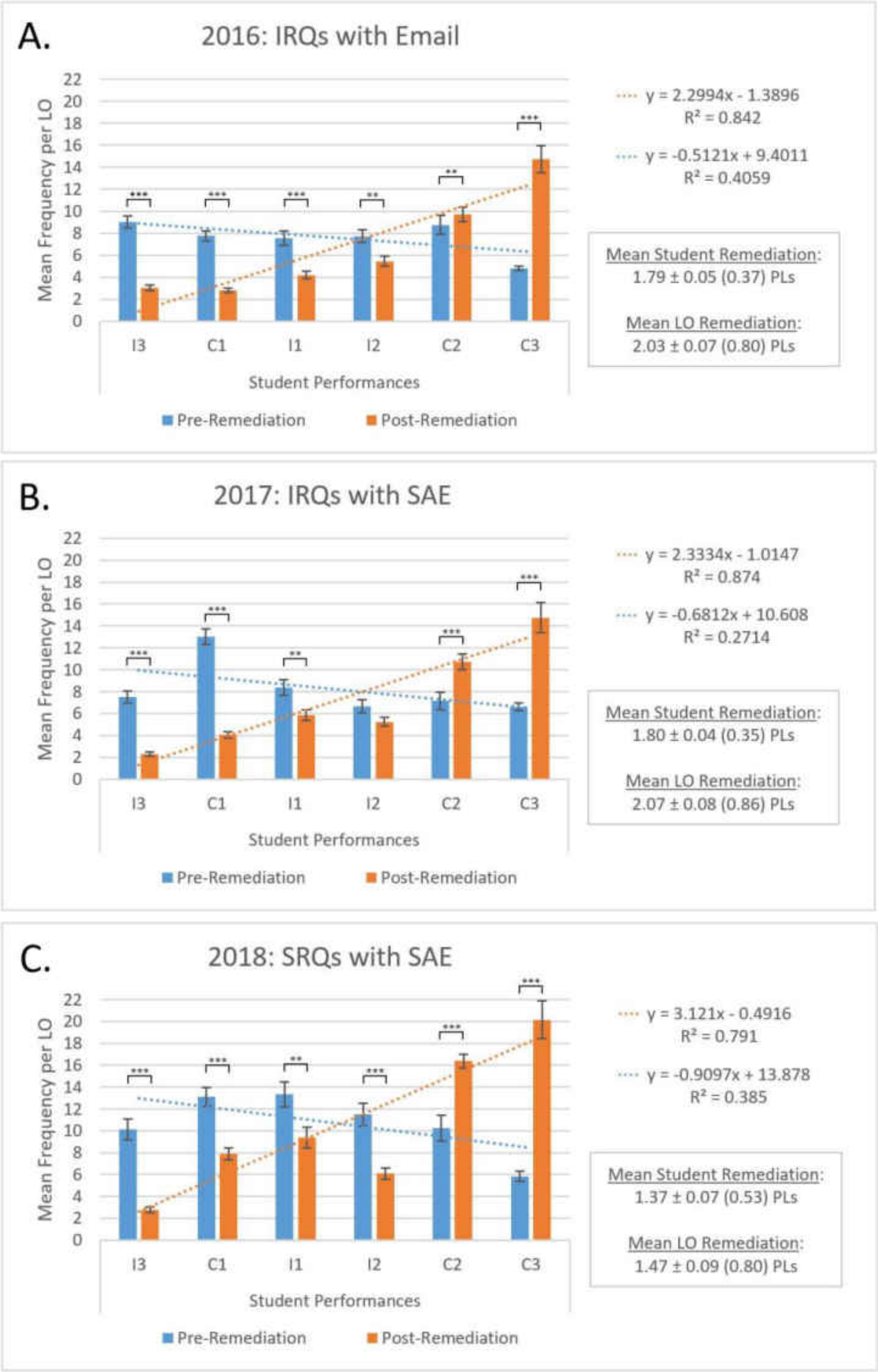
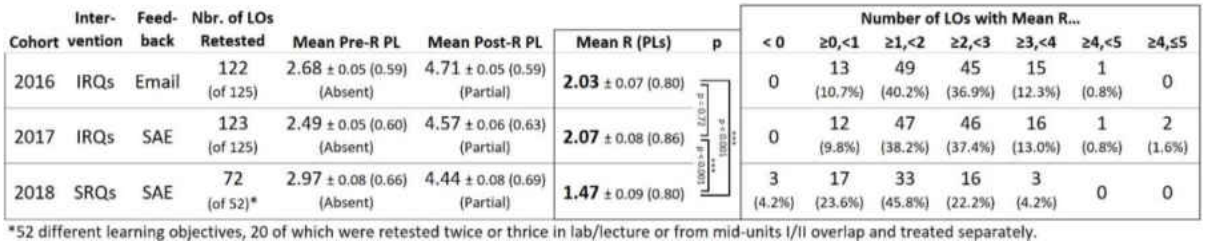


Figure II-5. Mean LO Pre-R and Post-R PL Frequencies. The 2016 (A), 2017 (B), and 2018 (C) remediation strategies resulted in a significant and positive impact on LO achievement as indicated by high pre-R frequencies and low post-R frequencies for less desirable PLs, and low pre-R frequencies and high post-R frequencies for more desirable PLs. Standard error is also presented. Individual LO pre-R and post-R PL frequencies can be viewed in Appendix C.

Most LOs (over 97% of 125 total) were retested in 2016 and 2017 due to the individualized nature of the remediation strategy (Table II-5). This resulted in comparable 2016 and 2017 mean LO remediation scores of 2.03 ± 0.80 PLs per LO and 2.07 ± 0.86 PLs per LO, respectively.

Table II-5. Mean LO Remediation. Nearly every LO was retested with the individualized remediation strategy in 2016 and 2017, but under half of them were retested with the standardized remediation strategy in 2018. While the mean LO remediation amount in 2018 was significantly less than that for 2016 or 2017, resulting in a negative shift in mean remediation range frequencies. Regardless, all years' interventions produced a mean remediation of one knowledge level (from absent knowledge to partial knowledge). Means are presented as "MEAN ± SEM (StDev)." Individual LO mean remediation scores and resulting achievement can be viewed in Appendix C.



Cohort	Inter-vention	Feed-back	Nbr. of LOs Retested	Mean Pre-R PL	Mean Post-R PL	Mean R (PLs)	p	Number of LOs with Mean R…						
								< 0	≥0,<1	≥1,<2	≥2,<3	≥3,<4	≥4,<5	≥4,≤5
2016	IRQs	Email	122 (of 125)	2.68 ± 0.05 (0.59) (Absent)	4.71 ± 0.05 (0.59) (Partial)	**2.03** ± 0.07 (0.80)	p = 0.72	0	13 (10.7%)	49 (40.2%)	45 (36.9%)	15 (12.3%)	1 (0.8%)	0
2017	IRQs	SAE	123 (of 125)	2.49 ± 0.05 (0.60) (Absent)	4.57 ± 0.06 (0.63) (Partial)	**2.07** ± 0.08 (0.86)	p < 0.001 ***	0	12 (9.8%)	47 (38.2%)	46 (37.4%)	16 (13.0%)	1 (0.8%)	2 (1.6%)
2018	SRQs	SAE	72 (of 52)*	2.97 ± 0.08 (0.66) (Absent)	4.44 ± 0.08 (0.69) (Partial)	**1.47** ± 0.09 (0.80)	p < 0.001 ***	3 (4.2%)	17 (23.6%)	33 (45.8%)	16 (22.2%)	3 (4.2%)	0	0

*52 different learning objectives, 20 of which were retested twice or thrice in lab/lecture or from mid-units I/II overlap and treated separately.

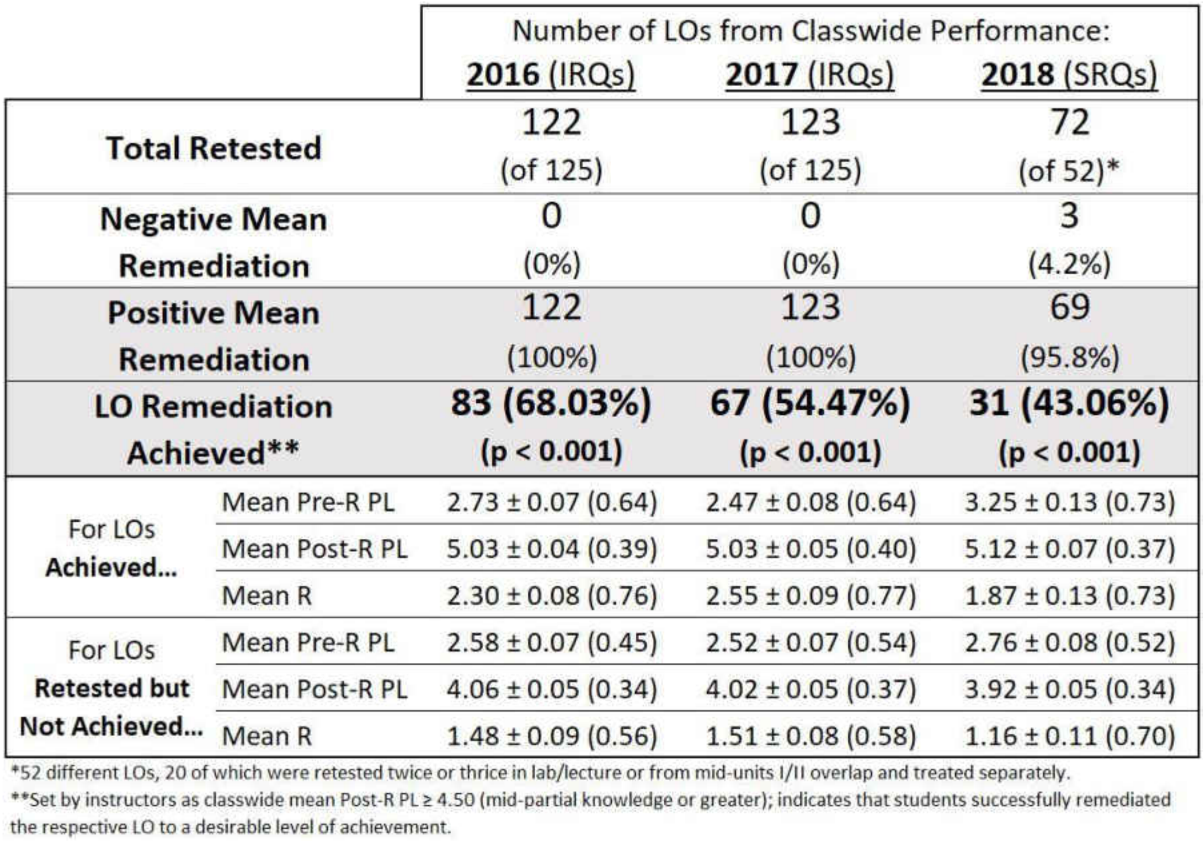
Alternatively, only 52 [nearly 42% of] LOs were retested in 2018 due to the standardized nature of the remediation strategy, 20 of which were retested twice or thrice by being treated as separate LOs between lecture and laboratory or between mid-units I and II overlap with unit exams. It was determined to treat these LOs separately since they were being tested through different assessment methods or at different stages in the curriculum, both of which are justifications that the instructors' used to guide their decisions when creating strategic plans for assessing each LO (see Appendix B). The 2018 remediation strategy produced less of an effect (1.47 ± 0.80 PLs per LO) on LO remediation (p < 0.001).

Table II-5 also shows the distribution of LO remediation effects beyond that of the mean. The individual LO remediation scores in 2016 and 2017 were quite similar; no LO achievement was negatively affected by the individualized remediation strategy, and the range distribution of positive effects for these two years resulted as nearly the same. However, the lower mean LO remediation effect from the standardized intervention in 2018 negatively shifted its range distribution of effects to include shifting three LOs further away from achievement rather than toward achievement. Additionally, while the high-end ranges for mean LO remediation respectively reached 4.33 PLs and 5.00 PLs per LO in 2016 and 2017, the high-end range in 2018 only reached 3.20 PLs per LO (not shown in Table II-5).

Overall, the results in Table II-5 suggest each LO retested was remediated by an entire knowledge level by reflecting a pre-R PL representing absent knowledge (no LO achievement) to reflecting a post-R PL representing partial knowledge (*better* achievement). However, because LO remediation is dependent on student RPs, the same principle addressed for student remediation applies here – that although any positive LO remediation is desirable, only positive LO RPs that result in LO achievement (*i.e.*, an associated individual or mean post-R PL ≥ 4.50) is *more* desirable.

Table II-6 shows that nearly all retested LOs were positively remediated by each cohort. Though, as seen with student RPs, fewer LOs were actually achieved due to the remediation. Most LOs were achieved from IRQs in 2016, less from IRQs in 2017, and least from SRQs in 2018. This suggests two possible conclusions: 1) more LOs are achieved through individualized remediation strategies, and 2) although SAEs do not seem to have any greater effect on student remediation in comparison to instructor feedback, self-assessment still influences overall LO achievement. Regardless, achieved LOs were associated with greater mean pre-R PLs, mean post-R PLs, and mean remediation scores than those that weren’t. Individual LO remediation data (including resulting achievement) is presented in Appendix C.

Table II-6. Direction of LO Remediation and Resulting Achievement. All retested LOs were positively remediated nearly unanimously for all years. However, the most LOs were achieved from IRQs in 2016, and the least LOs were achieved from SRQs in 2018. In general, LOs that were achieved were associated with greater mean Pre-R PLs, mean Post-R PLs, and mean remediation scores. Means are presented as "MEAN ± SEM (StDev)." Individual LO remediation breakdown (including resulting achievement and how many times each LO was retested) can be viewed in Appendix C.



		Number of LOs from Classwide Performance:		
		2016 (IRQs)	**2017** (IRQs)	**2018** (SRQs)
Total Retested		122 (of 125)	123 (of 125)	72 (of 52)*
Negative Mean Remediation		0 (0%)	0 (0%)	3 (4.2%)
Positive Mean Remediation		122 (100%)	123 (100%)	69 (95.8%)
LO Remediation Achieved**		**83 (68.03%) (p < 0.001)**	**67 (54.47%) (p < 0.001)**	**31 (43.06%) (p < 0.001)**
For LOs **Achieved...**	Mean Pre-R PL	2.73 ± 0.07 (0.64)	2.47 ± 0.08 (0.64)	3.25 ± 0.13 (0.73)
	Mean Post-R PL	5.03 ± 0.04 (0.39)	5.03 ± 0.05 (0.40)	5.12 ± 0.07 (0.37)
	Mean R	2.30 ± 0.08 (0.76)	2.55 ± 0.09 (0.77)	1.87 ± 0.13 (0.73)
For LOs **Retested but Not Achieved...**	Mean Pre-R PL	2.58 ± 0.07 (0.45)	2.52 ± 0.07 (0.54)	2.76 ± 0.08 (0.52)
	Mean Post-R PL	4.06 ± 0.05 (0.34)	4.02 ± 0.05 (0.37)	3.92 ± 0.05 (0.34)
	Mean R	1.48 ± 0.09 (0.56)	1.51 ± 0.08 (0.58)	1.16 ± 0.11 (0.70)

*52 different LOs, 20 of which were retested twice or thrice in lab/lecture or from mid-units I/II overlap and treated separately.

**Set by instructors as classwide mean Post-R PL ≥ 4.50 (mid-partial knowledge or greater); indicates that students successfully remediated the respective LO to a desirable level of achievement.

As nearly all LOs were retested by at least one student in 2016 and 2017, the frequencies with which each LO was retested were different. Individual LOs were retested within the entire cohort between 2 and 120 times in 2016 and between 1 and 153 times in 2017 depending on their frequency to each student's poorest pre-R PLs on each examination (Appendix C). This wide range suggests that students struggled with certain LOs more than others. These highly variable results support the concept of an individualized remediation strategy – *i.e.*, retesting students individually based on LOs for which they need to be retested whether that means retesting a student on an LO on which no one else needs to be retested or on an LO on which everyone needs to be retested.

LOs that received the highest attention for individualized remediation in 2016 and 2017 were used for the comparative standardized remediation strategy in 2018. While less LOs were retested overall due to the standardization, less LOs were proportionally achieved by students through remediation. This is mostly due to higher associated pre-R PLs, less range for remediation, and insufficient preparation for LOs being retested due to unawareness.

Alignment of Student Perceptions of Performances

In order to validate the efficacy of our ordering of CBA PLs (and the resulting student RP and LO achievement results), student perceptions of CBA PLs regarding their use and order were collected from their SAEs in 2017 and 2018. Because SAEs were implemented after each set of examinations, each student's perceptions were collected six times throughout the course.

After each set of examinations (lecture and laboratory), answer sheets, a printout of the examination questions, an answer key, and a one-page SAE were returned to students. The instructions on the SAE read:

> *"While examining your answers, grading marks, and confidence levels on your examination answer sheets, fill in the tables below with the 6 learning objectives you believe you most need to show remediation (improved performance) on in lecture and in laboratory as you prepare to become an Occupational Therapist. A purposeful and thoughtful completion of this self-assessment will earn you full credit for this*

assignment. Learning Objectives can be found in the Course Syllabus, and rankings for Learning Objectives can be found on the back of this page."

Students were previously taught how to interpret confidence-based PLs, including what knowledge level each PL represents. They were instructed to photograph the SAEs and use them to guide their remediation-related studying. The SAEs were collected, and the six LOs each student recorded for each examination were analyzed for alignment to the LOs on which they would be retested during the subsequent examination as determined individually by Snow's rank of CBA PLs in 2016 and 2017 and standardized by most frequently retested LOs from 2016 and 2017 in 2018.

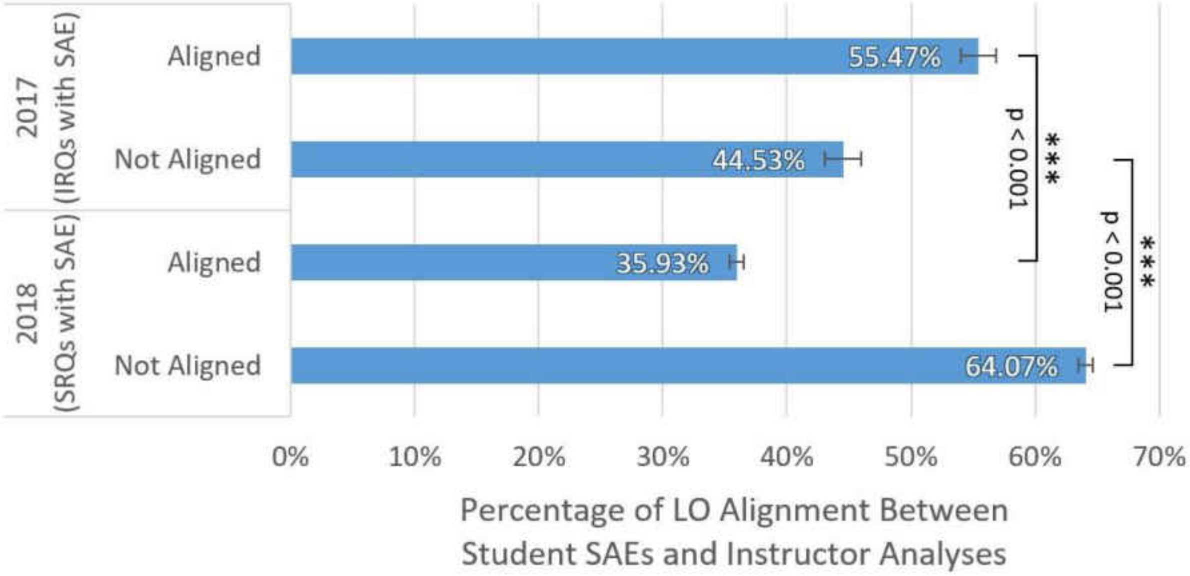


Figure II-6. LO Alignment between Student and Instructor Performance Perceptions. All IRQ/SRQ LOs were determined by the course instructors using Snow's Rank of Confidence-Based PLs. In 2017, each student's IRQs aligned with their respective SAE perceptions over half of the time. However, this alignment decreased significantly in 2018 when SRQs were introduced. Regardless, student perceptions in 2017 (when IRQs were implemented) and 2018 (when SRQs were implemented) did not indicate strong alignment to Snow's Rank of confidence-based PLs. Standard error is also presented.

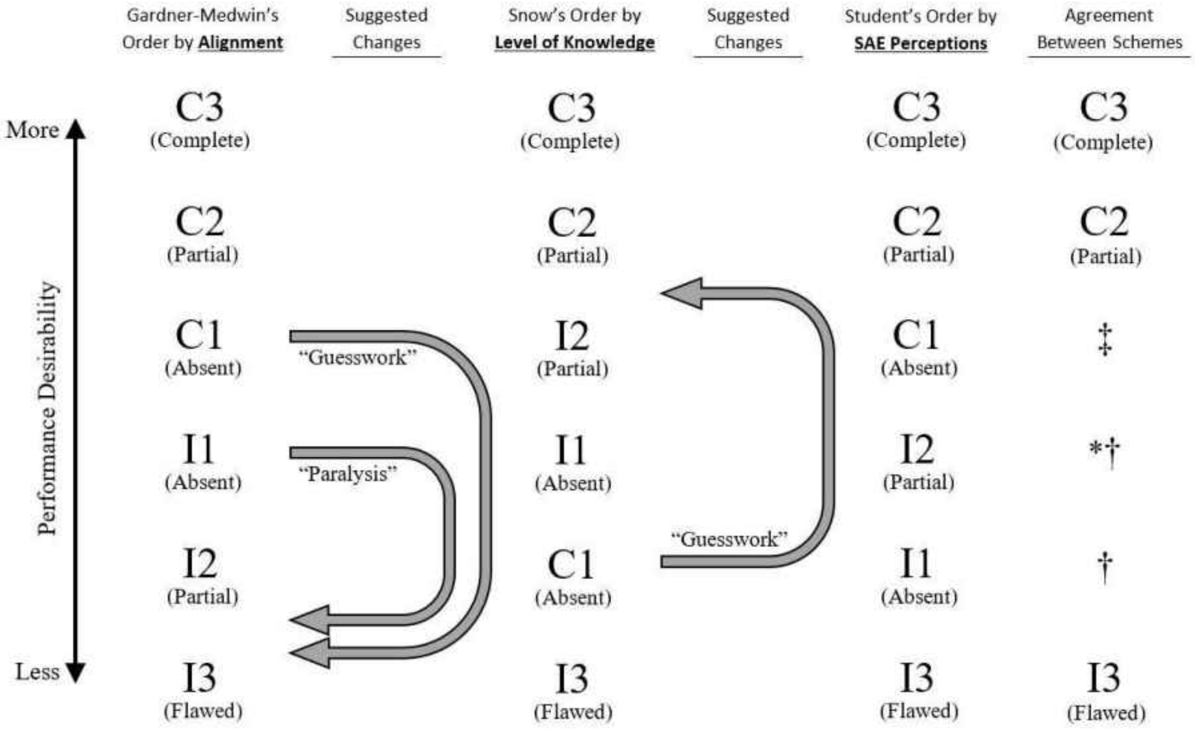
In 2017, student perceptions of LOs on which they most needed to be retested were only 55.47% aligned with those on which they actually needed to be retested. This alignment decreased to 35.93% in 2018 as a result of the standardized nature of the SRQs LOs. This result validates earlier discussions about how student awareness of which LOs they will be retested on impacts their learning; when they are retested individually, students demonstrate more alignment (and therefore more awareness) of their RQ LOs (Figure II-6), learn more (Tables II-3 and II-4), and achieve more LOs (Tables II-5 and II-6). While this alone is an important result of Figure II-6, it does not discount the impact of a lower possible remediation range from 2018's SRQs.

Assuming there would be further need for explaining misalignment between students' perceived worst performances from their SAEs and their *actual* worst performances, one last question on the SAEs asked the students to rank the confidence-based PLs in order of most-to-least need for remediation (*i.e.*, worst to best). This same question was also asked in an end-of-course survey for a total of 7 of these responses per student throughout the course. The results of this question are presented in Table II-7 in comparison to Gardner-Medwin's and Snow's rankings of confidence-based PLs.

In general, students agree with both Gardner-Medwin and Snow in that C3 (complete knowledge) and I3 (flawed knowledge) are respectively the most and least desirable PLs. Their ranking of the other PLs, however, presents both questions and answers. Although in different positions, the student's ranking of I2 over I1 aligns with Snow's ranking but not Gardner-Medwin's. However, students ranked C1 (absent knowledge), the guesswork performance, as being more desirable than any other incorrect performance. In fact, this position for C1 aligns with Gardner-Medwin's ranking despite instruction to each cohort about how guesswork, as indicated by correct but low confidence performances, is the second-worst confidence-based performance possible. This misperception was unexpected.

However, this provides a likely explanation for poor alignment between student’s perceptions of their six poorest performances and their six actual poorest performances (the poor alignments presented in Figure II-6) because C1 performances often identified LOs for IRQs and SRQs (Figure II-5).

Table II-7. Student Rank of Confidence-Based Performances (addendum to Table II-1). In general, students agree with both Gardner-Medwin and Snow regarding the best two and worst one confidence-based performances. Their rankings of I2 and I1 PLs mirrored Snow’s rank while their ranking of C1 mirrored Gardner-Medwin’s.



Gardner-Medwin’s Order by **Alignment**	Suggested Changes	Snow’s Order by **Level of Knowledge**	Suggested Changes	Student’s Order by **SAE Perceptions**	Agreement Between Schemes
C3 (Complete)		C3 (Complete)		C3 (Complete)	C3 (Complete)
C2 (Partial)		C2 (Partial)		C2 (Partial)	C2 (Partial)
C1 (Absent)	“Guesswork”	I2 (Partial)		C1 (Absent)	‡
I1 (Absent)	“Paralysis”	I1 (Absent)		I2 (Partial)	*†
I2 (Partial)		C1 (Absent)	“Guesswork”	I1 (Absent)	†
I3 (Flawed)		I3 (Flawed)		I3 (Flawed)	I3 (Flawed)

*Indirect agreement between Gardner-Medwin and Snow performance orders.

†Indirect agreement between Snow and Student performance orders.

‡Direct agreement between Gardner-Medwin and Student performance orders.

Overall, the student’s ranking of the confidence-based PLs represented aspects of both Gradner-Medwin’s and Snow’s rankings. Because each student was ranking these according to their own perceptions, large standard deviations (nearly two PLs for each) accompanied each PL in their presented mean ranking. The high standard deviations could mean that the students did not fully understand how to interpret the confidence-based performances, which could explain why they ranked C1 so highly. Additionally, given that the three main performances in questionable rank are within two PLs of precisely matching either Gardner-Medwin's or Snow's ranking, no conclusion can be made as to with which ranking scale the student's ranking agreed more.

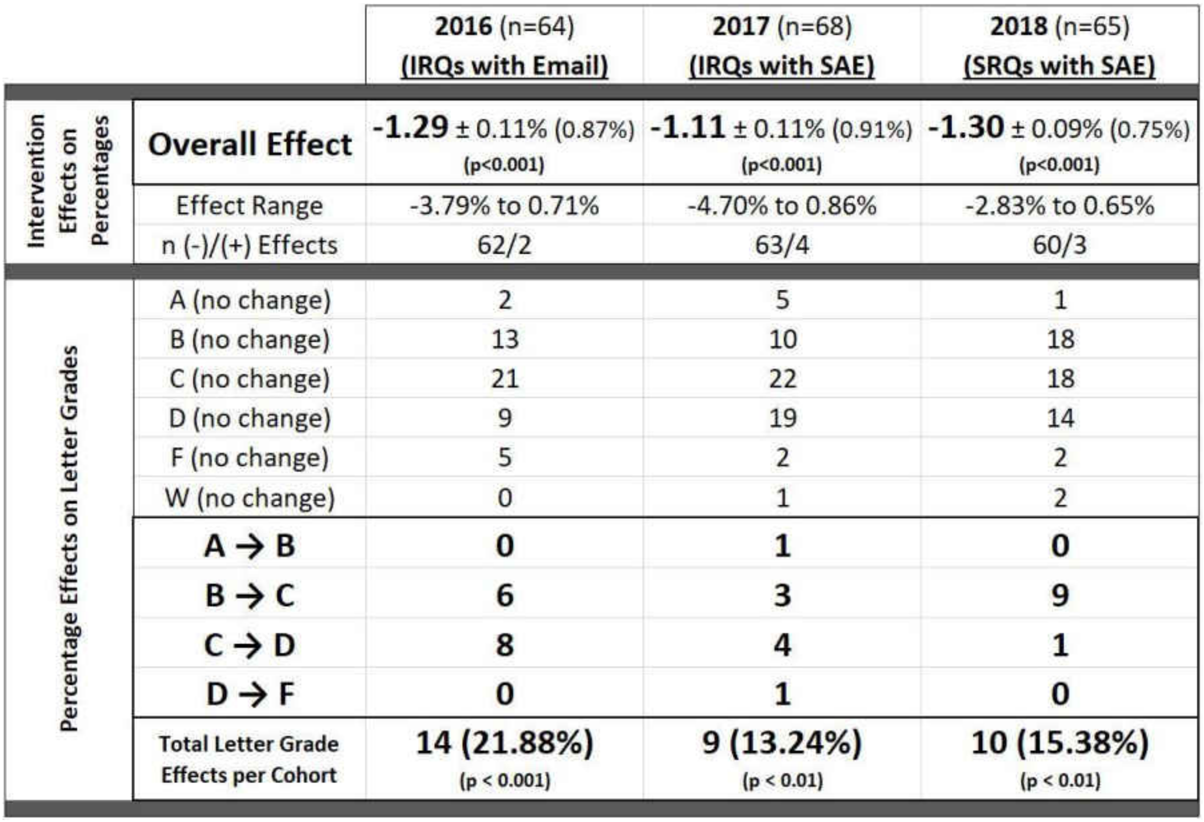
Given these results, however, we hypothesize that if the students had better understood how to interpret and rank performances based on the knowledge level they represent they would have attained more remediation, achieved more LOs, and would have had perceptions about performances better aligned to those of the instructors.

Effects on Student Grades

The major principle behind the use of our ranking scheme for CBA PLs is that it ranks performances based on knowledge levels for remediation purposes, but in doing so uses the conventional correctness-only "number correct" scoring criteria (unlike Gardner-Medwin's CBA scoring criteria) to be effective in detecting performances with low confidence levels. While the purpose of ranking the PLs by level of knowledge was to help students identify their poorest performances (primarily those indicating flawed and absent knowledge) and remediate those performances, the positive effects of this intervention should also be reflected in the students' grades since grades are *expected* to be accurate representations of student knowledge and competence.

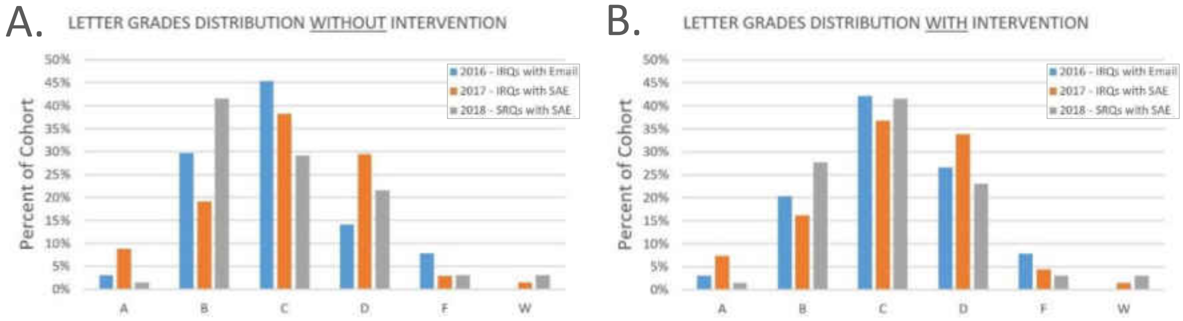
The IRQ's and SRQ's pre-R and post-R PLs, determined by a combination of correctness and confidence of students' responses (see Table II-1) but scored only by correctness (see Table II-2), caused percentage grades to decrease for nearly every student (see Table II-8). Although the mean percentage grade decrease was only slightly over 1% per student, it ultimately led to a change in letter grade for 14 students (21.9%) in 2016, 9 students (13.2%) in 2017, and 10 students (15.4%) in 2018. This effect was especially significant for the few students whose letter grade dropped below the minimum parameter of acceptable performance due to the intervention. In OT 422, a D is considered unacceptable but typically results in a probationary period until the student retakes the course, and an F typically results in dismissal from the program. As Table II-8 demonstrates, 14 total students among all three cohorts fell below the acceptable performance line, one of which qualified for dismissal. Students who retook the course were included in the following year's intervention analyses.

Table II-8. Effects of IRQ and SRQ Interventions on Percentage and Letter Grades. Percentage grades for nearly all students were found to decrease due to the interventions. Percentage grade decreases also resulted in letter grade changes for many students. Letter grades were determined from the percentage of points earned from 12 total examinations and 72 total remediation questions. The letter grade scales were as follows: A ≥ 93.45%; B = 85.45% to 93.44%; C = 77.45% to 85.44%; D = 69.45% to 77.44%; F ≤ 69.44%. W = withdrawal from the course.



		2016 (n=64) **(IRQs with Email)**	**2017** (n=68) **(IRQs with SAE)**	**2018** (n=65) **(SRQs with SAE)**
Intervention Effects on Percentages	**Overall Effect**	**-1.29** ± 0.11% (0.87%) (p<0.001)	**-1.11** ± 0.11% (0.91%) (p<0.001)	**-1.30** ± 0.09% (0.75%) (p<0.001)
	Effect Range	-3.79% to 0.71%	-4.70% to 0.86%	-2.83% to 0.65%
	n (-)/(+) Effects	62/2	63/4	60/3
Percentage Effects on Letter Grades	A (no change)	2	5	1
	B (no change)	13	10	18
	C (no change)	21	22	18
	D (no change)	9	19	14
	F (no change)	5	2	2
	W (no change)	0	1	2
	A → B	**0**	**1**	**0**
	B → C	**6**	**3**	**9**
	C → D	**8**	**4**	**1**
	D → F	**0**	**1**	**0**
	Total Letter Grade Effects per Cohort	**14 (21.88%)** (p < 0.001)	**9 (13.24%)** (p < 0.01)	**10 (15.38%)** (p < 0.01)

Figure II-7 shows the letter grade distribution shifts for all three cohorts. As expected from Table II-8, the overall percentage grade decreases caused a significant rightward shift in letter grade distributions. Because student population numbers varied between cohorts, letter grade distributions are reported in percentage of population instead of number of students so that accurate comparisons between cohort distributions can be made. Interestingly, we found that the shift in letter grades with intervention, which was slightly different between cohorts, made the final letter grade distributions more comparable between years (Figure II-7B) than without intervention (Figure II-7A).



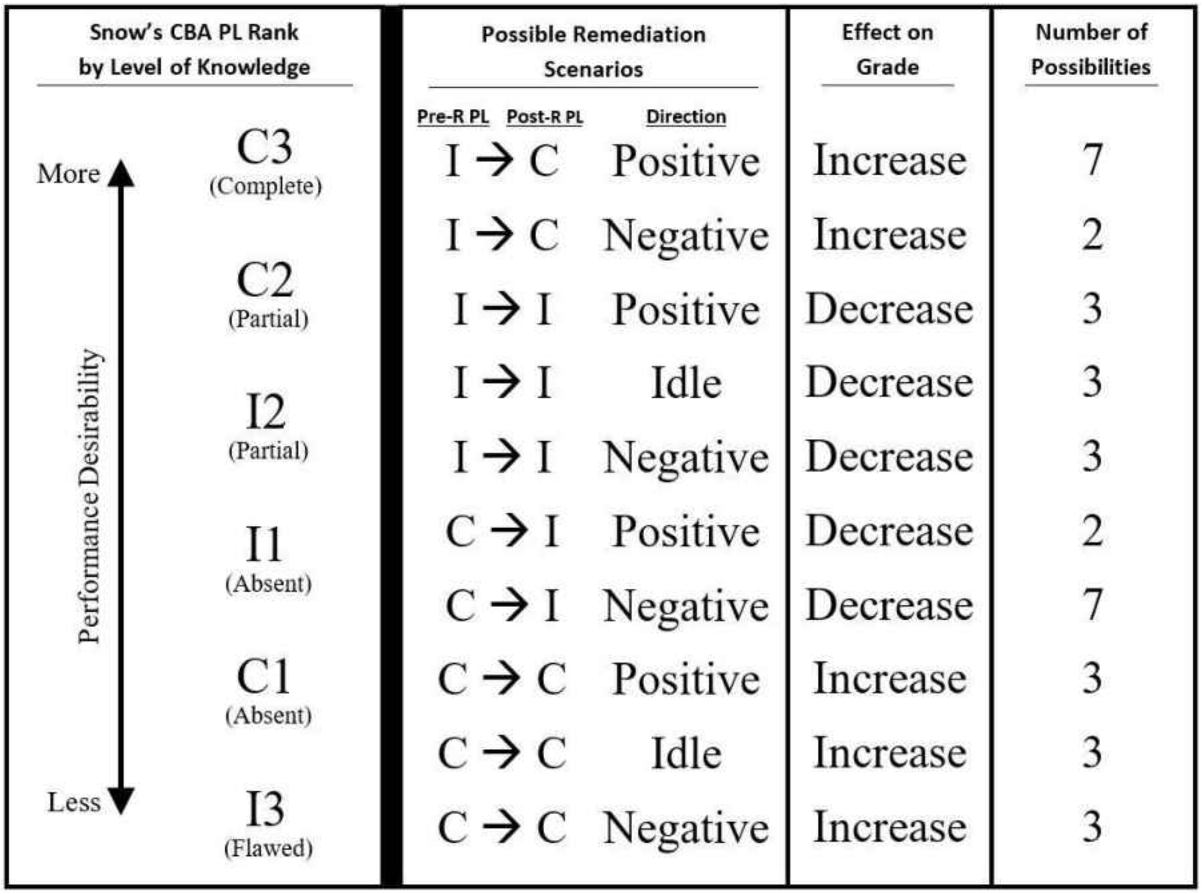
C. Kolmogorov-Smirnov (K-S) Test for Letter Grades Distribution Normality

	WITHOUT INTERVETNION				WITH INTERVETNION			
Cohort	D_{max}	$D_{critical}$ (α = 0.001)	Diff.	Distribution	D_{max}	$D_{critical}$ (α = 0.001)	Diff.	Distribution
2016	0.239	0.244	0.005	Normal	0.206	0.244	0.038	Normal
2017	0.222	0.238	0.016	Normal	0.213	0.238	0.025	Normal
2018	0.247	0.246	-0.001	Abnormal	0.207	0.246	0.039	Normal

Figure II-7. Remediation Intervention Effects on Letter Grade Distributions. Letter grade distributions with (B) and without (A) the intervention are presented for comparison. The intervention's rightward shift toward lower grades caused letter grades to be more normally distributed according to K-S test statistics (excluding W grades). Distributions are presented in percentages per cohort to better compare populations of different numbers (64 in 2016, 68 in 2017, and 65 in 2018).

The interventions' effects on grades poses an important question: why would an intervention that increases student learning and classwide achievement of LOs also cause student grades to decrease? The critical consideration is that the only performances directly tied to the interventions are the post-R performances; pre-R performances reflect performances on routinely administered unit examinations that affect grades regardless of intervention. Table II-9 lists all possible remediation scenarios and their respective post-R PL effects on grades. Understanding and using these scenarios, the answer to the question lies in the rationale for using a correctness-only scoring criteria for the employed CBA PL scheme.

Table II-9. Possible Remediation Scenarios and Their Effects on Grades. While Snow's CBA PL Rank determines the direction of remediation between one performance and another, only the correctness of the intervention's Post-R PL influences a student's grade.



Snow's CBA PL Rank by Level of Knowledge		Possible Remediation Scenarios		Effect on Grade	Number of Possibilities
		Pre-R PL Post-R PL	Direction		
More ▲	C3 (Complete)	I → C	Positive	Increase	7
		I → C	Negative	Increase	2
	C2 (Partial)	I → I	Positive	Decrease	3
Performance Desirability	I2 (Partial)	I → I	Idle	Decrease	3
		I → I	Negative	Decrease	3
	I1 (Absent)	C → I	Positive	Decrease	2
		C → I	Negative	Decrease	7
	C1 (Absent)	C → C	Positive	Increase	3
		C → C	Idle	Increase	3
Less ▼	I3 (Flawed)	C → C	Negative	Increase	3

The ten possible remediation scenarios are listed in Table II-9 next to Snow's CBA PL rank for convenient referencing. Remediation scenarios can be incorrect-to-correct, incorrect-to-incorrect, correct-to-incorrect, or correct-to-correct and result in positive, negative, and/or idle remediation directions. Regardless, only the correctness of post-R PLs dictates the remediation effect on students' grades. Given the location of each PL in rank, there are 15 possible scenarios that would result in either positive or negative remediation and 6 for idle remediation, yet their actual frequencies (see Table II-4) suggest that they do not occur proportionally. Furthermore, there are an equal number of remediation scenarios that would result in increasing and decreasing effects on grades, yet we know that grade-decreasing scenarios had to be more common than grade-increasing ones to cause the mean overall grade decrease. The actual performances contributing to these remediation scenario categories were not expected to be equally proportional. Because incorrect performances are more common within the lower ranks of Snow's CBA PLs and correct answers are more common in the higher ranks, and we strategically chose low-rank PLs to serve as the basis for remediation, we expected that the remediation scenarios would be unequally frequent among the remediation strategies, favoring those that would be more common with pre-R PLs toward the bottom of the scale.

Consequently then, positively remediating I3 (the first-chosen PL for remediation) with any of the mean student remediation scores (see Table II-3) would result in an incorrect post-R PL and contribute to the decrease in grades. Further, students were also often retested on C1 pre-R performances (see Figure II-5). Although C1 is one of the least desirable PLs, students were still given full credit for choosing/recording the correct answer (see Table II-2) despite its clear representation of absent knowledge from the student. Once again, according to the mean student remediation scores (see Table II-3) and their positive direction, students are more likely to remediate the C1 pre-R PL to an incorrect (but better) post-R PL in which they would receive no credit. A *correct-to-incorrect positive* remediation like this that results in inversely-awarded credit (i.e. credit is awarded for a poor pre-R performance on the unit examination but not for a better post-R performance on the respective remediation question) is *only possible given a pre-R PL of C1*.

Because credit is inversely-awarded in this scenario, the intervention causes a decrease in grade despite a positive remediation from C1. As such, it is as if the deserved but unawarded credit for the incorrect post-R performances in these scenarios compensates for the undeserved but awarded credit for the respective C1 Pre-R performances. All five other correct-to-incorrect remediation scenarios result in negative remediation. Any of these *correct-to-incorrect negative* remediation scenarios, in addition to *incorrect-to-incorrect idle* remediation scenarios, result in decreased grades from incorrect post-R performances and are justified by their respective negative/idle remediation directions. However, in these cases, the incorrect post-R performances are either less desirable or equally desirable than the respective pre-R performance, much unlike the scenario regarding C1, and we also know negative and idle remediation scenarios happen quite infrequently (see Table II-4) compared to positive ones. Relatedly, as only a minor category of negative remediation scenarios, *incorrect-to-correct negative* remediation scenarios (*i.e.*, I2 or I1 to C1) did not occur often, nor did any *correct-to-correct idle* remediation performances. The C1 to C1 remediation scenario was especially rare due to the greater likelihood of choosing an incorrect answer if having to randomly guess once again.

Alternatively, any of the *incorrect-to-correct positive* remediation scenarios results in credit earned from choosing/constructing the correct answer as a result of positive remediation. Here, the credit earned would rightfully increase the student's grade. However, this tends not to happen as often due to the majority of the correct PLs at the very top of the ranking scale, resulting in more post-R incorrect performances contributing to the mean grades decrease. All of these reasons contribute to explaining the overall decrease in grades.

The evidence demonstrates that correctness-only “number correct” scoring criteria significantly decreased grades. However, although this effect may seem unfortunate, *it does not suggest the effect is negative*. In fact, we believe it to be an acceptable effect as it suggests a more accurate representation of student knowledge, at least to some degree, understanding that guesswork (*i.e.*, C1 performances) falsely increases grades determined by correctness-only scoring criteria. However, because we implemented a novel confidence-based remediation strategy and used correctness-only scoring criteria to retest most guesswork performances, we accept a decrease in overall grades understanding it was largely due to C1-to-I1/I2 positive remediation performances. In fact, other studies have suggested much harsher grading criteria for correcting guesswork (*i.e.*, C1 performances) and misinformation (*i.e.*, I3 performances) without a remediation intervention. For these reasons, we believe the resulting effect on the grades more reasonably and accurately represent student learning and knowledge achievement.

Alignment of Student Perceptions of Intervention Effects

Lastly, because we previously noted significant student perceptions regarding interpretations of the confidence-based performance levels that explained some outcomes of this study, we also examined student perceptions of how they believed the interventions were affecting their overall grades as well as their mastery and retention of course material. We asked each cohort the following questions (see survey results in Figure II-8):

1) *What effect do you believe the individualized questions had on your overall grade?*

A. They increased my grade.

B. They made no change to my grade.

C. They decreased my grade.

2) *What effect do you believe the individualized questions had on your mastery and retention of the content?"*

A. They helped me to master and retain more content.

B. They did not affect my mastery and retention of the content.

C. They caused me to master and retain less content.

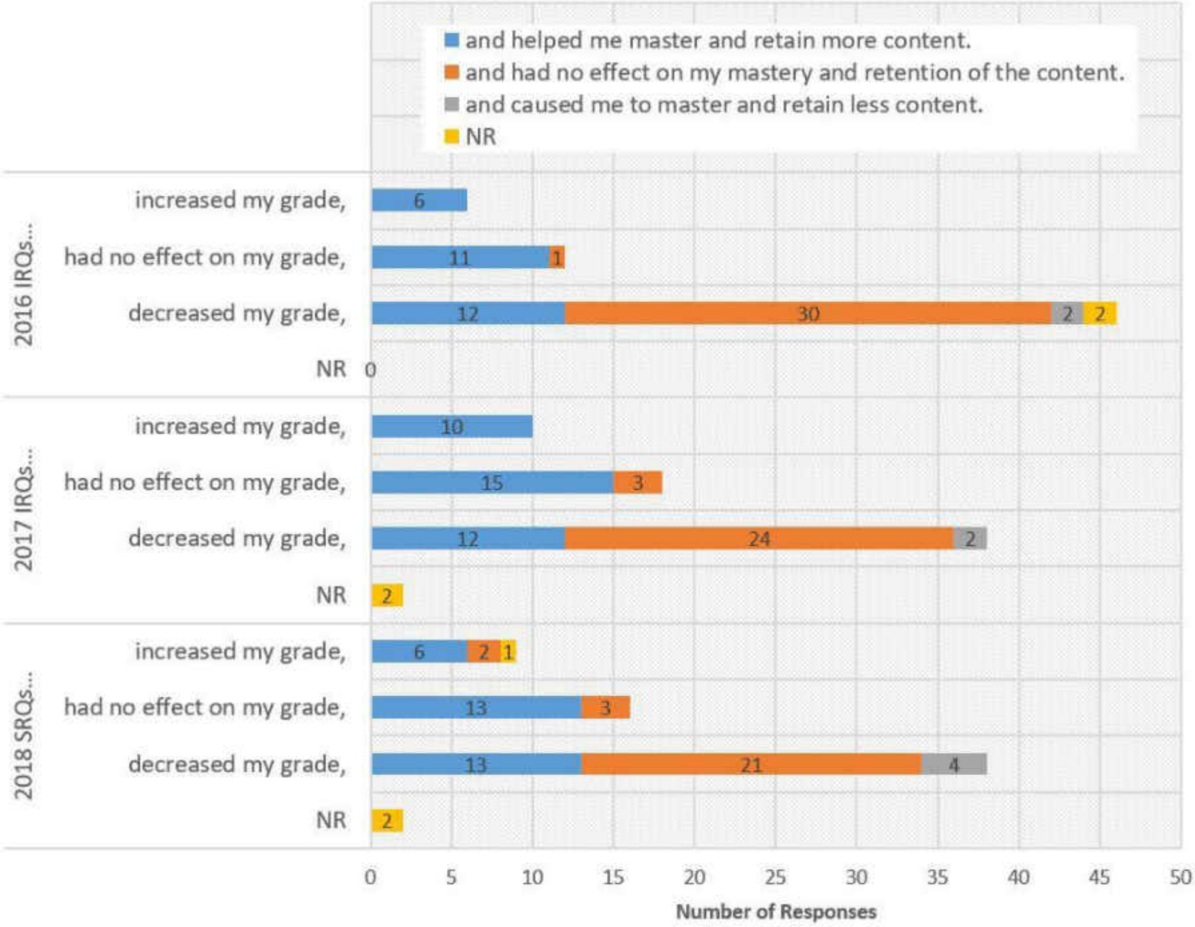


Figure II-8. Survey Results of Intervention Impact on Learning and Grades. In an end-of-the-semester course survey, students were asked what effect they believed the intervention had on their learning and grades. Only 12-13 students per cohort had perceptions that directly agreed with the actual resulting data. NR = No Response.

When reviewing survey question results separately, a majority of each cohort believed the interventions were causing their grade to decrease. However, each cohort was split in agreement about whether or not the interventions helped them master and retain more content or had no effect on their mastery and retention of the content. Only 12-13 students per cohort believed that the interventions were both helping them master and retain more content while also decreasing their grades. While contradictory at first glance, these data suggest students did not have an understanding of how the interventions were designed to both help them learn and better assess their knowledge. No significant differences in these perceptions were found to exist between cohorts.

Discussion and Conclusions

The central idea throughout this project is about how to best assess knowledge. Given that *better performances* should be associated with *more desirable knowledge levels*, correctness cannot be used alone to assess knowledge as we know that positive RPs can and often do result from remediating correct but undesirable pre-R PLs to incorrect but more desirable post-R PLs. In the same light confidence cannot be used alone to accurately assess knowledge. Alone each can present both valuable yet falsely convincing data, but together they assess knowledge and present accurate interpretations of that knowledge.

The individualized, LO-based remediation interventions employed in this study demonstrate how confidence-based assessment strategies can be used to more meaningfully assess student performance. As a result of the methods used to implement the individualized and standardized RQs, Pre-R performances were more associated with lower PLs in comparison to post-R performances. This result suggested that retesting poorly-performed LOs with RQs enhanced student ability to demonstrate what they learned with better aligned levels of correctness and confidence, and this led to an increase in class-wide LO achievement.

Although each had a significant impact on their respective cohorts, IRQs were found to have more of an impact on learning and LO achievement than SRQs due to the greater possible remediation ranges and awareness of LOs retested for IRQs. For these reasons, it was important to learn that low remediation scores are *not necessarily* undesirable if students are performing well to begin with. They are only undesirable if they are caused by poor post-R performances given a greater range for remediation. Additionally, while any positive RP is desirable, it does not necessarily mean a satisfactory PL was reached or the respective LO was achieved.

Nearly all students demonstrated mean positive remediation. However, students who displayed lower levels of remediation on retested LOs were not necessarily those that received lower grades; in fact, students with lower grades tended to achieve higher remediation from the individualized remediation intervention. These results suggest that final grades awarded according to a predetermined scale based solely on a total number of available points may not be strictly valid indicators of student learning and achievement of competency.

Similar to their effect on learning, each intervention decreased student grades due to post-R PLs being more commonly incorrect in the more likely remediation scenario possibilities. However, although it may seem unfortunate at first, this effect is accepted by the instructors as correction for using correctness-only scoring criteria in order to detect pre-R C1 (guesswork) PLs for remediation. Students seemed to be aware that the IRQs and SRQs were decreasing their grade while also helping them master and retain more content (or at least having no impact on their learning); however, few of them correctly noticed both seemingly contradictory effects that the data show, suggesting that students may not have understood how the interventions were designed to help them learn and better assess their levels of knowledge in the course LOs. Students believed they were learning more *at the expense* of their grade.

Whether students received feedback about their performance from instructors or self-assessed their own performance did not seem to impact student learning, LO achievement, or grades. For this reason, the 2018 SRQ performances can be compared to both years. SAE results suggest students in all years may have not have had as good of an understanding of the mid-range PLs as the instructors had hoped, and this could mean that more potential for learning and achievement could have been reached if their perceptions would have been better aligned with the instructors'. As a result, a limitation for the SAE-related interventions was that instructors chose questions based on Snow's rank of CBA PLs and not the results of the students' SAEs. As such, the misalignment in student vs. instructor choice for poorest-performed-on LOs expectedly would have influenced their remediation performances in a negative manner, again suggesting a potential for better learning and achievement if students had better understood how to interpret and use confidence-based performances according to their respective knowledge levels.

To best understand the data, we consider other factors that may have influenced remediation efforts and results. While the effect of remediation is dependent on post-R PLs, the type of remediation is dependent on pre-R PLs. For example, students tasked with remediating pre-R I3 performances take a very different approach to their remediation efforts than students tasked with remediating pre-R C1 performances; remediating I3 (flawed knowledge) performances typically involves finding and re-learning the small piece of knowledge that was learned incorrectly, whereas remediating C1 (absent knowledge) performances involves learning (for the first time) the material for the respective LO. Therefore, one could conclude that remediating some pre-R PLs takes more effort than others.

Similarly, individual LOs were retested within the entire cohort between 2 and 120 times in 2016 and between 1 and 153 times in 2017. This suggests that certain LOs were more difficult for students to achieve. The important aspect of LO-based assessments is that they are intentionally and strategically measuring LO achievement.

The LO remediation analysis presented in this study is done with one significant limitation. LO remediation, as presented, is only according to the remediation performances based on a single poor pre-R performance, when in fact each LO is tested by many questions (see Appendix B). A scenario could exist where the mean overall performance for an LO reached a level of achievement from the routine unit examination questions despite a single pre-R performance identifying that LO for remediation. Ideally, the remediation strategies implemented would not have retested an LO based on a single-question poor performance but rather its overall performance. We were unable to do this due to the infeasibility of managing mean performances per LO across multiple assessment types, e.g., lecture and laboratory, whereas handling individual performances per LO on lecture examinations was more manageable.

As aforementioned, other researchers have studied similar aspects of multidimensional assessments that support the information presented by this study. Kate Hevner's initial use of confidence in assessment led her to showing how her confidence-based true/false assessments led to greater reliability than correctness only methods.[22] Gardner-Medwin's entire collection of work, from initial studies to validity and reliability studies, especially confirms the claims of confidence-based assessments[3–5,23,24] and further strengthens the results presented in this study as well.

Other researchers have studied related topics this study that are a bit more indirect in relation. General student remediation, especially in medical education, is a topic studied my many educators. Identifying underperforming students and helping them improve their performance is an ongoing concern and challenge for medical institutions. Karen Hauer *et. al.* examined how US medical schools approach remediation of students exhibiting poor, unsatisfactory clinical skills. Hauer found that most medical institutions' remediation efforts consisted of three steps: 1) identifying/diagnosing learning insufficiencies, 2) intervention of remedial activities, and 3) re-testing of the previously-failed material.[25] These findings reflect the steps taken in the study presented in this dissertation – learning deficits were identified, feedback was presented to the students to guide remedial learning, and students were re-tested on the relevant material. Hauer even goes on to describe how many remediation examinations are often shorter and more formative than the original examinations. Interestingly, Varun Saxena *et. al.* worked with Hauer to also examine how confident institutions were about their remediation efforts and identified both strengths and weaknesses of general remediation techniques,[26] slightly opposite to who we collected confidence from in this study but still collecting it in the same manner and for similar purpose.

Furthermore, in 2011, Steven Durning *et. al.* examined how self-regulation can be used to help medical educators with assessment and remediation. Durning used a method called the Self-Regulated Learning – Microanalytic Assessment and Training (SRL-MAT) to not only identify underperforming students but also determine the cause of their knowledge deficits in real time during the assessment of clinical skills. Durning describes how this method can and should be combined with summative evaluation of performances to enhance remediation efforts to a more meaningful and effective level.[27] Durning compared the methodology he used to ones similar to those used in this dissertation. In doing so, he described how purely summative assessments of remediation performance could benefit from using the SRL-MAT or other real time metanalytic assessment techniques. Hauer, Saxena, and Durning all recognize the substantial challenges medical institutions face regarding student remediation efforts, specifically those related to resources and dedication of time from faculty. Perhaps the more educator scholars work to identify best methods of remediation, more efficient methods for implementation will also ensue.

In conclusion, this dissertation study demonstrates how the novel use of a confidence-based assessment method can be used to create an individualized remediation strategy to positively influence student learning and LO achievement by more meaningfully assessing student knowledge. Future use of this intervention method should prove more impactful upon addressing the limitations addressed. Better understanding student confidence and attitudes toward learning may also inform other strategies to successfully remediate poor academic performance.

References

1. Hunt, D. P. The concept of knowledge and how to measure it. *J. Intellect. Cap.* **4**, 100–113 (2003).

2. Ghadermarzi, M., Yazdani, S., Pooladi, A., Bahram-Rezaei, M. & Hosseini, F. A Comparative Study between the Conventional MCQ Scores and MCQ with the CBA Scores at the Standardized Clinical Knowledge Exam for Clinical Medical Students. *J. Med. Educ.* **14**, 31–37 (2015).

3. Gardner-Medwin, A. R. Confidence assessment in the teaching of basic science. *ALT-J* **3**, 80–85 (1995).

4. Gardner-Medwin, A. R. Confidence-Based Marking: Encouraging rigour through assessment. in *University of Bristol. Proceedings of The Physiological Society* **J Physiol 567P**, **WA10**, (Physiological Society, 2005).

5. Gardner-Medwin, A. & Curtin, N. Certainty-Based Marking (CBM) for reflective learning and proper knowledge assessment. in *Re-Engineering Assessment Practices (REAP) Int. Online Conference on Assessment Design for Learner Responsibility, Proceedings for Raising students' meta-cognition (self-assessment) abilities* (2007).

6. Farrell, G. & Leung, Y. Convergence of validity for the results of a summative assessment with confidence measurement and traditional assessment. in *12th CAA International Computer Assisted Assessment Conference : Proceedings of the Conference on 8th and 9th July 2008 at Loughborough University.* 195–204 (Loughborough University, 2008).

7. Adams, T. M. & Ewen, G. W. The Importance of Confidence in Improving Educational Outcomes. in *University of Wisconsin System's 25th Annual Conference*

on Distance Teaching and Learning 1–5 (2009).

8. Kissel, H., Miller, B. J. & Young, H. *Writing Objectives*. (James Madison University).

9. Williams, B. *Writing Objectives*. (Penn State University).

10. Dalto, J. ABCD: The Four Parts of a Learning Objective. (2013). Available at: https://www.convergencetraining.com/blog/abcd-the-four-parts-of-a-learning-objective.

11. University of Connecticut. Writing Learning Objectives. (2014). Available at: https://kb.ecampus.uconn.edu/2014/07/31/writing-cognitive-objectives/.

12. Mager, R. F. *Preparing Instructional Objectives*. (Fearon Publishers, 1962).

13. Shabatura, J. Using Bloom's Taxonomy to Write Effective Learning Objectives. *Teaching Innovation & Pedagogical Support* (2018). Available at: https://tips.uark.edu/using-blooms-taxonomy/. (Accessed: 16th March 2019)

14. International Assembly for Collegiate Business Education (IACBE). Bloom's Taxonomy of Educational Objectives and Writing Intended Learning Outcomes Statements. (2016).

15. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. (David McKay Company, Inc., 1956).

16. Dalto, J. How to Write SMART Learning Objectives. (2013). Available at: https://www.convergencetraining.com/blog/how-to-write-smart-learning-objectives.

17. White, R. ABCD's of SMART Objectives. *Louisiana State University* Available at: https://www.slideshare.net/bwhitelsu/abcds-of-smart-objectives. (Accessed: 19th March 2017)

18. Bryan, W., DiMartino, J. & Center for Secondary School Redesign. *Writing Goals and Objectives: A Guide for Grantees of the Smaller Learning Communities Program*. (United States Department of Education. Academy for Educational Development, 2010).

19. Bush, M. Reducing the need for guesswork in multiple-choice tests. *Assess. Eval. High. Educ.* **40**, 218–231 (2015).

20. Jaradat, D. & Sawaged, S. The Subset Selection Technique for Multiple-Choice Tests: An Empirical Inquiry. *J. Educ. Meas.* **23**, 369–76 (1986).

21. Alnabhan, M. An Empiracle Investigation of the Effects of Three Methods of Handling Guessing and Risk Taking on the Psychometric Indices of a Test. *Soc. Behav. Personal. an Int. J.* **30**, 645–652 (2002).

22. Hevner, K. A Method of Correcting for Guessing in True-False Tests and Empirical Evidence in Support of IT. *J. Soc. Psychol.* **3**, 359–362 (1932).

23. Gardner-Medwin, A. R. & Gahan, M. Formative and Summative Confidence-Based Assessment. 147–155 (2003).

24. Gardner-Medwin, A. Confidence-Based Marking - towards deeper learning and better exams. In: Bryan, C and Clegg, K, (eds.) Innovative Assessment in Higher Education. in 141–149 (Routledge, Taylor and Group; Francis, 2006).

25. Hauer, K. E., Teherani, A., Irby, D. M., Kerr, K. M. & O'Sullivan, P. S. Approaches to medical student remediation after a comprehensive clinical skills examination. *Med. Educ.* **42**, 104–112 (2008).

26. Saxena, V., O'Sullivan, P. S., Teherani, A., Irby, D. M. & Hauer, K. E. Remediation techniques for student performance problems after a comprehensive clinical skills

assessment. *Acad. Med.* **84**, 669–676 (2009).

27. Durning, S. J. *et al.* Viewing ‘strugglers’ through a different lens: how a self-regulated learning perspective can help medical educators with assessment and remediation. *Acad. Med.* **86**, 488–495 (2011).

CHAPTER III

COMPARABILITY OF STUDENT POPULATIONS AND PERFORMANCES BETWEEN DISTRIBUTED LEARNING SITES

Abstract

Background

Delivering a comparable education to distributed learning sites poses many challenges, especially for one uniquely distributed Anatomy for Occupational Therapists course (OT 422) at the University of North Dakota. OT 422 is taught simultaneously between two learning sites – a home site in Grand Forks, ND on the University of North Dakota campus and a satellite site in Casper, WY on the Casper College campus. The Grand Forks and Casper campuses respectively enroll approximately 70% and 30% of each cohort annually. In addition to differences in population sizes and demographics, the lecture portion of the course is delivered simultaneously by the same instructor across sites via live two-way video conferencing from the home site, and although the teaching laboratories are also delivered simultaneously, they are delivered by different instructors without inter-site collaboration. The instructors use learning objective-based assessments to ensure comparable educational experiences are delivered to each site.

Methods

The instructors established OT 422 course learning objectives (LOs) according to the ABCD method[1–5] and to meet SMART criteria[6–10]. LO-based blueprints for assessment were established (Appendix B), which guided the development of six high-stakes lecture and

laboratory examinations. Both sites were administered the same multiple choice question (MCQ) lecture examinations. However, due to the sites receiving comparable (but different) laboratory experiences from different instructors and site resources, each site was administered different (but comparable) constructed-response question (CRQ) laboratory examinations. Each lecture and laboratory examination tested new material in addition to a set of remediation questions for each student based on poor performances on the previous examination. Student overall and remediation-specific performances, including their resulting grades, were collected from both sites' examinations. LO retesting frequencies and grades were examined for comparability between sites and between lecture and laboratory. Additionally, the Motivated Strategies for Learning Questionnaire (MSLQ)[11] was administered to better understand student motivations and strategies for learning. Site-specific motivations and learning strategies were then examined for comparability to their respective remediation performances and grades.

Results

Data was collected and analyzed from three cohorts of students from years 2016, 2017, and 2018. Two motivation scales (Self Efficacy for Learning and Performance, and Test Anxiety) and three learning strategies scales (Rehearsal, Metacognitive Self-Regulation, and Peer Learning) were found to be consistently different between Grand Forks and Casper students ($p < 0.01$). Two other motivations scales (Task Value and Control Beliefs) and two other learning strategies scales (Time and Study Environment, and Effort Regulation) – the four of which received the highest scores of the 15 total MSLQ scales – were correlated to increasing post-remediation (Post-R) performance levels (PLs) for both Grand Forks and Casper populations ($p < 0.05$). In general, Casper students tended to exhibit slightly more desirable pre- and post- remediation PLs than Grand Forks students. We hypothesize this may be due to how students allocate their time; Casper students reported spending more time studying for OT 422 ($p < 0.01$), and Grand Forks students reported spending more time working at an outside job ($p < 0.05$). Even so, no significant differences resulted in overall final percentage or letter grades between sites. Students at both sites did receive higher grades on lecture examinations than laboratory examinations though, likely due to the expected nature of MCQs vs. CRQs. Additionally, LOs retested in lecture were more comparable between sites than LOs retested in laboratory. This is also due in part to the nature of MCQs vs. CRQs but also due to differences, although comparable ones, in each site's laboratory examinations.

Conclusions

Delivering even a theoretically perfectly comparable education to distributed learning sites does not ensure comparable outcomes between sites. Different populations can, and often do, exhibit general differences in demographics, motivations for learning, learning strategies, and other behaviors, each of which can expectedly impact learning outcomes. We have shown how a novel LO-based assessment method can serve as an evidence-based strategy for delivering individualized but comparable education experiences to students across distributed learning sites. As a result, the individualized nature of this comparable assessment strategy is believed to better satisfy the expectedly different needs of each population and minimize the resulting effects of site differences on population outcomes. Grand Forks and Casper site performances were different, yet learning outcomes reflected by grades were comparable. These outcomes rely primarily on well-written learning objectives and a blueprint for incorporating them into the course's learning assessments.

Introduction

As an outgrowth of the advancement of globalization, electronic communication, and other changing socioeconomic forces increasing student mobility,[12] institutions across the country have provided learning experiences to distributed populations by establishing other physical distributed learning sites for additional face-to-face options or online experiences utilizing and emphasizing electronic communication (*i.e.*, *distance education* or *distance learning*). As a result, populations previously prevented from pursuing learning experiences due to time or location constraints have easier been able to pursue them. This way of delivering/receiving education, "…when the teacher and student are situated in separate locations and learning occurs through the use of technologies (such as video and internet), which many be part of a wholly distance education program or supplementary to traditional instruction," is known as *distributed education* or *distributed learning*.[13]

As described by Oblinger, expanding accessibility to learning experiences is important for many different individuals and situations. State/company employees may need education/training, underserved populations may be prevented from pursuing comparable learning experiences to others, and even for those who are fortunate to freely pursue educational experiences, many academic programs are not flexible enough to accommodate work and family responsibilities. Relatedly, with the growing human population and increasing desire to pursue post-secondary education, institutions are expecting future enrollment numbers to exceed their existing physical campus constraints. All of these limitations significantly impact individuals' ability to pursue educational experiences and can have broader effects on city, state, and country. Developing/expanding distributed education opportunities offers one solution for these constraints. Furthermore, Oblinger goes on to explain how distributed education could also be used to capitalize on more lucrative but emerging higher-education market opportunities, such as executive education or education for working adults.[14] Ultimately, one of the main goals of distributed education is to make learning experiences available to *anyone, anytime, anywhere*.

Implementing distributed learning experiences also poses significant challenges. While ongoing IT support, costs for additional resources and faculty, and student accessibility to support resources present only a few minor challenges,[15–17] the major challenge with distributed education lies in its delivery. Ensuring a comparable education is delivered between sites or across delivery platforms is crucial. While institutions are aware of this and its importance, accrediting bodies also enforce these principles. For example, the Liaison Committee of Medical Education (LCME) reviews and certifies (*i.e.*, "accredits") the quality of medical programs located in the United States and Canada, and within its 12 standards for accreditation, Standard 8.7 (Comparability of Education/Assessment) addresses specifically comparability of educational experiences between distributed learning populations. It states, "A medical school ensures that the medical curriculum includes comparable educational experiences and equivalent methods of assessment across all locations within a given course and clerkship to ensure that all medical students achieve the same medical education program [learning] objectives."[18] Other accrediting bodies dictate the needs for similar behaviors about distance education specifically. For example, Standard III (Program Outcomes, Curricula, and Materials) for the Distance Education Accrediting Commission (DEAC) states, "The effective design of program outcomes, curricula, and supplemental materials results in cohesive educational offerings and evaluation methods of student learning that are clearly connected to the stated [learning objectives]." Subheading H

(Examinations and Other Assessments) of this standard states, "Examinations and other assessment techniques provide adequate evidence of the achievement of stated learning [objectives]. The institution implements grading criteria that it uses to evaluate and document student attainment of learning [objectives]."[19] The LCME and DEAC accreditation standards for ensuring comparable education experiences for distributed learning populations, as well as those enforced by many other accrediting agencies, as described by Oblinger,[15] all reference a common foundational element for establishing comparability – learning objectives.

Evidence of comparability for delivering distributed education may not be overly difficult to produce if well-written course learning objectives (LOs) and a blueprint for assessing them across sites have been established and implemented into the course. Well-written LOs are developed with the ABCD method[1–5] and according to SMART criteria.[6–10] If these guidelines are followed and an LO-based blueprint for assessment is properly established (see Appendix B), LOs can serve as fundamental guides for providing comparability between curriculum design, instruction delivery, and assessment for courses offered across distributed sites. As a result, this process serves as a model for evidence-based assessment. It is for these reasons that LOs (and their strategic development, implementation, and assessment in distributed courses) are a crucial component to accrediting agencies for demonstrating comparability of experience and outcomes in courses across distributed campus sites.

Comparable education does not, however, ensure comparable outcomes. Different populations can, and often do, exhibit general differences in demographics, motivations for learning, learning strategies, and other behaviors, each of which can expectedly impact their ability to achieve stated LOs. In fact, empirical evidence shows motivations and learning strategies impact performance.[20,21] Therefore, learning about each population – even each student individually – can help instructors understand and explain population outcomes differences within a comparable educational environment. One way of doing this would be to collect information about each of the population's behaviors that could influence their ability to achieve LOs. In the present study, we collected information about each population's motivations and strategies for learning using the Motivated Strategies for Learning Questionnaire (MSLQ), an 81-question survey tool developed in 1991 by Paul Pintrich to assess six scales of student motivations for learning and nine scales of their use of different learning strategies.[11]

Value, expectancy, and affect are the three general constructs that form the basis of motivation scales.[11] Value components address how much of a student's motivation to learn comes from their desire to learn and master material (intrinsic goal orientation), to earn good grades or the approval of others (extrinsic goal orientation), and to fulfill an importance or usefulness (task value). Expectancy components address how much a student believes that the learning outcomes that will result from a learning experience are contingent on their own efforts alone (control beliefs) and to what degree students expect that they will succeed at academic tasks (self-efficacy for learning and performance). Lastly, affective components address how much students worry about taking examinations and how they believe anxiety affects their academic performance.[11,20]

Learning strategies scales are categorized into two constructs: cognitive and metacognitive strategies, and resource management strategies. Cognitive and metacognitive strategies address how much each student uses common strategies for learning. These strategies (scales) include reviewing information until they have memorized it for the short term (rehearsal), summarizing and making connections between material for the long term (elaboration), prioritizing and outlining information (organization), using existing knowledge to evaluate new information (critical thinking), and practicing control and awareness of one's own level of thoughts, knowledge, and reasoning (metacognitive self-regulation). Resource management strategies address how well each student regulates resources that are helpful for learning. These scales measure how well students regulate their study time and setting (time and study environment), persist through difficult or boring tasks (effort regulation), seek help from other students (peer learning), or pursue assistance from an instructor when needed (help seeking).[11,20]

Other similar socio-cognitive assessment tools have been created for similar purposes; the Learning and Study Strategies Inventory (LASSI) developed in 1987 by Claire Weinstein to assess students' awareness about their use of learning and study strategies (more generally than the MSLQ),[22] the Multidimensional Self-Concept Scale (MSCS) questionnaire developed in 1992 by Bruce Bracken to evaluate six subscales of self-concept,[23] and the Academic Self-regulated Learning Scale (A-SRL-S) established in 2010 by Carlo Mango to measure self-regulation in higher education.[24] Despite each of these tools' advantages, the MSLQ is most closely associated with course-specific performances.[20]

While the MSLQ is found to be reasonable valid and reliable, the biggest criticism it receives is that its "expected" results can vary greatly for students between courses. However, Pintrich acknowledges these variable results and in fact explains they support the socio-cognitive paradigm he used to create the MSLQ. According to this framework, motivations and learning strategies are dynamic, contextually bound, and controlled by the student, and therefore the student will and should express these motivations and learning strategies differently between courses depending on their interest, relative self-efficacy, etc. in the nature of the course.[20] Other reasons for choosing to use the MSLQ include its customizability, wide use, easy (electronic) administration, and extensive validation and reliability studies.[25–27] Its results offers significant feedback that can help students improve their performances and also help instructors mentor students accordingly.

The objective of this study was to employ evidence-based assessment practices in a distributed University of North Dakota OT 422 (Anatomy for Occupational Therapists) course that delivers equivalent and simultaneous learning experiences to populations in two face-to-face sites (a home site in Grand Forks, ND and a satellite site in Casper, WY). The course is unique because each site employs a simultaneous, common lecture time (taught by one instructor live via webcam) and simultaneous yet disparate laboratory experiences (taught by different instructors with different resources). Both student populations received the same multiple-choice question (MCQ) lecture examinations but different yet comparable constructed-response question (CRQ) laboratory examinations. Both populations also received a set of remediation questions on each examination. The purpose of this study is to determine comparability between Grand Forks and Casper population/behavior characteristics and resulting performances with special attention given to the LO-based properties of the similar high stakes MCQ lecture examinations and different (but comparable) high stakes CRQ laboratory examinations. We hypothesize that student motivations, learning strategies, and academic performances in a human anatomy curriculum are comparable across distributed campus sites. If differences in motivations or learning strategies are discovered between sites, they will be examined for resulting impact on site-specific remediation performances. We additionally hypothesize that Grand Forks and Casper students academically struggle more differently with laboratory LOs than lecture LOs since laboratories are conducted autonomously on the campuses whereas lectures are delivered synchronously using distance technology.

Materials and Methods

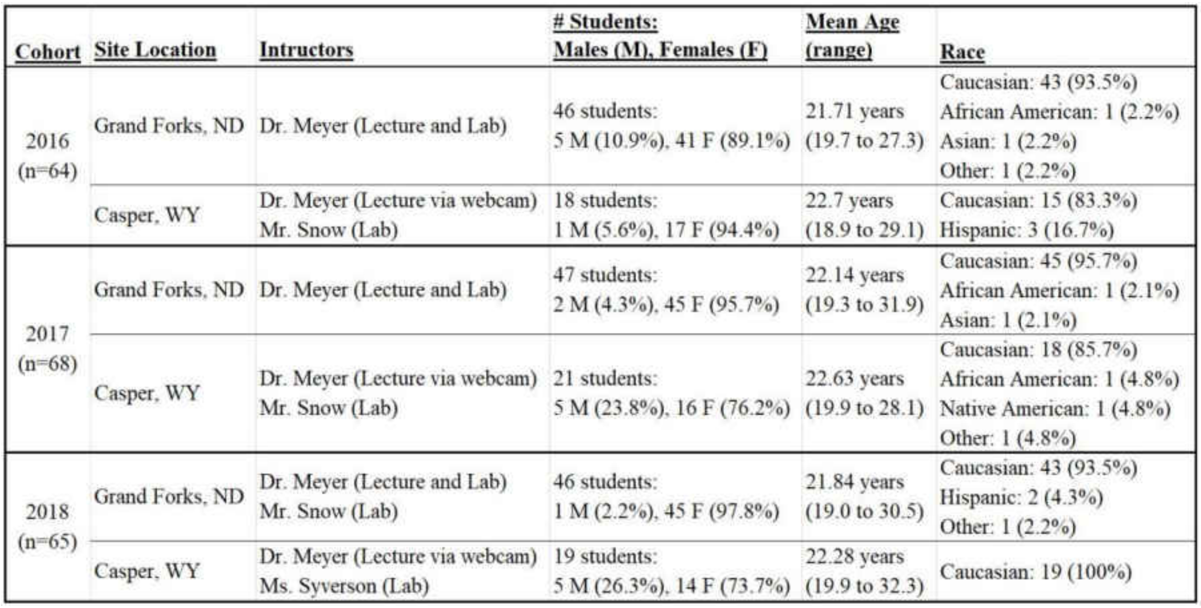
Research Design

A post hoc site comparability (Grand Forks vs. Casper populations) study was conducted for the 2016, 2017, and 2018 cohorts of University of North Dakota's OT 422 course (Anatomy for Occupational Therapists). The independent variables examined for comparability were demographic characteristics, six motivations scales, nine learning strategies scales, remediation performances, study/work time allocation, percentage grades, and letter grades. Remediation performances and percentage grades were also examined as dependent variables to certain motivations and learning strategies scales. While the main results of this study focus on comparative differences between Grand Forks and Casper populations, we also examine percentage grades and LO performance differences between the sites' common lecture and disparate laboratory learning experiences. This study was approved via the procedures of the University of North Dakota (UND) Institutional Review Board (IRB) and ruled as level 4 exempt (IRB-201809-062).

Subjects and Setting

Three cohorts of OT 422 students at the UND School of Medicine and Health Sciences served as subjects (see Table III-1). The majority of the students were female: 58 of 64 (90.6%) in 2016, 61 of 68 (89.7%) in 2017, and 59 of 65 (90.8%) in 2018. Additionally, most students are 21- to 23-year-old Caucasian individuals. OT 422 is a distributed course with one larger population of students at a home site in Grand Forks, ND on UND's campus and the other smaller population of students at a satellite site in Casper, WY on Casper College's campus; on average, the Grand Forks student population consists of 70.6% of each cohort whereas the Casper student population consists of 29.4% of each cohort. Multiple instructors teach the course. Table III-1 shows the distributions of instructors for the two sites as well as demographic differences between the site-specific student populations for the years studied.

Table III-1. Instructors and Student Demographics for Three OT 422 Cohorts. Information is presented as site-specific for comparison. Instructors remained consistent for 2016 and 2017 and changed slightly with the addition of a new Casper adjunct instructor in 2018. Grand Forks to Casper student ratios were 1:2.56 in 2016, 1:2.24 in 2107, and 1:2.42 in 2018. Both student populations were overwhelmingly female students. Additionally, most students were Caucasian individuals between 21 and 23 years old.



Cohort	Site Location	Intructors	# Students: Males (M), Females (F)	Mean Age (range)	Race
2016 (n=64)	Grand Forks, ND	Dr. Meyer (Lecture and Lab)	46 students: 5 M (10.9%), 41 F (89.1%)	21.71 years (19.7 to 27.3)	Caucasian: 43 (93.5%) African American: 1 (2.2%) Asian: 1 (2.2%) Other: 1 (2.2%)
	Casper, WY	Dr. Meyer (Lecture via webcam) Mr. Snow (Lab)	18 students: 1 M (5.6%), 17 F (94.4%)	22.7 years (18.9 to 29.1)	Caucasian: 15 (83.3%) Hispanic: 3 (16.7%)
2017 (n=68)	Grand Forks, ND	Dr. Meyer (Lecture and Lab)	47 students: 2 M (4.3%), 45 F (95.7%)	22.14 years (19.3 to 31.9)	Caucasian: 45 (95.7%) African American: 1 (2.1%) Asian: 1 (2.1%)
	Casper, WY	Dr. Meyer (Lecture via webcam) Mr. Snow (Lab)	21 students: 5 M (23.8%), 16 F (76.2%)	22.63 years (19.9 to 28.1)	Caucasian: 18 (85.7%) African American: 1 (4.8%) Native American: 1 (4.8%) Other: 1 (4.8%)
2018 (n=65)	Grand Forks, ND	Dr. Meyer (Lecture and Lab) Mr. Snow (Lab)	46 students: 1 M (2.2%), 45 F (97.8%)	21.84 years (19.0 to 30.5)	Caucasian: 43 (93.5%) Hispanic: 2 (4.3%) Other: 1 (2.2%)
	Casper, WY	Dr. Meyer (Lecture via webcam) Ms. Syverson (Lab)	19 students: 5 M (26.3%), 14 F (73.7%)	22.28 years (19.9 to 32.3)	Caucasian: 19 (100%)

The students are enrolled in OT 422 during the same summer they begin the Entry-level Master's Program in Occupational Therapy. Students are required to have completed established prerequisite coursework to enter the program. However, some students enter the program having already earned their bachelor's degree or after having left the academic system for some time (*i.e*., a non-traditional student). Additionally, while the majority of students enter the professional program from undergraduate programs at UND, some students apply and transfer from other institutions.

OT 422 is a 5-credit, twelve-week course that meets for a lecture and laboratory five days per week (Monday through Friday) for approximately 15 scheduled contact hours/week. The OT 422 course is delivered through a hybrid of traditional lecturing and active learning. Each class day begins with a PowerPoint lecture presentation (approximately 1 hour in length) covering the anatomical topic of the day followed by a dissection-based, active learning cadaver laboratory (approximately 2 hours in length). The lecture is delivered simultaneously across sites by the home site instructor via live two-way video conferencing and simultaneously subsequent laboratories are delivered by different instructors with different resources and without inter-site collaboration. The laboratory component is student-directed, but instructors and teaching assistants facilitate learning to ensure learning progress and expectations are met. Because of the uniquely distributed character of this course, the instructors deliberately used evidence-based practices for establishing LO-based assessment strategies (see LOs for OT 422 in Appendix A and LO Proportions and Assessment Strategy Tables in Appendix B) that would ensure comparability of instruction delivery and assessment between sites.

Routine Examination Procedure

For each cohort, formal assessment of student learning was conducted using six lecture examinations and six laboratory examinations that determined the majority of the students' final grades. All examination scores were recorded in Blackboard, the course learning management system. Lecture and laboratory examinations were administered in pairs approximately every 8 class days. Each examination tested new material in addition to a set of remediation questions for each student based on poor performances on the previous examination (see Chapter II). The methods of examination reflected the respective instructors, teaching methods, and site resources – both sites were simultaneously administered the same multiple choice question (MCQ) lecture examinations, but each site administered different constructed-response question (CRQ) laboratory examinations. To ensure comparability between different laboratory examinations, LO Proportions and Assessment Strategy Tables (see Appendix B) were used to guide site-specific construction of examinations while ensuring their comparability.

With each examination, students were given a paper answer sheet with numbered answer lines to record their answers to the corresponding questions. An additional line next to each answer line was provided for students to record their level of confidence in their answer to each question. Answer sheets were collected and scored against a pre-established key according to Snow's confidence-based scoring criteria, awarding up to two credit points per question. During the subsequent class day, scored answer sheets were returned to students for review and reflection of their performances. Answer sheets were accompanied by a copy of the examination for lecture, and in laboratory an examination answer key that included the question, an indication of the anatomical structure tagged on the cadaver for that question, and the correct answer. During this examination review time, the 2017 and 2018 cohorts completed a self-assessment exercise (SAE) to identify the six LOs which they believed they performed poorest on in lecture and in laboratory. The answer sheets for all cohorts (with the SAEs in 2017 and 2018) were again collected, organized, and prepared for instructor analysis for determining remediation questions for each student to be included on the subsequent examination (see Chapter II).

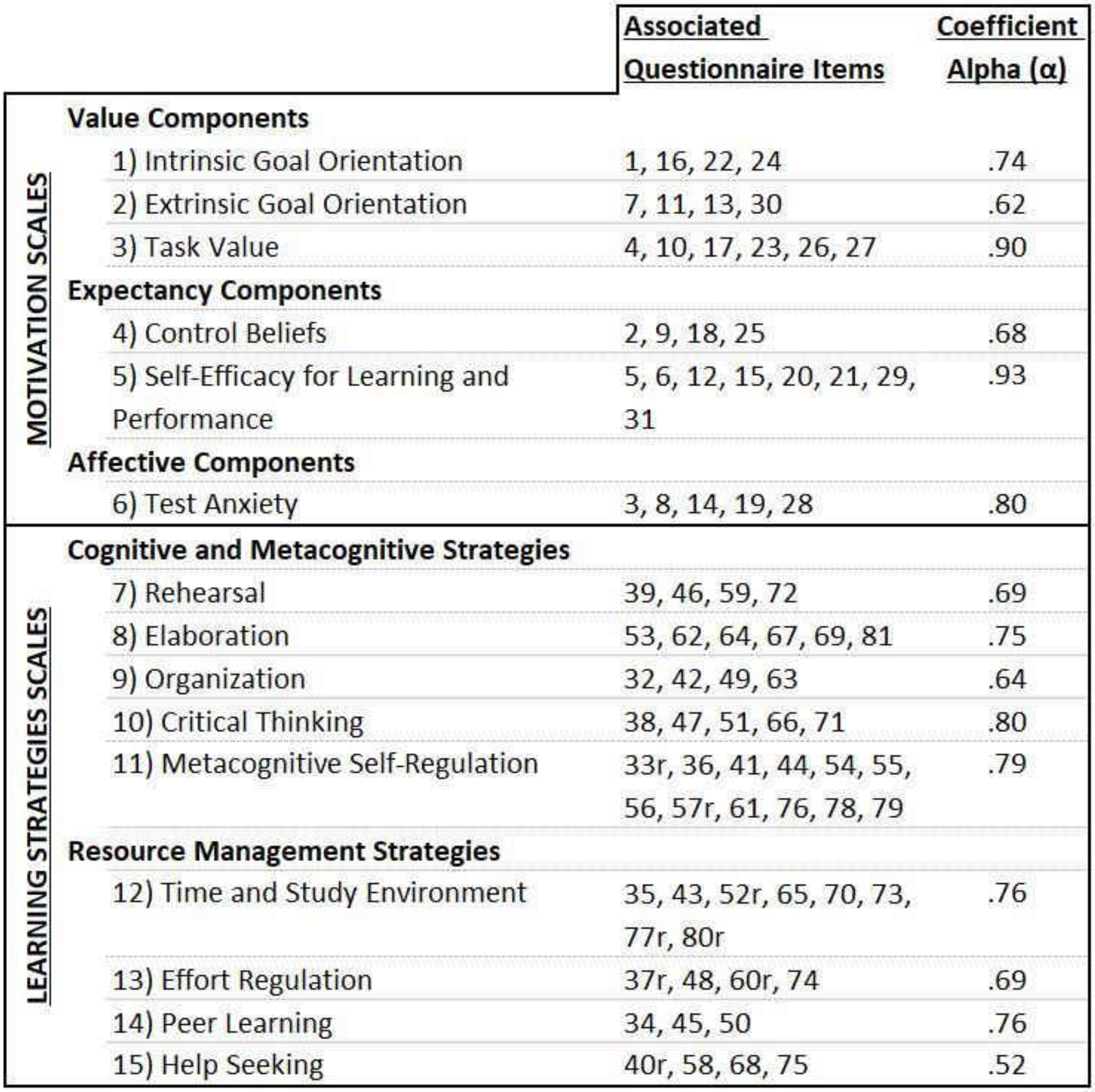
Motivated Strategies for Learning Questionnaire (MSLQ)

The Motivated Strategies for Learning Questionnaire (MSLQ) was administered to each of the 2016, 2017, and 2018 cohorts. Because the MSLQ collects information about course-specific motivations and learning strategies, the MSLQ was administered to these cohorts approximately half-way through the 12-week OT 422 course – a point in the course at which students understand their motivations and learning strategies for the course. Qualtrics survey software was used to administer the survey electronically and collect the results. The survey's 81 questions take students approximately 20-30 minutes to complete. Responses to the 81 MSLQ items are recorded with a 7-level Likert-type scale of agreement (i.e., 1 = not at all true of me and 7 = very true of me).

While other similar socio-cognitive assessment tools have been created for similar purposes, the MSLQ was primarily chosen to be used in this study because of its empirical ties to course-specific performances.[20] The MSLQ was also chosen for its popularity, customizability, extensive validity and reliability studies, and ease of use.[20,21,25–27] Table III-2 shows the mapping of each of the 81 MSLQ items to their respective scales. Also listed for each scale is its accepted coefficient alpha which was refined by Pintrich over 5 years of study and indicates expected consistency between item responses for each scale.[11] Full questionnaire items are listed in Appendix D.

In addition to the MSLQ, a course survey developed by the instructors was administered on the last class day for each cohort, collecting student perceptions of the remediation interventions and other course-related elements.

Table III-2. MSLQ Item Mapping and Coefficient Alphas per Scale.[11] The Motivated Strategies for Learning Questionnaire (MSLQ) assesses 6 motivation scales and 9 learning strategies scales through a series of 81 questions responded to with a 7-level Likert-type response of agreement to self. Some questions items (as noted by "r") are worded negatively (as opposed to the rest being worded positively) and require their numerical answers to be reversed prior to analysis.



		Associated Questionnaire Items	Coefficient Alpha (α)
MOTIVATION SCALES	**Value Components**		
	1) Intrinsic Goal Orientation	1, 16, 22, 24	.74
	2) Extrinsic Goal Orientation	7, 11, 13, 30	.62
	3) Task Value	4, 10, 17, 23, 26, 27	.90
	Expectancy Components		
	4) Control Beliefs	2, 9, 18, 25	.68
	5) Self-Efficacy for Learning and Performance	5, 6, 12, 15, 20, 21, 29, 31	.93
	Affective Components		
	6) Test Anxiety	3, 8, 14, 19, 28	.80
LEARNING STRATEGIES SCALES	**Cognitive and Metacognitive Strategies**		
	7) Rehearsal	39, 46, 59, 72	.69
	8) Elaboration	53, 62, 64, 67, 69, 81	.75
	9) Organization	32, 42, 49, 63	.64
	10) Critical Thinking	38, 47, 51, 66, 71	.80
	11) Metacognitive Self-Regulation	33r, 36, 41, 44, 54, 55, 56, 57r, 61, 76, 78, 79	.79
	Resource Management Strategies		
	12) Time and Study Environment	35, 43, 52r, 65, 70, 73, 77r, 80r	.76
	13) Effort Regulation	37r, 48, 60r, 74	.69
	14) Peer Learning	34, 45, 50	.76
	15) Help Seeking	40r, 58, 68, 75	.52

Data Collection and Analysis

All performance and survey data for the 2016, 2017, and 2018 cohorts were transferred to Microsoft Excel for analysis and anonymized. Student t-tests were used compare data sets. Pearson correlation tests were used to test for correlations between data.

Lastly, Kolmogorov-Smirnov (K-S) tests were used to compare letter grade distributions between populations. Differences were determined to be significant if $p < 0.05$. The following variables were examined:

a) Means in overall motivations and learning strategies scores for Grand Forks and Casper populations were compared. The results of the 31 MSLQ items surveying the 6 motivations scales were combined for each site and compared. Similarly, the results of the 50 MSLQ items surveying the 9 learning strategies scales were combined for each site and compared.

b) Means for each of the 15 MSLQ scales for Grand Forks and Casper populations were compared. The results of the combined questions pertaining to each MSLQ scale were compared between sites.

c) Correlation between MSLQ scales and remediation performances were compared for Grand Forks and Casper site populations. Each population's individual student MSLQ scale scores were tested for correlation with their respective post-remediation performance levels (PLs).

d) Remediation performances were compared for Grand Forks and Casper populations. For each student's 72 remediation questions, mean population-specific pre-remediation (Pre-R) PLs and post-remediation (Post-R) PLs were compared.

e) Hours spent studying vs. hours spent working were compared for Grand Forks and Casper populations.

f) Percentage grades were compared between Grand Forks and Casper populations and letter grades were compared to normal curves. Overall percentages were also compared between lecture and laboratory performances for each population.

g) Retesting frequencies for each of 125 LOs was examined for comparability between Grand Forks and Casper. LOs were examined for differences in retesting frequencies between lecture and laboratory examinations because of their respective similar and different learning environments between the populations.

Results

Motivations and Learning Strategies

Combined MSLQ data collected from the 2016, 2017, and 2018 OT 422 student cohorts show that Casper students exhibit stronger motivations for learning than Grand Forks ($p < 0.001$) but comparable learning strategies (Figure III-1). Both populations exhibited greater motivations for learning than learning strategies ($p < 0.01$). Even so, the differences in these behaviors is not great, likely due to this data representing a combined six separate motivations scales and nine learning strategies scales. Slight but significant differences suggest there could be bigger differences between the populations for certain individual motivations and/or learning strategies scales.

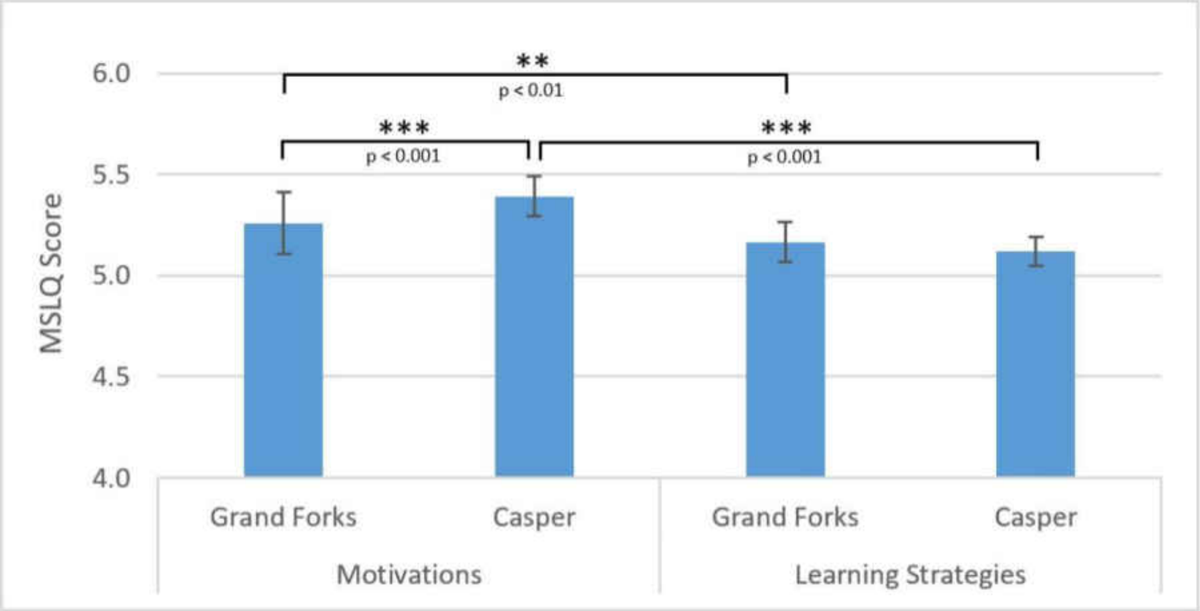


Figure III-1. General Motivations and Learning Strategies. Data is presented as Grand Forks vs. Casper to compare populations. Casper students exhibited slightly greater motivations for learning, but both populations exhibited greater use of motivations for learning than learning strategies. Standard error is also presented. As presented here, general motivations consist of a combined six separate motivations scales, and learning strategies consist of a combined nine separate learning strategies scales. Data presented is from combined MSLQ results from collected from 2016, 2017, and 2018 cohorts. Responses to the 81 MSLQ items are recorded with a 7-level Likert-type scale of agreement to self (*i.e.,* 1 = not at all true of me and 7 = very true of me). Standard error is also presented.

When examining the MSLQ data for population-specific differences in individual scales, Grand Forks and Casper populations exhibit significant differences in two motivations scales. Casper students exhibit greater self-efficacy for learning and performance (p < 0.001) than Grand Forks students, but they also exhibit greater test anxiety (p < 0.01) (see Figure III-2A). Interestingly, these two scales were also the lowest-scoring motivational scales for both populations. Although there were no significant differences between populations for them, Task Value and Control Beliefs were the highest motivational scales.

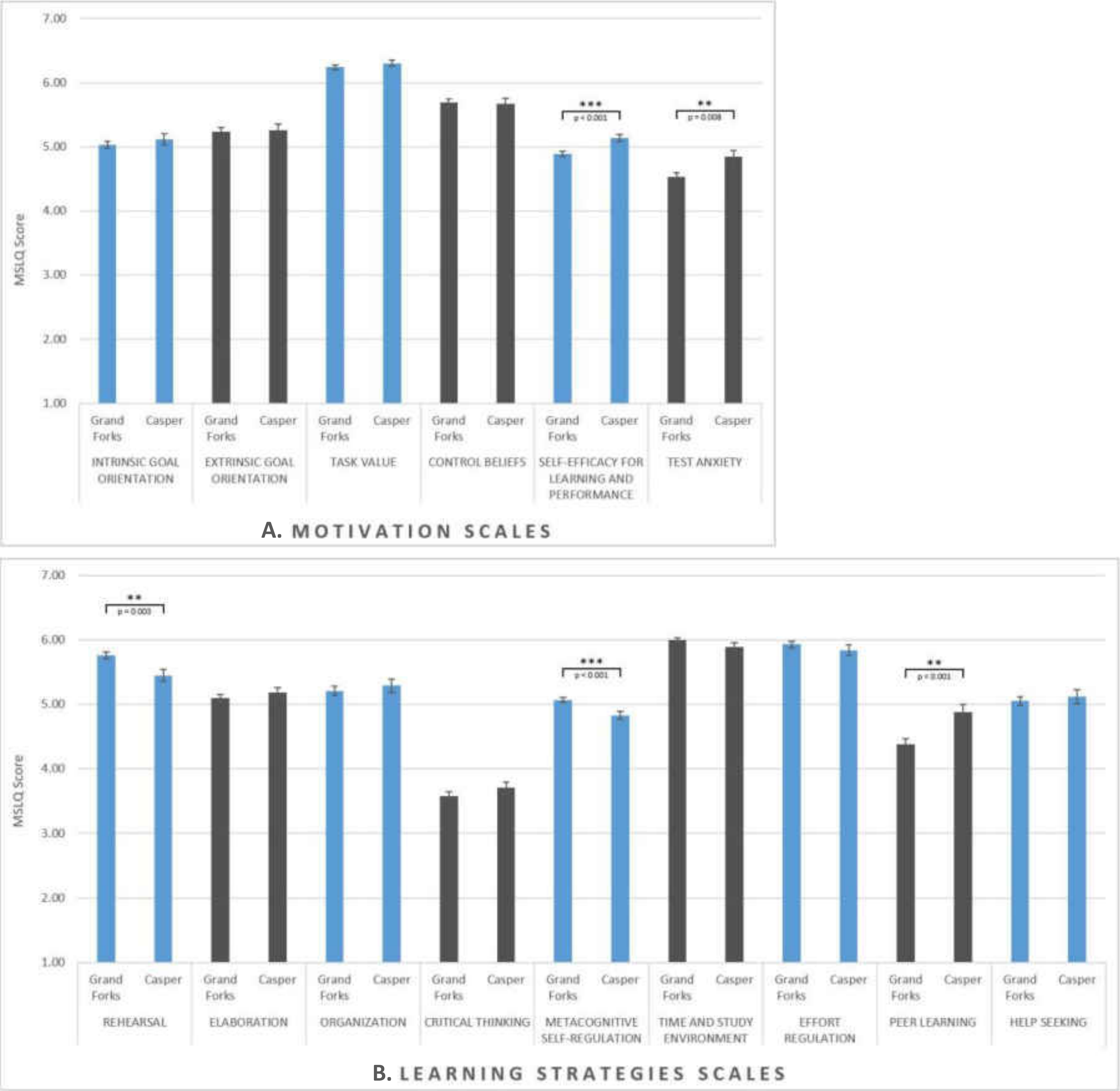


Figure III-2. Individual Motivations and Learning Strategies Scales. Data is presented as Grand Forks vs. Casper to compare populations. Populations differ in two motivational scales (A) and three learning strategies scales (B). Highest and lowest scoring scales should also be noted. Data presented is from combined MSLQ results from collected from 2016, 2017, and 2018 cohorts. Responses to the 81 MSLQ items are recorded with a 7-level Likert-type scale of agreement to self (i.e., 1 = not at all true of me and 7 = very true of me). Standard error is also reported.

Figure III-2B displays population-specific scale comparisons for the nine learning strategies scales. Grand Forks students use rehearsal ($p < 0.01$) and metacognitive self-regulation learning strategies ($p < 0.001$) more than Casper students, but Casper students used peer learning more than Grand Forks students ($p < 0.01$). These scales were among the least-emphasized learning strategies used for both populations, with critical thinking being the least-emphasized of all 15 scales. While the OT 422 instructors emphasize critical thinking, the OT 422 course is very fact- and structure- based. Consequently, this is likely why scales like rehearsal, time and study environment, and effort regulation are higher than the rest.

The motivations and learning strategies scales data reveal important population behaviors that would have otherwise been indistinguishable. Collecting and analyzing this information can influence instruction, curriculum development/changes, assessment strategies, and explain differences in population-specific outcomes. In fact, having implemented a confidence-based assessment strategy generally accepted to be more accurate than any single-dimensional assessment technique, no significant difference in course-wide mean confidence levels (2.47 ± 0.21 vs. 2.44 ± 0.17) or mean correctness levels (80.13 ± 7.71% vs. 80.84 ± 5.52%) respectively could be demonstrated for Grand Forks and Casper populations (this course-wide performance data only from 2016 cohort).

To further understand how OT 422 students' motivations and learning strategies impacted their performance, we compared each sites' students' mean MSLQ scale scores to their respective mean post-remediation (post-R) performance levels (PLs). We were especially interested in determining if increasing motivations and learning strategies were correlated with increasing post-R PLs for student remediation performances since post-R PLs best represent learning through remediation. Figure III-3 (see Appendix E) shows these correlation results for each of the 15 scales for Grand Forks vs. Casper populations.

For the motivations scales (Figure III-3A-F), two scales, Task Value (Figure III-3C) and Control Beliefs (Figure III-3D), were correlated with increasing post-R PLs for both

Grand Forks and Casper populations with small to medium strength. Intrinsic Goal Orientation (Figure III-3A) and Self-Efficacy for Learning and Performance (Figure III-3E) indicated positive correlations as well, but the respective ANOVAs only indicated significant correlations for those associated with the Grand Forks populations. Additionally, Test Anxiety (Figure III-3F) was the only motivations scale to be negatively correlated with increasing post-R PLs ($p < 0.001$ for Grand Forks but $p > 0.05$ for Casper). No correlation or difference was found for or between Grand Forks and Casper populations for the Extrinsic Goal Orientation scale (Figure III-3B).

Only one of the nine learning strategies scales, Time and Study Environment (Figure III-3L) was found to be correlated with increasing post-R PLs for both Grand Forks ($r = 0.19$, $p < 0.05$) and Casper populations ($r = 0.28$, $p < 0.05$). Elaboration (Figure III-3H), Metacognitive Self-Regulation (Figure III-3K), and Effort Regulation (Figure III-3M) were the only other learning strategies that showed correlation to post-R PLs, but only for Grand Forks populations. All of the other learning strategies scales (Rehearsal (Figure III-3G), Organization (Figure III-3I), Critical Thinking (Figure III-3J), Peer Learning (Figure III-3N), and Help Seeking (Figure III-3O)) showed no correlation to post-R PLs for either Grand Forks or Casper populations.

Despite less number of scales, motivations seemed to drive more desirable post-R performances than learning strategies. Notably, the two greatest motivations scales (Task Value and Control Beliefs) and two greatest learning strategies scales (Time and Study Environment, and Effort Regulation), as noted from Figure III-2, all resulted in significant (or approaching significant) correlations to high post-R PLs for both Grand Forks and Casper populations (see Figure III-3C, D, L, and M).

Significance in scale correlations to post-R PLs was found to exist nearly three times more for the Grand Forks population than the Casper population, despite no differences in mean scores between the populations for most of the scales. In contrast to the Grand Forks data, Casper data often suggests weak correlations that do not reach significance (see Figure III-3A, E, F, H, I, K, and M). Furthermore, population differences for the Self-Efficacy for Learning and Performance, Test Anxiety, and Metacognitive Self-Regulation scales were associated with population-specific differences in correlations (see Figure III-3E, F, and K), but the other two scales that reflected differences between populations (Rehearsal and Peer Learning) showed no differences in population-specific correlations (see Figure III-3G and N). This suggests that all scales may not equally or directly influence performance.

Remediation Performances

A confidence-based, individualized remediation intervention was implemented for the 2016 and 2017 cohorts. For this remediation strategy, students reported their level of confidence in their answers to each examination question using a 3-point Likert-type scale (1 = low confidence, 2 = medium confidence, and 3 = high confidence). All examination questions were specifically linked to course learning objectives (LOs). After each examination, each student received feedback regarding the LOs on which they performed poorest based on their confidence in and correctness of their answers. With the following examination, each student was administered six Individualized Remediation Questions (IRQs) retesting each of the six identified LOs as identified by using a pre-determined scale of six possible confidence-based performances ordered by representative levels of knowledge. Students in the 2018 cohort were given Standardized Remediation Questions (SRQs) based on the previous two years' IRQs to compare the impact of individualized remediation strategy to a standardized strategy. Examinations were scored and remediation performances were analyzed. Here, the pre-remediation (pre-R) and post-remediation (post-R) performance levels for the remediation questions are examined for population-specific differences. The following data represents performance data from 72 remediation questions administered to each student.

As previously described, correlations between MSLQ scales and mean post-R PLs were tested since they best represent how motivations and learning strategies can impact post-feedback learning abilities. However, mean pre-R PLs also provide important information about population behaviors. Figure III-4A shows that Casper students had more desirable pre-R PLs than Grand Forks students ($p < 0.05$). Because pre-R PLs are determined from the worst performances from all of the non-remediation examination questions, this means that Casper students displayed overall more desirable PLs on all non-remediation examination questions than Grand Forks students. This is verified by the breakdown of means for individual pre-R PLs; compared to Grand Forks students, Casper students had less frequency of the two least desirable pre-R PLs (I3 and C1) and higher pre-R frequency for three (I1, I2, and C2) of the remaining four more desirable PLs (see Figure III-4B). With this notable difference between overall performance behaviors between Grand Forks and Casper populations, we were especially interested in determining if each population displayed similar behaviors in their post-feedback learning and post-R performances.

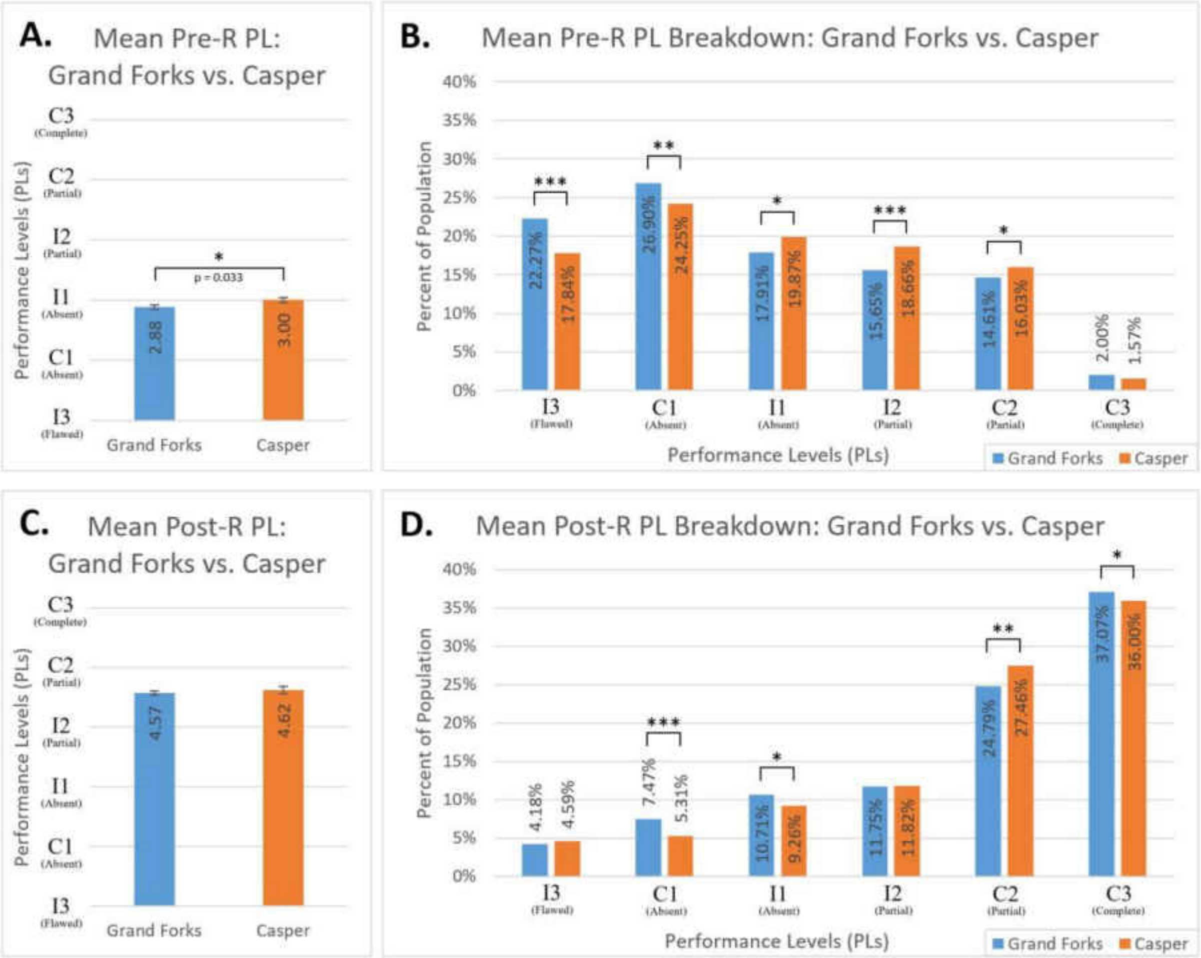


Figure III-4. Mean Pre-R and Post-R PLs and PL Frequencies. Combined data from 2016 and 2017 is presented as percentage of Grand Forks vs. Casper populations to accurately compare population-specific differences in mean pre-R (A) and post-R (C) PLs as well as individual pre-R (B) and post-R (D) PL mean frequencies. Casper students performed better overall (A), resulting in more desirable individual pre-R PLs for them than as displayed by Grand Forks students (B). Although differences are seen between Grand Forks and Casper individual post-R PL mean frequencies (D), post-R performances were found to be comparable between sites (C).

Grand Forks and Casper populations exhibited certain significant differences in individual post-R PL frequencies (see Figure III-4D). Grand Forks students more often remediated to one of two PLs (C1 or I1) indicating absent knowledge. However, Grand Forks students also more often remediated to C3 (complete knowledge) ($p < 0.05$), with Casper students more often remediating to C2 (partial knowledge) ($p < 0.01$). This result verifies that post-R PLs are independent of respective pre-R PLs. Because both populations exhibited comparable mean post-R PLs after first exhibiting different pre-R PLs, this suggests that Grand Forks students achieved greater remediation than Casper students. According to the data, Grand Forks 2016 and 2017 students achieved a mean remediation of 1.82 ± 0.39 PLs, whereas Casper 2016 and 2017 students achieved a slightly less mean remediation of 1.72 ± 0.28 PLs (not shown), however these results were statistically insignificant. A significant difference in population-specific mean pre-R PLs suggests each population may require different learning efforts to reach similar remediation and post-R PLs.

Remediating each of the different knowledge levels – as determined by pre-R PLs – requires different amounts and types of effort. For example, remediating absent knowledge (*i.e.,* C1 or I1 performances) requires a student to learn the relevant material with no pre-existing knowledge on which to draw, whereas remediating flawed knowledge (*i.e.,* I3 performances) requires a student to identify and correct the flaw in his/her existing knowledge. As such, remediating absent knowledge likely requires more time and effort than remediating flawed knowledge, especially having received feedback including confidence-based performance levels, the original questions/answers, and respective LOs. In support of this, we believe how students allocate their time between studying and working – the two greatest time commitments we believe these students manage during OT 422 – may impact their overall and remediation performances and explain any population-specific differences in mean pre-R and post-R PLs. For example, we hypothesize that Grand Forks students displayed poorer mean pre-R PLs because they spent more time working and less time studying. Anticipating these and other possible population-specific differences, we surveyed the 2016, 2017, and 2018 cohorts about their time allocated to working jobs vs. studying for OT 422. The survey results are presented in Figure III-5.

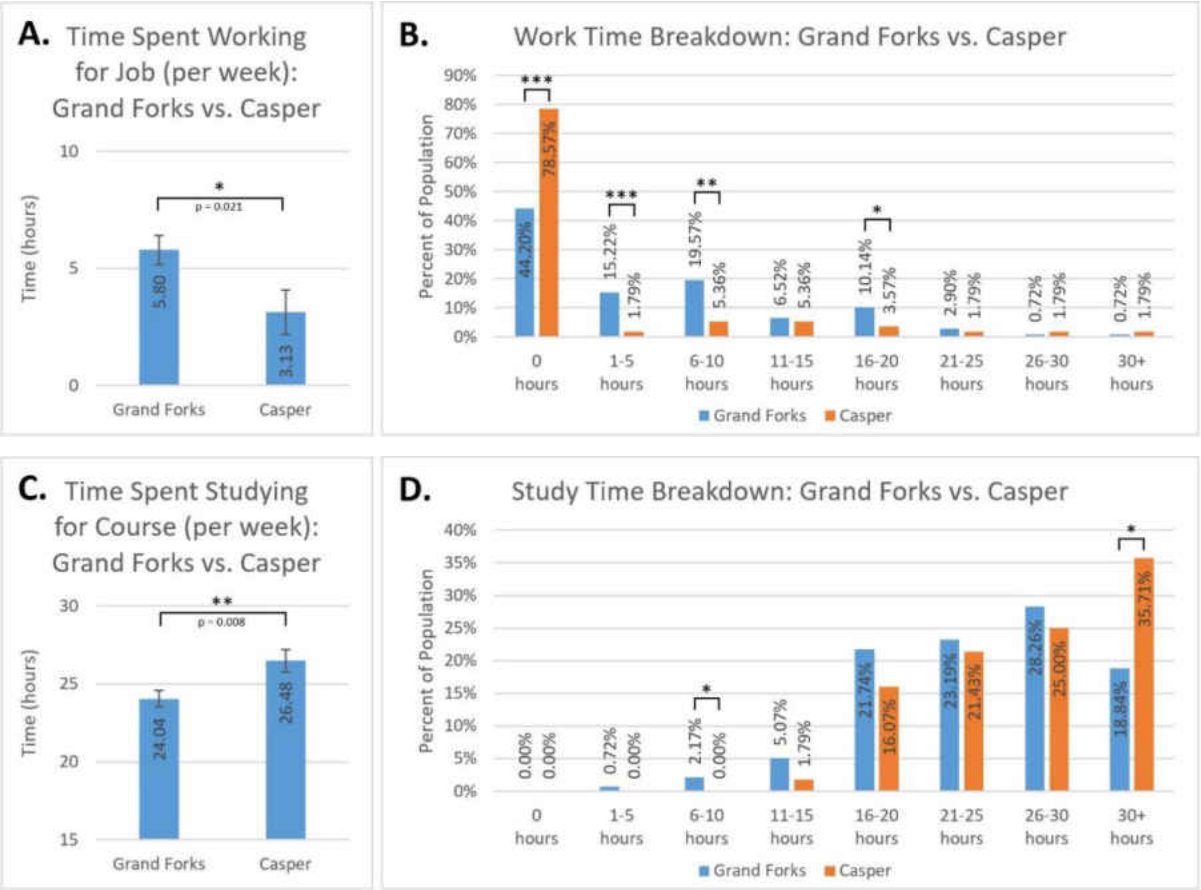


Figure III-5. Student Time Spent Working for Job vs. Studying for OT 422. Combined data from 2016, 2017, and 2018 is presented as mean hours working (A) and studying (C) as well as percentage of Grand Forks vs. Casper populations for each time-range survey option (B and D). Compared to Grand Forks students, Casper students worked less hours at jobs ($p < 0.05$) and spent more time studying for OT 422 ($p < 0.01$).

On average, the majority (78.57%) of Casper students and 44.20% of Grand Forks students did not work at outside jobs during OT 422 (their first summer in the OT program) (see Figure III-5B). Therefore, over 34% more students work in Grand Forks than Casper during the time period in which they are taking OT422. In analyzing the classwide data, those in Grand Forks were found to work approximately 6 hours/week – nearly double the average hours per week than those who worked in Casper ($p < 0.05$) (see Figure III-5A). When comparing the students who worked jobs (55.80% of Grand Forks students vs. 21.43% of Casper students), Grand Forks students worked an average of 10.39 ± 6.82 hours whereas

Casper students worked an average of 14.58 ± 8.07 hours. Therefore, although the overall Grand Forks population worked more hours per week on average at a job in comparison to the overall Casper population, Casper students who are employed work more hours per week than Grand Forks students who are employed.

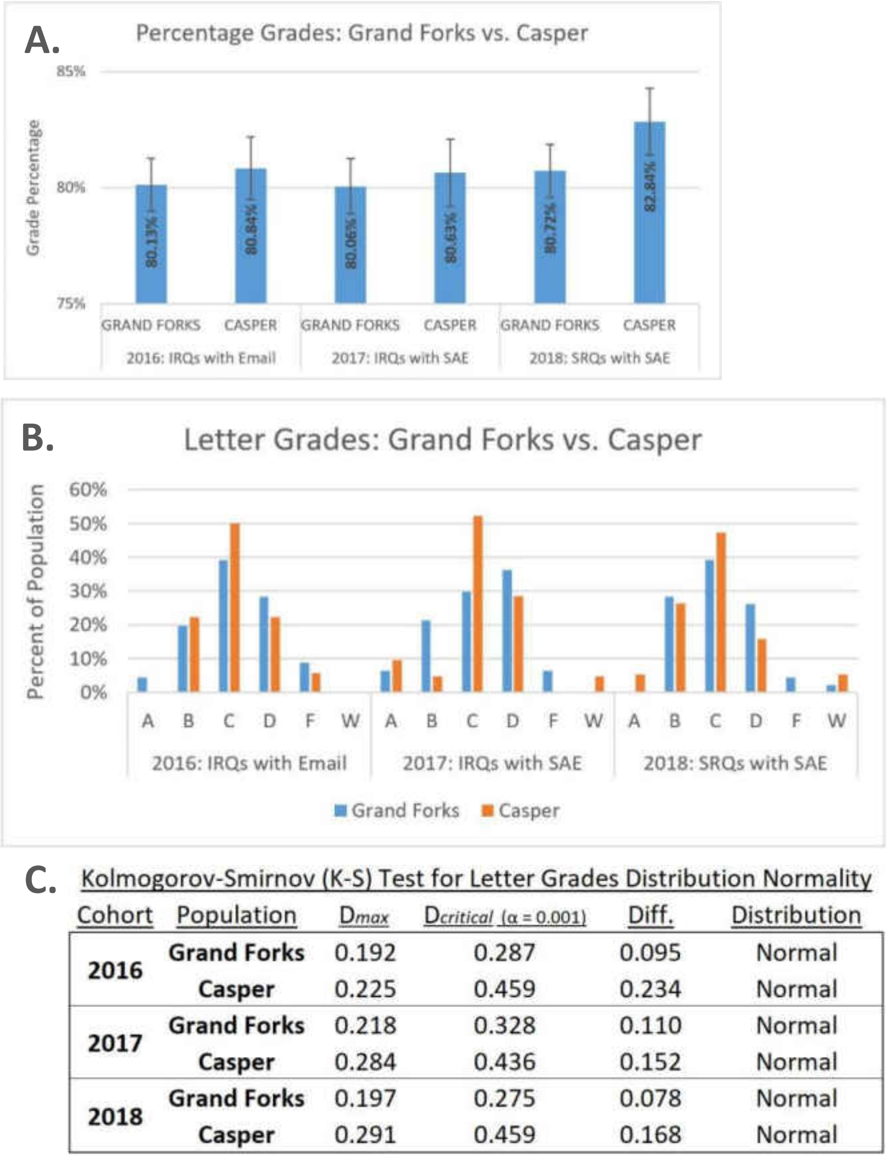
Because available hours in each week are limited, students who work more hours will have less hours left available to study for OT 422. Grand Forks students reported studying approximately 24 hours per week outside of class – about 2.5 hours less per week on average than Casper students ($p < 0.01$) (see Figure III-5C). As significant as this is, it does not suggest that Grand Forks students do not study enough. In fact, nearly all of both populations expectedly indicated that they spent at least 16 hours per week studying for OT 422 (see Figure III-5D). Nonetheless, we believe the reason Grand Forks students study less is that they, on average, worked more hours than Casper students (see Figure III-5 A and B). This may also explain why Grand Forks students displayed less-desirable pre-R PLs than Casper students (see Figure III-4 A and B). However, because the working students in Grand Forks worked less hours per week than those who worked in Casper, they had more time to spend studying for the remediation questions than those who were working more hours in Casper. In fact, data from another survey question asking about hours spent studying for the remediation questions specifically showed that working students in Grand Forks spent slightly more time (1.64 hours vs. 1.31 hours) studying for the remediation questions than working students in Casper. These results suggest one possible explanation for why Grand Forks students performed comparably to Casper students on post-R performances (see Figure III-4 C and D) despite having poorer pre-R performances and requiring more time to successfully remediate pre-R performances (Figure III-4 A and B).

These data allow us to better understand and explain population-specific performance differences, but more data needs to be collected to come to definitive conclusions about how these time allocations impact population-specific pre-R and post-R performances. While a class-wide regression analysis verified a relationship between work hours and study hours (r = 0.16, p < 0.05), the strength of the correlation was small. Furthermore, regression analysis produced insignificant and inconclusive results for site-specific analyses. These results are likely due to the low and/or different number of data points and widely ranging responses.

Grades

While the remediation strategies were based on confidence and correctness, examination scores were recorded based on correctness only. This was necessary to detect guesswork (C1 performances). Correct responses, for which full credit was awarded, are more associated with desirable (*i.e.*, "better") confidence-based performance levels, and incorrect responses, for which no credit was awarded, are more associated with undesirable (*i.e.,* "worse") confidence-based performances. Therefore, the accuracy of assessing confidence-based performances can be reflected in correctness-only scoring.

As presented in Figure III-6, average overall final percentage grades and resulting letter grades earned by the 2016, 2017, and 2018 student cohorts exhibited no population-specific differences or differences between years for Grand Forks and Casper populations. Despite similar overall mean percentages (see Figure III-6A), widely-ranging percentages resulted in a broad distribution of final letter grades (Figure III-6B).



Kolmogorov-Smirnov (K-S) Test for Letter Grades Distribution Normality

Cohort	Population	D_{max}	$D_{critical}$ $(\alpha = 0.001)$	Diff.	Distribution
2016	**Grand Forks**	0.192	0.287	0.095	Normal
	Casper	0.225	0.459	0.234	Normal
2017	**Grand Forks**	0.218	0.328	0.110	Normal
	Casper	0.284	0.436	0.152	Normal
2018	**Grand Forks**	0.197	0.275	0.078	Normal
	Casper	0.291	0.459	0.168	Normal

Figure III-6. Overall Site-Specific Percentage and Letter Grades. Grand Forks and Casper populations exhibited no significant differences between their overall final percentage grades (A) or final letter grades (B) for any of the three years (2016, 2017, and 2018). Only examination-related grades were included in this analysis. The interpretations of final letter grades from percentage grades were made according to the following scale: A ≥ 93.45%; B = 85.45% to 93.44%; C = 77.45% to 85.44%; D = 69.45% to 77.44%; F ≤ 69.44%. W = withdrawal from the course. Each population's letter grades are normally distributed according to K-S test statistics (excluding W grades) (C).

Finding no statistically significant differences, the marked difference in the number of B and C letter grades earned by students in 2017, especially for Casper students stood out from the data (see Figure III-6B). At this time, we have no explanation for why this occurred particularly since the K-S tests suggest the distributions of letter grades between Grand Forks and Casper populations are comparable. Casper population's overall mean percentage grade also stood out as slightly higher than the others despite statistical insignificance. We hypothesize a change in Casper's lab instructor in 2018 contributed to this note.

Grand Forks and Casper students achieved comparable performances on all lecture exams and all lab exams with one inverse exception for the Casper students in 2018 (see Figure III-7). Excluding this abnormality, both populations performed significantly better (3.23% higher on average) in lecture than in lab. Aside from the noted exception, these results reflect expected performance differences between MCQ lecture exams and CRQ lab exams. Lecture and lab exams were comparable in length and content (see Appendix B).

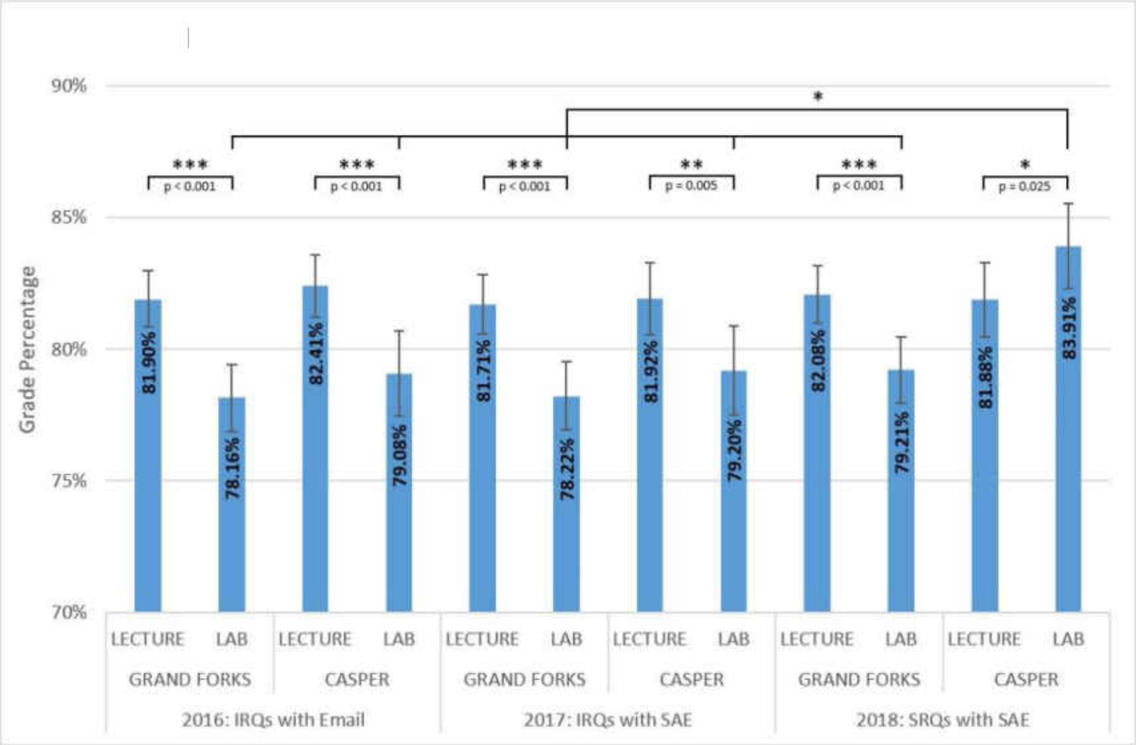


Figure III-7. Overall Site- and Learning Environment- Specific Percentage Grades. Grand Forks and Casper populations performed significantly better on lecture MCQ examinations than laboratory CRQ examinations with one inverse exception for Casper students' abnormally higher laboratory performances in 2018. Both populations performed comparably between all lecture examinations and all other laboratory examinations. Standard error is also presented.

Learning Objectives Retested

Figure III-8 (see Appendix F) shows how often each of the 125 course LOs were individually retested by 2016 and 2017 Grand Forks and Casper populations for lecture (Figure III-8 A and B) and laboratory (Figure III-8 C and D). Grand Forks and Casper populations were found to demonstrate poorer performances on generally the same LOs in both lecture in laboratory ($r = 0.63$ to 0.89, $P < 0.001$). This finding suggests that some LOs – those most often retested due to poor initial performances – may be more difficult for students to master.

Most LOs were retested by both populations each year in both lecture and laboratory settings (see Figure III-8). Despite the common learning environment, instructor, lectures, and examinations for lecture, Grand Forks and Casper students exhibited differences with regard to which LOs caused them greatest difficulty in achieving. Other factors, such as receiving the lectures synchronously but via videoconference (Casper), experiencing different methods of learning/reinforcing the lecture material in similar but disparate laboratory learning environments, or differences in population-specific behaviors such as motivations and learning strategies may explain the observed differences.

LO retesting frequencies were found to be different between examination types (*i.e.*, lecture vs. laboratory) and between site populations (i.e. Grand Forks vs. Casper) (compare Figure III-8 A and B to C and D). These differences support the rationale for using of individualized remediation strategies, especially for distributed campus populations. In the case of OT 422, greater differences for Grand Forks vs. Casper populations in laboratory LO retesting frequencies likely result from different (but comparable) site-specific laboratory examination questions, and therefore reflect the individual needs of each population. These results justify providing an individualized education to individual students and/or distributed student populations.

Discussion and Conclusions

The results of this study of education delivered across multiple campus sites gave us specific insight to the student populations within OT 422 and how they responded to a novel confidence-based, individualized remediation strategy. Given the lack of diversity in the student populations (the vast majority of OT 422 students are 22-year-old Caucasian females), this study did not investigate minority populations (*i.e.*, males, older or younger populations, or those of non-Caucasian race) compared to the majority. However, while the literature shows that demographic characteristics – even slight ones – affect student performances, low numbers of diverse students in one population likely would have yielded inconclusive results.

Aside from demographic characteristics, the present study identified population-specific behavioral characteristics. Why a student pursues a learning experience, what the student wants to get out of the experience, how interested the student is in learning the subject matter, *etc.* (*i.e.,* motivations), in addition to responding to different instructional/assessment methods, time constraints, course difficulty levels, etc. (i.e. learning strategies), all influence MSLQ results. Accordingly, each student may exhibit very different scores on the same scales for the same course depending on what his/her own motivations and learning strategies are for the given experience. Although greater motivations and learning strategies are associated with better performance,[20,21] a “one size fits all” expectation for student motivations and learning strategies is inappropriate. Rather, promoting how effectively students employ *their own* motivations and learning strategies leads to better outcomes and justifies the implementation of individualized but comparable learning experiences.

Because MSLQ results are specific to the course in which it is administered, we should not expect these same OT 422 student cohorts to exhibit the same mean MSLQ scale scores in different courses. Population-wide MSLQ results explain student population behaviors and help course instructors better understand course quality and outcomes. For example, high *rehearsal* scores and low critical thinking scores indicate the OT 422 course requires the acquisition of basic, factual knowledge. Unexpectedly high or low scale scores would inform the instructor, who then could use that information to make changes in the course with the intent to better support students' use of that scale to increase their learning and performances.

We acknowledge that certain results presented are informative but not empirically conclusive. The data presented about time allocation to working vs. studying, although important and informative, cannot be definitively causally associated with remediation performances. Although we hypothesize a relationship, it is impossible to distinguish the number of hours specifically spent studying for remediation questions from "hours spent studying for the course," as reflected in the course survey. Any perception-based survey response confined to a single or limited number of survey questions must reasonably be questioned for reliability. Unlike the MSLQ scales which survey using multiple questions on a given topic to achieve reliability (see coefficient alphas in Table III-2), the time allocation data were collected from one survey question, and the response options limited the reliability of that data by listing ranges of time as response options. As a result, the *actual* work and study hours students spend are comparable to those they reported in response to survey questions is unknown. Examining how much time students allocate to working jobs highlights one of many possible external factors that can take time away from studying and result in decreased learning and/or performance. Considering these limitations, we can only speculate on their association.

The principal issue this chapter addresses is the challenge associated with delivering a comparable education across multiple campus sites. Conversations about how to deliver comparable learning experiences naturally lead to discussing how to ensure comparable outcomes; however, ensuring comparable delivery of educational experiences across distributed campuses does not guarantee comparable outcomes. Distributed campuses exists to reach different populations of students, and as the present study demonstrates, different populations of students have different learning needs. Therefore, expecting two different student populations to exhibit the comparable outcomes without considering the characteristics of individuals as well as collectively to the group is naive. Rather an important consideration in delivering comparable learning experiences is the ability to explain differences in outcomes between populations.

Alternatively, ensuring comparable outcomes between populations likely involves compromising comparability in learning experiences if the learning needs of the individual are given a higher degree of consideration. Delivering comparable education or ensuring comparable outcomes among multiple distributed populations poses a challenge since instructors cannot control how students will learn, and different populations learn and perform differently despite comparable learning experiences, as shown in the present study. However, we have shown how learning and performance outcome differences resulting from standardized and comparable education between distributed populations can be made more comparable with individualized yet comparable learning experiences.

Our data support these claims. Student populations and their performances were different, yet outcomes on the basis of remediation resulted in comparable student learning and final grades. These outcomes support the novel ordering and use of confidence-based performances for remediation strategies. The analysis of lecture vs. laboratory LO retesting frequencies particularly supports the need for a comparable but individualized education. Furthermore, although this chapter focused on population-wide differences in learning behaviors and outcomes, the interventions were administered at the student level – specific to each individual student's performances and learning – and population-wide conclusions do not translate to individual students. Accordingly, comparability and outcome differences would be better explained through comprehensive comparisons of individual students to one another.

Although providing comparable education to distributed populations poses significant challenges to institutions, comparability studies between distributed populations regarding confidence-based assessments, remediation efforts, or even individualized examinations are rare. One of the closest in similarity to the one presented in this dissertation is a study done by David Pike *et. al.* that showed how a competency based curriculum grounded in the Keller method resulted in no significant differences in mean module examination scores or final course grades for two distributed populations at Texas Tech University.[28] Pike even used an anatomy course as one of the focus populations. The course was taught simultaneously between a home site in Amarillo, TX and a satellite site in Abilene, TX. Additionally, more students attended the home site, the majority of students were white (and one site included significantly more females), and faculty were located at both sites despite the lectures being live broadcasted from the home site to the satellite site via two-way video conferencing. Pike also presented how the students in the two populations were roughly equivalent in education upon entry to the program. Although the methodology differed in some ways, Pike's study supports how a strategic, competency-based assessment of learning can help create comparable learning experiences and outcomes for populations across distributed campuses.

Other studies have presented other related aspects of delivering comparable education to distributed campuses. Chris Lovato *et. al.* completed a comparability study of two preclinical training courses in the University of British Columbia's undergraduate medical program that is distributed across three separate sites.[29] After analyzing student examination scores, survey responses, and tutor performance, regarding the two focus courses, Lovato found that the three student populations' performances and perceptions were comparable across the distributed campus sites. More recently, Jonathan Tummons *et. al.* examined specifically how technology (cameras, microphones, and other related communication technologies) can be used to effectively deliver comparable educational experiences to two medical student populations across distributed campuses for one Canadian university.[30] Lovato also indicated the importance of communications technology in his work, and even presented student satisfaction with both audio and visual elements of the communications technology employed in the courses they studied as part of his comparability results.[29]

While the study presented in this dissertation generates additional questions, we believe it also presented a new outlook and understanding of distributed populations. While certain findings support what is already known, comparing population performance and outcomes as a result of our remediation strategy is novel. A strategic LO-based approach to administering an individualized confidence-based remediation strategy ensured comparability in delivering a professional education to Grand Forks and Casper occupational therapy student populations by recognizing and valuing student diversity, and by doing so, lessening the impact of diversity on achievement of learning outcomes.

References

1. Kissel, H., Miller, B. J. & Young, H. *Writing Objectives*. (James Madison University).

2. Williams, B. *Writing Objectives*. (Penn State University).

3. Dalto, J. ABCD: The Four Parts of a Learning Objective. (2013). Available at: https://www.convergencetraining.com/blog/abcd-the-four-parts-of-a-learning-objective.

4. University of Connecticut. Writing Learning Objectives. (2014). Available at: https://kb.ecampus.uconn.edu/2014/07/31/writing-cognitive-objectives/.

5. Mager, R. F. *Preparing Instructional Objectives*. (Fearon Publishers, 1962).

6. Dalto, J. How to Write SMART Learning Objectives. (2013). Available at: https://www.convergencetraining.com/blog/how-to-write-smart-learning-objectives.

7. White, R. ABCD's of SMART Objectives. *Louisiana State University* Available at: https://www.slideshare.net/bwhitelsu/abcds-of-smart-objectives. (Accessed: 19th March 2017)

8. Shabatura, J. Using Bloom's Taxonomy to Write Effective Learning Objectives. *Teaching Innovation & Pedagogical Support* (2018). Available at: https://tips.uark.edu/using-blooms-taxonomy/. (Accessed: 16th March 2019)

9. International Assembly for Collegiate Business Education (IACBE). Bloom's Taxonomy of Educational Objectives and Writing Intended Learning Outcomes Statements. (2016).

10. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. (David McKay Company, Inc., 1956).

11. Pintrich, P. R., Smith, D. A., Garcia, T. & McKeachie, W. J. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). (1991).

12. Matheos, K. & Archer, W. From distance education to distributed learning surviving and thriving. (2004).

13. Harvey, L. Analytic Quality Glossary. Available at: http://www.qualityresearchinternational.com/glossary/.

14. Oblinger, D. & Kidwell, J. Distance Learning: Are We Being Realistic? *Educ. Rev.* **35**, 30–34 (2000).

15. Oblinger, D., Barone, C. & Hawkins, B. *Distributed Education and Its Challanges: An Overview*. (American Council on Education, 2001).

16. Brian Hawkins. Distributed Learning and Institutional Restructuring. *Educom Rev.* **34**, (1999).

17. McGee, P., Carmean, C. & Jafari, A. Distributed Learning: Making Systems that Work. in *EdMedia + Innovate Learning - World Conference on Educational Multimedia, Hypermedia & Telecommunications* (eds. C. Montgomerie & J. Seale) **2007**, 1360–1364 (Association for the Advancement of Computing in Education (AACE), 2007).

18. Liaison Committee on Medical Education. Functions and structure of a medical school. (2019). Available at: http://lcme.org/publications/#Standards.

19. Commission, D. E. A. Accredidation Handbook. Available at: https://www.deac.org/UploadedDocuments/Handbook/DEAC_Accreditation_Handbook.pdf.

20. Duncan, T. G. & McKeachie, W. J. The Making of the Motivated Strategies for

Learning Questionnaire. *Educ. Psychol.* **40**, 117–128 (2005).

21. Robbins, S. B. *et al.* Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychol. Bull.* **130**, 261–288 (2004).

22. Weinstein, C. E., Palmer, D. & Schulte, A. C. Learning and Study Strategies Inventory (LASSI). *Clear. FL H H Publ.* (1987).

23. Bracken, B. A. & Pro-Ed (Firm). *MSCS : Multidimensional Self Concept Scale*. (Pro-Ed, 1992).

24. Mango, C. Assessing and Developing Self-regulated Learning. in *The Assessment Handbook* 26–42 (Philippine Educational Measurement and Evaluation Association, 2009).

25. ERTURAN İLKER, G., ARSLAN, Y. & DEMİRHAN, G. A Validity and Reliability Study of the Motivated Strategies for Learning Questionnaire. *Educ. Sci. Theory Pract.* **14**, (2014).

26. Credé, M. & Phillips, L. A. A meta-analytic review of the Motivated Strategies for Learning Questionnaire. *Learn. Individ. Differ.* **21**, 337–346 (2011).

27. Pintrich, P. R., Smith, D. A. F., Garcia, T. & Mckeachie, W. J. Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educ. Psychol. Meas.* **53**, 801–813 (1993).

28. Fike, D. S., McCall, K. L., Raehl, C. L., Smith, Q. R. & Lockman, P. R. Achieving Equivalent Academic Performance Between Campuses Using a Distributed Education Model. *Am. J. Pharm. Educ.* **73**, 88 (2009).

29. Lovato, C. & Murphy, C. Comparability of student performance and experiences in UBC's distributed MD undergraduate program: The first 2 years of implementation.

BC Med. J. **50**, 380–383 (2008).

30. Tummons, J., Fournier, C., Kits, O. & MacLeod, A. Using technology to accomplish comparability of provision in distributed medical education in Canada: an actor–network theory ethnography. *Stud. High. Educ.* **43**, 1912–1922 (2018).

CHAPTER IV

SUMMARY

Discussion and Conclusions

We established the theoretical framework for this dissertation in Chapter I and laid the foundation for evidence-based assessment and how it drives and measures learning. We then presented two essential components for measuring knowledge, information trueness and belief justification, which show how correctness-only "number correct" assessments are deficient in their ability to accurately measure all knowledge levels and certain performances such as guesswork and misinformation. Multi-dimensional assessments such as confidence-based assessments can accurately detect all levels of knowledge. While accurately assessing knowledge and performance is crucial in any learning experience, we hypothesized how it is particularly important for documenting comparability of education between distributed campus sites, and how tools like the MSLQ can help to explain population performance outcome differences. All of these concepts were used to build the theoretical and experimental framework for this project.

In Chapter II, we showed how a confidence-based, individualized remediation strategy increased student learning. This conclusion was supported by achievement of positive remediation, *i.e.,* reaching higher knowledge levels, of poorest performances in addition to increased LO achievement. As a result of novel ordering sequencing of the six confidence-based performances based on knowledge level, the resulting mean positive remediation put students who retested guesswork (C1) performances into better but incorrect performance levels and resulted, for some students, in a lower letter grade. Although at first glance this may seem punitive, we consider the decrease in grade a positive result to the remediation strategy due to its ability to detect and remediate guesswork performances – something correctness-only “number correct” assessment methods are unable to do.

The successful, beneficial use of the novel confidence-based, individualized remediation strategy shown in Chapter II presents many important principles that can be reasonably adopted into practice, but as it currently stands, implementing the entire confidence-based, individualized remediation project in whole requires an unreasonable amount of time to manage huge data sets and individualized analyses. Principles that can be readily adopted into any course include applying evidence-based principles to writing proper learning objectives, establishing LO proportions and assessment strategies, and using them to build course materials, guide assessment development, and drive student learning. These principles are the underpinnings for application of advanced assessment methods, such as confidence-based assessments, in order to give better meaning and purpose to assessment results. Lastly, applying strategies of standardized remediation is worthwhile and more practical than an individualized approach, even though our data demonstrate how an individualized remediation approach is better for student learning and LO achievement.

Students who self-assess their own academic performance demonstrated no difference in LO achievement via remediation in comparison to those who receive instructor feedback. While SAEs were not found to increase remediation in comparison to instructor-based email feedback, pre-R and post-R performances were notably lower in 2017 (IRQs with SAE). Given comparability between population motivations and learning strategies, we conclude lower performances in 2017 were caused *indirectly* by the SAEs; since students misjudged poorest performances in their SAEs, they likely prepared to retest LOs that differed from those chosen by their instructor for their next IRQs. This was not the case for students in 2016, when the instructor's formative feedback email each student received identified their poorest performances and the LOs on which they would be subsequently retested. In hindsight, a cohort of students that received no formative feedback could have served as a control group for studying the effectiveness of formative feedback on remediation outcomes. Unlike the remediation strategy, we are unable to factor out the feedback intervention to see its direct effect on students in the same year. Despite our insignificant findings regarding the impact of student self-assessments, based on the positive effect of SAEs reported in the literature[1–3], SAEs can be designed to impact learning comparably to or even better than formative feedback provided by instructors. Clearly, providing formative feedback to students enhances student learning.

Finally, student motivation, learning strategies, and academic performances in OT 422 curriculum were comparable across distributed campus sites. Even though our methods demonstrated differences in motivations, learning strategies, and performances between Grand Forks and Casper populations, the populations did not demonstrate sufficient diversity to support a conclusion that the populations were different in comparison to one another, and any differences that were found to exist did not affect overall learning outcomes and final grade distributions. However, results of this study were significant to our understanding of the populations in the study and how they were affected by the remediation interventions.

Discovering the differences within the populations regarding course-specific student motivations, learning strategies, and performances was valuable for the instructors to better understand their course and student needs. This principle, including the methods we employed, could be easily implemented into any course and the results readily analyzed. The advantages of the MSLQ include its ability to highlight differences in motivations and learning strategies between students/populations specific to the course in which it was administered.[4] In comparison to population-specific motivations and learning strategies characteristics, one of the most important comparison results we obtained from our study was the differences in percentage grades and LO performances found between each population's lecture and laboratory learning experiences. These results were expected given the nature of the respective learning environments and assessment methods employed, but these analyses informed the instructors of the magnitude of the differences. Being better informed about a course and its students is one of the most important principles to be taken from the comparability study, as this level of understanding is critical for making data-informed changes to enhance student learning and course quality.

The results presented in Chapter III were also important for justifying our efforts in providing an individualized (but comparable) LO-based education to the Grand Forks and Casper populations. Accordingly, the methods employed for ensuring a comparable delivery of education to populations across multiple campus sites are likely the most important principles that can be taken away from Chapter III. While the results presented are important for better understanding the uniqueness of the distributed OT 422 course, the methods we employed to ensure comparability can be used in any course offered across distributed sites and with varied delivery methods. Similarly, these methods could be used to ensure comparability of the same courses taught between different institutions as we have shown how courses can be taught separately and differently but still comparably based on a common set of learning objectives.

Project Limitations

The biggest limitation we faced with this project was the management and analysis of extensive data sets and complex methods, particularly in conjunction with the administration of paper-based examinations. Completing individual analyses (separately for lecture and laboratory) for each student at each site and preparing individualized exams within 8 days was necessary to carry out the experimental design of the study but impractical for implementation in a normal classroom or teaching laboratory setting. Gardner-Medwin and others have created CBA-capable software for administering CBAs[5], but they are incapable of supporting the remediation strategy we employed. Our intervention could be easier managed on a smaller scale (*i.e.*, for less than 20 students). This would, however, compromise statistical power.

LOs were chosen to be remediated in this study based on poor performance on only one examination question testing that LO. In reality, as shown in the LO proportions and assessment strategy tables (see Appendix B), multiple questions were used to test each LO. Because of this, a better justification for remediation would be on the basis of a student's collective performance on all examination questions testing that LO not just one question. In this manner, our strategy likely retested LOs despite students' otherwise acceptable performance on other questions testing the same LO.

Despite the rigorous mapping of LOs to examination items and blueprinting of examinations, we did not focus specifically in this study on proper item writing technique or post-examination item analysis. For example, although lecture examinations were the same for each population, an inconsistent number of answer options characterized MCQ examination items. Additionally, difficulty indices between items – even those testing the same LO – were inconsistent. Although LO-based guides for each site's laboratory examination question development were utilized, question topics were comparable but there was no data on comparability of difficulty level. We believe these limitations had an impact on how LOs were selected for remediation at both campus sites.

We must also acknowledge the limitation of studying human subjects in research. So many factors not addressed or accounted for could influence results in unexpected and undetected ways, no matter how much care and attention is given to accounting for all variables. Even though we acknowledge this limitation, we are confident we have accounted for those that had the greatest potential of influencing outcomes of the study.

Future Directions

Despite the limitations, we plan to apply the findings of this project to additional large-scale courses to advance of the project and refine the methods. To do this, we will develop a web-based software application to facilitate item writing, automate data collection, conduct performance analysis, and create and administer IRQs with examinations. This is a costly entrepreneurial venture requiring specialized expertise and experience. Even so, we see value for continuing this study in this way to enhance the value it adds to education. Having a more efficient manner for managing and analyzing the large data sets would also allow us to overcome the limitation of retesting LOs based on only one question's performance and instead retesting on overall performance.

Further study about the accuracy and consistency of confidence is intended. Understanding how confidence is interpreted by each student is important for fully understanding student performances. Additionally, it is important to address any inconsistencies, misconceptions, and even unintended errors in recording accurate confidence levels. For example, it would be worth identifying if and when any student recorded high confidence for being "highly confident that they do not know the answer" instead of low confidence for being "not confident in the correct answer." Additionally, an advanced learner with well-developed metacognition, understanding that they would be unaware of any misconceptions they *could have*, may believe it would be unwise of them to *ever* record high confidence. Albeit rare given the guidance for choosing confidence levels, this would indicate that a student with the best metacognition may only ever display low-to-medium confidence in their responses. While these and other related isolated occurrences are expected to be rare, implementing a tool or method to detect them and normalize them for accurate student-to-student comparisons could provide a better insight for understanding student metacognitive thought processes.

Another issue to further explore is the proper uses of "*confidence*" in comparison to "*certainty*" in conjunction with correctness to best identify knowledge levels in multidimensional assessments. While some educator-scholars use these terms interchangeably, these two terms, while similar, have different meanings/uses.[6] While the definition of each is similar and they are often interchanged, we interpret the difference in these terms to reflect timing. "Confidence" appertains to a belief in one's ability *to be correct*, whereas "certainty" pertains to a belief *in the correctness* of someone/something. Accordingly, certainty is determined situationally in accordance with information currently available at hand, whereas confidence is established beforehand independent of the current situation. Therefore, according to their definitions, the "confidence" levels we collected in this study to determine performance and knowledge levels were, in reality, “certainty” levels. We do not believe this impacted the results of our study; our use of a *technically incorrect* term did not change the meaning of the data we were collecting. In fact, Gardner-Medwin, the investigator credited with the establishment of Confidence-Based Assessments, used the term “confidence” similarly when he established CBAs. However, he subsequently acknowledged the difference between “*confidence*” and “*certainty*”, and now uses the term “Certainty-Based Assessments.”[7]

While these two terms are obviously quite similar and often do correlate with one another, separating them according to their actual meanings could present some important advantages. For example, a student can be highly confident in their ability to answer a question about an LO beforehand, but then when presented with questions testing that LO they can exhibit low certainty. This student would accurately represent an *over confident* student. *Over confident* and *under confident* are often terms used incorrectly though, as they are often derived from comparing at-hand “confidence” (certainty) levels in comparison to response correctness. For example, students who select an incorrect answer option with high confidence are thought to be over confident. This is an invalid conclusion though, as misinformation/flawed knowledge (as interpreted by I3 performances) and being over-confident are two completely different things. Confidence can only be validated by certainty, and certainty can only be validated by correctness. Therefore, *over-confidence* and *under-confidence* are terms that can only be accurately determined from initial confidence compared to subsequent certainty; they cannot be determined from comparing initial confidence or subsequent certainty to correctness. Therefore, over- and under-confidence as behaviors should not directly affect grades since they are only reflections of past experiences and current expectations.[8] If these two behaviors are not compared to each other, all we can technically detect is which students are *more confident* or *more certain* and *more correct* than others, as we showed in this study, in addition to being able to identify students' performance and knowledge levels.

Therefore, we propose to take this project to the next level and collect data on confidence, certainty, and correctness to compare over/under-confident student behaviors to resulting performance levels and take both into consideration for different remediation methods. Studies have shown that over confidence is more associated with certain personality characteristics, males, students holding lower GPAs, and students primarily enrolled in lower-division courses.[9] We believe this data could enhance our intended future studies as well as add to existing knowledge.

We suspect using and teaching the difference between confidence and certainty would help students parse their understanding of confidence in believed correctness from interpersonal (self-) confidence – the belief in one's own abilities and/or qualities to succeed – and display more accurate confidence levels.[10,11] However, substantial changes for data collection, management, and analysis would need to be made to do this successfully and in a practical, efficient way.

The principles of confidence/certainty-based assessments in remediation could be applied to many related variables. Studies of question response times have shown how fast an answer is chosen/constructed reflects on overall knowledge and behavior.[12–14] Based on this and the results presented in this study, discovering how response timeliness correlates with different populations, learning environments, remediation, general pre-R/post-R performance behaviors, confidence/certainty, correctness, motivations, learning strategies, and demographic characteristics would better explain student learning and performance achievements. Other relevant variables pertaining to different parts of this study, such as the methods of intervention, multi-campus comparability, or feedback, could also be further studied. For example, we would be interested in implementing the remediation intervention from this project into online distance educational experiences and comparing outcomes to those collected from face-to-face learning in the present study, and we would be interested in seeing how our LO-based course/assessment development efforts could make same courses taught by different institutions more comparable. The MSLQ could be used to compare a single population's motivations and learning strategies for different course subjects or even the same subjects at different difficulty levels. In fact, some of this data already exists and could help formulate informed hypotheses.[4]

The possibilities for expanding on this project and looking at additional variables are numerous. Any additional study relative to the outcomes of this study would increase the likelihood of enhancing student learning and identifying predictive indicators for student readiness and likelihood for success in the educational experiences.

References

1. Mcmillan, J. H. & Hearn, J. Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement What Is Student Self-Assessment?

2. Andrade, H. & Du, Y. Student responses to criteria-referenced self-assessment. *Assess. Eval. High. Educ.* **32**, 159–181 (2007).

3. Andrade, H. & Valtcheva, A. Promoting Learning and Achievement Through Self-Assessment. *Theory Pract.* **48**, 12–19 (2009).

4. Duncan, T. G. & McKeachie, W. J. The Making of the Motivated Strategies for Learning Questionnaire. *Educ. Psychol.* **40**, 117–128 (2005).

5. Gardner-Medwin, A. R. & Gahan, M. Formative and Summative Confidence-Based Assessment. 147–155 (2003).

6. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).

7. Gardner-Medwin, A. & Curtin, N. Certainty-Based Marking (CBM) for reflective learning and proper knowledge assessment. in *Re-Engineering Assessment Practices (REAP) Int. Online Conference on Assessment Design for Learner Responsibility, Proceedings for Raising students' meta-cognition (self-assessment) abilities* (2007).

8. Clayson, D. E. Performance Overconfidence: Metacognitive Effects or Misplaced Student Expectations? *J. Mark. Educ.* **27**, 122–129 (2005).

9. Nowell, C. & Alston, R. M. I Thought I Got an A! Overconfidence Across the Economics Curriculum. *J. Econ. Educ.* **38**, 131–142 (2007).

10. Markman, A. Confidence and Certainty From Advisors. *Psychology Today* (2018). Available at: https://www.psychologytoday.com/us/blog/ulterior-

motives/201805/confidence-and-certainty-advisors.

11. Gaertig, C. & Simmons, J. P. Do People Inherently Dislike Uncertain Advice? *Psychol. Sci.* **29**, 504–520 (2018).

12. Wise, S. L. & Kong, X. Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Appl. Meas. Educ.* **18**, 163–183 (2005).

13. Lee, Y.-H. & Jia, Y. Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments Educ.* **2**, 8 (2014).

14. Lee, Y.-H. & Chen, H. A review of recent response-time analyses in educational testing. *Psychol. Test Assess. Model.* **53**, 359–379 (2011).

APPENDICES

Appendix A

Learning Objectives (LOs) for OT 422

Unit I: Introduction to the Human Nervous System and the Lower Extremity

At the time of the Unit I Examination and without the use of any aides, OT 422 students will be ≥ 78% proficient in their ability to…

1. *identify the basic components of the Central Nervous System (CNS) and the Peripheral Nervous System (PNS).*
2. *describe the function and components of the layers of fascia and connective tissue in the lower limb.*
3. *define important anatomical terms (muscle functions, regions, types of actions, attachment site terms, anatomical directions, etc.).*
4. *describe the lymphatics of the lower limb.*
5. *identify the resulting nerves and their derivations (from the spinal cord) of the sacral and lumbar plexuses.*
6. *identify the function(s) of the nerves that result from the sacral and lumbar plexuses.*
7. *describe the location and function of the vasculature of the (A) gluteal region, (B) thigh, (C) leg, and (D) foot.*
8. *identify the muscles of the (A) gluteal region, (B) thigh, (C) leg, and (D) foot.*
9. *identify the proximal and distal attachments of the muscles of the (A) gluteal region, (B) thigh, (C) leg, and (D) foot.*
10. *identify the action(s) of the muscles of the (A) gluteal region, (B) thigh, (C) leg, and (D) foot.*
11. *identify the innervations of the muscles of the (A) gluteal region, (B) thigh, (C) leg, and (D) foot.*
12. *describe the locations and functions of bursa.*
13. *describe the joints and joint components of the lower limb.*
14. *identify the ligaments and aponeuroses of the lower limb.*
15. *describe the locations of anatomical structures in the lower limb compared to one another using anatomical terms.*
16. *describe components of specific anatomical landmarks, compartments, layers, regions, spaces, etc. of the lower limb.*
17. *identify bones and bony structures of the lower limb.*
18. *describe the location and function of the retinacula of the lower limb.*
19. *apply knowledge of the anatomy of the lower limb to OT-related clinical situations.*
20. *predict the effect(s) from the interaction(s) of multiple organ systems on the lower limb.*

Unit II: Upper Extremity

At the time of the Unit II Examination and without the use of any aides, OT 422 students will be ≥ 78% proficient in their ability to…

21. *identify the basic components of the pectoral region and upper limb.*
22. *identify the location and function of fascia and connective tissue in the upper limb.*

23. *describe the lymphatics of the upper limb.*
24. *identify the components and their derivations (from the spinal cord) of the brachial plexus.*
25. *identify the function(s) of the nerves that result from the brachial plexus.*
26. *describe the location and function of the vasculature of the (A) pectoral region and scapula, (B) axilla and arm (C) forearm and (D) hand.*
27. *identify the muscles of the (A) pectoral girdle, (B) arm (C) forearm and (D) hand.*
28. *identify the proximal and distal attachments of the muscles of the (A) pectoral girdle, (B) arm (C) forearm and (D) hand.*
29. *identify the action(s) of the muscles of the (A) pectoral girdle, (B) arm (C) forearm and (D) hand.*
30. *identify the innervations of the muscles of the (A) pectoral girdle, (B) arm (C) forearm and (D) hand.*
31. *describe the locations and functions of tendon sheaths and dorsal expansions.*
32. *describe the joints and joint components of the upper limb.*
33. *identify the ligaments and aponeuroses of the upper limb.*
34. *describe the locations of anatomical structures in the upper limb compared to one another using anatomical terms.*
35. *describe components of specific anatomical landmarks, compartments, layers, regions, spaces, etc. of the upper limb.*
36. *identify the locations and functions of the bones and bony structures of the upper limb.*
37. *describe the location and function of the retinacula of the upper limb.*
38. *apply knowledge of the anatomy of the upper limb to OT-related clinical situations.*
39. *predict the effect(s) from the interaction(s) of multiple organ systems on the upper limb.*

Unit III: Head, Neck and Back

At the time of the Unit III Examination and without the use of any aides, OT 422 students will be ≥ 78% proficient in their ability to…

40. *identify the basic components and functions of the spinal column, neck, and head.*
41. *identify the location and function of fascia and connective tissue in the neck.*
42. *identify the components and their derivations (from the spinal cord) of the cervical plexus.*
43. *identify the function(s) of the nerves that result from the cervical plexus and the brainstem (cranial nerves).*
44. *describe the location and function of the vasculature of the (A) back, (B) head (C) neck.*
45. *identify the muscles of the (A) back, (B) head (C) neck.*
46. *identify the proximal and distal attachments (or superior and inferior attachments) of the muscles of the (A) back, (B) head (C) neck.*
47. *identify the action(s) of the muscles of the (A) back, (B) head (C) neck.*
48. *identify the innervations of the muscles of the (A) back, (B) head (C) neck.*
49. *describe the joints and joint components of the back, head, and neck.*

50. identify the ligaments and aponeuroses of the back, head, and neck.
51. describe the locations of anatomical structures in the head, neck, and back compared to one another using anatomical terms.
52. describe components of specific anatomical landmarks, compartments, layers, regions, spaces, etc. of the back, head, and neck.
53. identify the locations and functions of the bones and bony structures of the back, head, and neck.
54. describe the location and function of the cartilages in the neck.
55. apply knowledge of the anatomy of the back, head, and neck to OT-related clinical situations.
56. predict the effect(s) from the interaction(s) of multiple organ systems on the back, head, and neck.

Unit IV: Thorax and Abdomen

At the time of the Unit IV Examination and without the use of any aides, OT 422 students will be ≥ 78% proficient in their ability to…

57. identify the basic components and functions of the thorax and abdomen.
58. identify the nerves and their functions in the thorax and abdomen.
59. describe the location and function of the vasculature of the (A) thorax and (B) abdomen.
60. identify the muscles of the (A) thorax, (B) abdomen, and (C) pelvis and perineum.
61. identify the proximal and distal attachments (or origins and insertions) of the muscles of the (A) thorax, (B) abdomen, and (C) pelvis and perineum.
62. identify the action(s) of the muscles of the (A) thorax, (B) abdomen, and (C) pelvis and perineum.
63. identify the innervations of the muscles of the (A) thorax, (B) abdomen, and (C) pelvis and perineum.
64. describe the locations and functions of the components of thoracic wall.
65. describe the locations of anatomical structures in the thorax and abdomen compared to one another using anatomical terms.
66. describe components of specific anatomical landmarks, compartments, layers, regions, spaces, etc. of the thorax, abdomen, and pelvis.
67. identify the locations and functions of the bones and bony structures of the thorax, abdomen, and pelvis.
68. describe the location and function of the pleural membranes and peritoneum.
69. describe the locations and functions of the components of the mediastinum and internal aspect of the posterior thoracic wall.
70. identify the location and function of the thoracic and abdominal viscera (except the heart).
71. describe the locations and functions of the components of the heart (including the pericardium and conduction system).
72. describe the locations and functions of the components of fetal circulation.
73. describe the locations and functions of the components of the posterior abdominal wall and diaphragm.

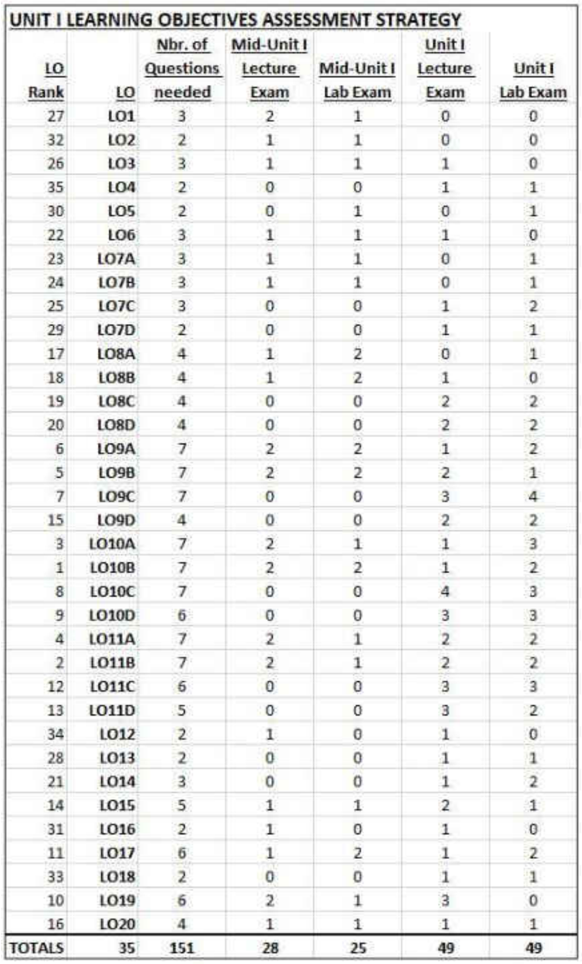
74. *describe the locations and functions of the components of the male and female reproductive systems.*
75. *apply knowledge of the anatomy of the thorax, abdomen, and pelvis to OT-related clinical situations.*
76. *predict the effect(s) from the interaction(s) of multiple organ systems on the thorax, abdomen, and pelvis.*

APPENDIX B

LO Proportions and Assessment Strategy Tables for OT 422

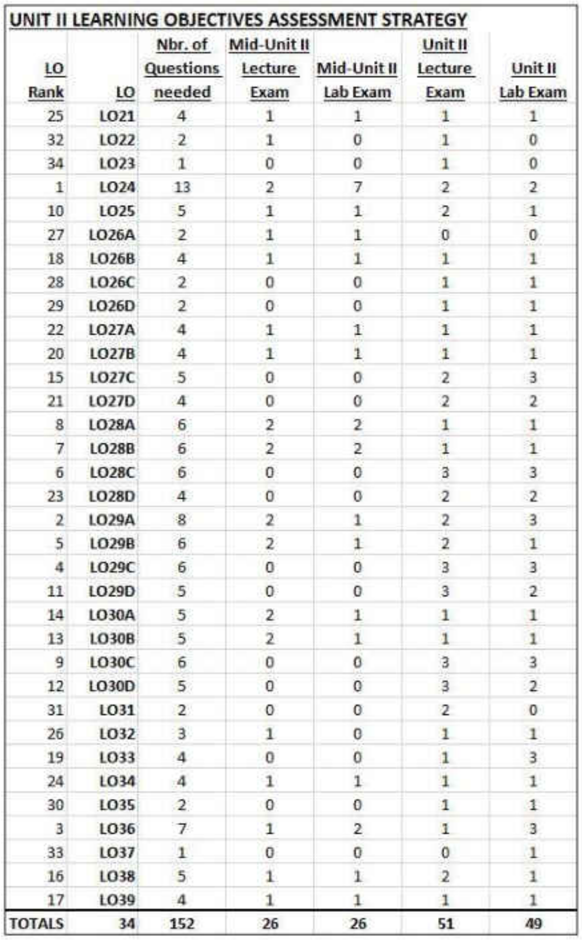
The following methods were employed:

Step 1: Determine amount and type of assessment necessary to judge student proficiency on each LO.
Step 2: Determine amount of items necessary for each LO in each assessment.
Step 3: Total assessment columns to determine necessary assessment length.
Note: If the assessment is too long, consider breaking it into multiple assessments. Do not alter the amount of assessment per LO for the reason of fitting the needed assessment amounts to the assessment tools.
Step 4: Use the resulting information to guide the development of the curriculum and assessment items.
Step 5: (Optional) Rank each LO to signify which LOs are most important regarding the course curriculum and intended outcomes.



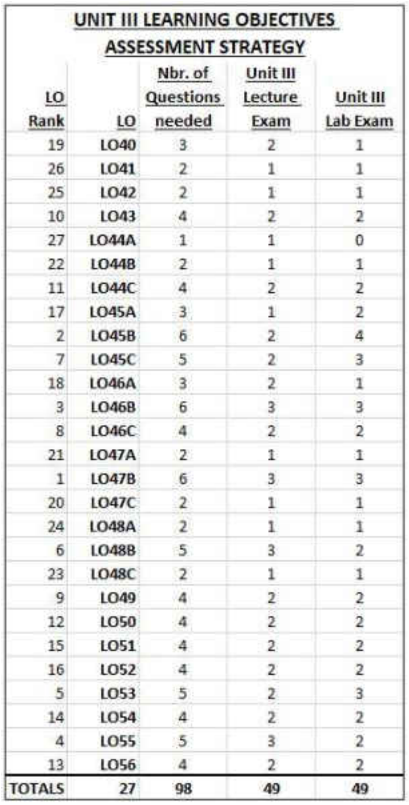
UNIT I LEARNING OBJECTIVES ASSESSMENT STRATEGY

LO Rank	LO	Nbr. of Questions needed	Mid-Unit I Lecture Exam	Mid-Unit I Lab Exam	Unit I Lecture Exam	Unit I Lab Exam
27	LO1	3	2	1	0	0
32	LO2	2	1	1	0	0
26	LO3	3	1	1	1	0
35	LO4	2	0	0	1	1
30	LO5	2	0	1	0	1
22	LO6	3	1	1	1	0
23	LO7A	3	1	1	0	1
24	LO7B	3	1	1	0	1
25	LO7C	3	0	0	1	2
29	LO7D	2	0	0	1	1
17	LO8A	4	1	2	0	1
18	LO8B	4	1	2	1	0
19	LO8C	4	0	0	2	2
20	LO8D	4	0	0	2	2
6	LO9A	7	2	2	1	2
5	LO9B	7	2	2	2	1
7	LO9C	7	0	0	3	4
15	LO9D	4	0	0	2	2
3	LO10A	7	2	1	1	3
1	LO10B	7	2	2	1	2
8	LO10C	7	0	0	4	3
9	LO10D	6	0	0	3	3
4	LO11A	7	2	1	2	2
2	LO11B	7	2	1	2	2
12	LO11C	6	0	0	3	3
13	LO11D	5	0	0	3	2
34	LO12	2	1	0	1	0
28	LO13	2	0	0	1	1
21	LO14	3	0	0	1	2
14	LO15	5	1	1	2	1
31	LO16	2	1	0	1	0
11	LO17	6	1	2	1	2
33	LO18	2	0	0	1	1
10	LO19	6	2	1	3	0
16	LO20	4	1	1	1	1
TOTALS	35	151	28	25	49	49



UNIT II LEARNING OBJECTIVES ASSESSMENT STRATEGY

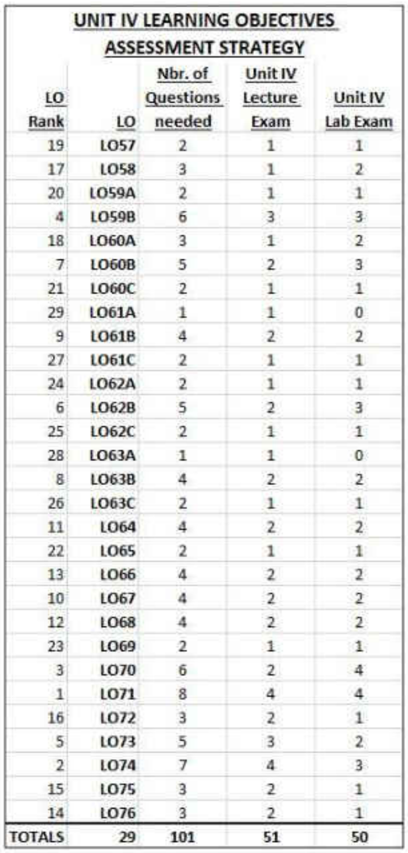
LO Rank	LO	Nbr. of Questions needed	Mid-Unit II Lecture Exam	Mid-Unit II Lab Exam	Unit II Lecture Exam	Unit II Lab Exam
25	LO21	4	1	1	1	1
32	LO22	2	1	0	1	0
34	LO23	1	0	0	1	0
1	LO24	13	2	7	2	2
10	LO25	5	1	1	2	1
27	LO26A	2	1	1	0	0
18	LO26B	4	1	1	1	1
28	LO26C	2	0	0	1	1
29	LO26D	2	0	0	1	1
22	LO27A	4	1	1	1	1
20	LO27B	4	1	1	1	1
15	LO27C	5	0	0	2	3
21	LO27D	4	0	0	2	2
8	LO28A	6	2	2	1	1
7	LO28B	6	2	2	1	1
6	LO28C	6	0	0	3	3
23	LO28D	4	0	0	2	2
2	LO29A	8	2	1	2	3
5	LO29B	6	2	1	2	1
4	LO29C	6	0	0	3	3
11	LO29D	5	0	0	3	2
14	LO30A	5	2	1	1	1
13	LO30B	5	2	1	1	1
9	LO30C	6	0	0	3	3
12	LO30D	5	0	0	3	2
31	LO31	2	0	0	2	0
26	LO32	3	1	0	1	1
19	LO33	4	0	0	1	3
24	LO34	4	1	1	1	1
30	LO35	2	0	0	1	1
3	LO36	7	1	2	1	3
33	LO37	1	0	0	0	1
16	LO38	5	1	1	2	1
17	LO39	4	1	1	1	1
TOTALS	34	152	26	26	51	49



UNIT III LEARNING OBJECTIVES

ASSESSMENT STRATEGY

LO Rank	LO	Nbr. of Questions needed	Unit III Lecture Exam	Unit III Lab Exam
19	LO40	3	2	1
26	LO41	2	1	1
25	LO42	2	1	1
10	LO43	4	2	2
27	LO44A	1	1	0
22	LO44B	2	1	1
11	LO44C	4	2	2
17	LO45A	3	1	2
2	LO45B	6	2	4
7	LO45C	5	2	3
18	LO46A	3	2	1
3	LO46B	6	3	3
8	LO46C	4	2	2
21	LO47A	2	1	1
1	LO47B	6	3	3
20	LO47C	2	1	1
24	LO48A	2	1	1
6	LO48B	5	3	2
23	LO48C	2	1	1
9	LO49	4	2	2
12	LO50	4	2	2
15	LO51	4	2	2
16	LO52	4	2	2
5	LO53	5	2	3
14	LO54	4	2	2
4	LO55	5	3	2
13	LO56	4	2	2
TOTALS	27	98	49	49



UNIT IV LEARNING OBJECTIVES

ASSESSMENT STRATEGY

LO Rank	LO	Nbr. of Questions needed	Unit IV Lecture Exam	Unit IV Lab Exam
19	LO57	2	1	1
17	LO58	3	1	2
20	LO59A	2	1	1
4	LO59B	6	3	3
18	LO60A	3	1	2
7	LO60B	5	2	3
21	LO60C	2	1	1
29	LO61A	1	1	0
9	LO61B	4	2	2
27	LO61C	2	1	1
24	LO62A	2	1	1
6	LO62B	5	2	3
25	LO62C	2	1	1
28	LO63A	1	1	0
8	LO63B	4	2	2
26	LO63C	2	1	1
11	LO64	4	2	2
22	LO65	2	1	1
13	LO66	4	2	2
10	LO67	4	2	2
12	LO68	4	2	2
23	LO69	2	1	1
3	LO70	6	2	4
1	LO71	8	4	4
16	LO72	3	2	1
5	LO73	5	3	2
2	LO74	7	4	3
15	LO75	3	2	1
14	LO76	3	2	1
TOTALS	29	101	51	50

APPENDIX C

Individual LO Pre-R and Post-R PL Frequencies and Resulting Achievement

2016: IRQs with Email

		Pre-R PL Frequencies												Post-R PL Frequencies															
Unit	LO#	I3	C1	I1	I2	C2	C3	P1	P3	P2	Record Error	Total	LO#	I3	C1	I1	I2	C2	C3	P1	P3	P2	Record Error	Total	Mean Pre-R PL	LO Achieved? (Mean PL ≥ 4.50)	Mean Post-R PL	LO Achieved? (Mean PL ≥ 4.50)	Mean R
Unit I	LO1	6	17	19	11	7	0	0	0	0	0	60	LO1	3	4	3	10	15	25	0	0	0	0	60	2.93	N	4.75	Y	1.82
	LO2	8	12	11	6	4	0	0	0	1	0	42	LO2	3	8	2	3	16	8	0	1	0	1	42	2.66	N	4.13	N	1.47
	LO3	9	8	4	5	2	0	0	0	0	0	28	LO3	4	1	1	6	4	12	0	0	0	0	28	2.39	N	4.46	N	2.07
	LO4	1	20	22	7	0	0	0	0	0	0	50	LO4	5	1	17	21	4	2	0	0	0	0	50	2.70	N	3.48	N	0.78
	LO5	4	4	8	1	1	0	1	0	0	0	19	LO5	1	0	4	4	7	0	1	1	1	0	19	2.50	N	4.00	N	1.50
	LO6	3	7	18	15	4	0	0	0	0	1	48	LO6	3	3	7	9	12	10	1	2	1	0	48	3.21	N	4.23	N	1.01
	LO7A	4	25	17	9	13	0	0	0	0	1	69	LO7A	3	8	13	15	24	5	0	0	0	1	69	3.03	N	3.94	N	0.91
	LO7B	6	6	14	4	3	0	0	1	0	0	34	LO7B	4	0	3	8	9	10	0	0	0	0	34	2.76	N	4.41	N	1.65
	LO7C	19	9	3	6	0	0	0	0	0	0	37	LO7C	0	4	10	1	11	11	0	0	0	0	37	1.89	N	4.41	N	2.51
	LO7D	0	3	1	1	0	0	0	0	0	0	5	LO7D	0	0	1	0	4	0	0	0	0	0	5	2.60	N	4.60	Y	2.00
	LO8A	5	8	7	6	0	0	0	0	0	0	26	LO8A	0	0	0	1	0	25	0	0	0	0	26	2.54	N	5.92	Y	3.38
	LO8B	7	12	19	9	3	0	0	0	0	0	50	LO8B	2	1	6	3	17	17	0	3	1	0	50	2.78	N	4.80	Y	2.02
	LO8C	11	7	2	9	0	0	0	0	0	0	29	LO8C	0	1	0	2	5	21	0	0	0	0	29	2.31	N	5.55	Y	3.24
	LO8D	6	1	0	2	0	0	0	0	0	0	9	LO8D	1	0	0	1	2	5	0	0	0	0	9	1.78	N	5.00	Y	3.22
	LO9A	21	8	15	7	5	1	0	0	0	0	57	LO9A	6	2	3	6	10	30	0	0	0	0	57	2.47	N	4.79	Y	2.32
	LO9B	13	5	25	6	14	0	0	0	1	0	64	LO9B	0	2	7	4	21	27	1	1	1	0	64	3.05	N	5.05	Y	2.00
	LO9C	12	12	7	7	3	0	0	0	0	0	41	LO9C	4	1	2	4	11	14	0	3	2	0	41	2.44	N	4.64	Y	2.20
	LO9D	11	4	2	1	1	0	0	0	0	0	19	LO9D	1	0	5	4	2	1	0	1	5	0	19	1.79	N	3.69	N	1.90
	LO10A	4	9	4	2	9	0	0	1	0	1	30	LO10A	1	0	3	6	4	12	1	1	2	0	30	3.11	N	4.85	Y	1.74
	LO10B	12	13	24	27	23	3	0	1	0	2	105	LO10B	3	4	4	22	21	34	2	2	13	0	105	3.44	N	4.77	Y	1.33
	LO10C	8	7	3	4	6	0	0	0	0	0	28	LO10C	1	4	2	2	9	8	1	1	0	0	28	2.75	N	4.46	N	1.71
	LO10D	15	13	6	4	3	0	0	0	0	0	41	LO10D	0	3	7	3	14	9	2	1	2	0	41	2.20	N	4.53	Y	2.33
	LO11A	13	2	5	5	1	0	0	0	0	1	27	LO11A	1	0	0	0	2	24	0	0	0	0	27	2.19	N	5.74	Y	3.55
	LO11B	7	11	10	8	15	3	0	0	0	2	56	LO11B	2	2	1	3	23	20	1	1	3	0	56	3.41	N	5.02	Y	1.61
	LO11C	18	11	8	3	1	0	0	0	0	0	41	LO11C	2	7	3	2	9	17	0	0	1	0	41	1.98	N	4.50	Y	2.52
	LO11D	20	5	2	4	0	0	0	0	0	0	31	LO11D	0	3	0	10	6	12	0	0	0	0	31	1.68	N	4.77	Y	3.10
	LO12	9	4	1	5	2	0	0	0	0	0	21	LO12	1	1	1	0	7	11	0	0	0	0	21	2.38	N	5.10	Y	2.71
	LO13	14	5	1	2	0	0	0	0	0	0	22	LO13	0	2	5	5	3	6	0	0	1	0	22	1.59	N	4.29	N	2.69
	LO14	9	11	7	6	0	0	0	0	0	0	33	LO14	1	3	0	1	11	16	0	1	0	0	33	2.30	N	5.06	Y	2.76
	LO15	19	19	11	4	17	0	0	1	2	0	73	LO15	11	2	9	10	18	23	0	0	0	0	73	2.73	N	4.25	N	1.52
	LO16	6	8	12	13	5	0	0	0	0	1	45	LO16	4	2	1	9	6	23	0	0	0	0	45	3.07	N	4.78	Y	1.71
	LO17	26	13	14	20	7	0	0	0	0	0	80	LO17	1	1	0	3	15	51	0	6	3	0	80	2.61	N	5.58	Y	2.96
	LO18	3	7	0	0	1	0	0	0	0	0	11	LO18	0	2	1	2	4	0	0	1	1	0	11	2.00	N	3.89	N	1.89
	LO19	17	16	42	22	10	0	0	0	0	1	108	LO19	4	4	17	22	18	28	0	14	1	0	108	2.93	N	4.40	N	1.47
	LO20	19	10	32	28	8	0	0	0	0	0	97	LO20	4	6	18	21	29	7	3	5	4	0	97	2.96	N	4.01	N	1.05
Unit II	LO21	3	12	13	14	11	0	0	1	2	0	56	LO21	1	5	5	5	10	30	0	0	0	0	56	3.34	N	4.93	Y	1.59
	LO22	2	7	12	10	6	0	0	0	0	0	37	LO22	3	4	4	10	7	9	0	0	0	0	37	3.30	N	4.11	N	0.81
	LO23	1	2	1	0	0	0	0	0	0	0	4	LO23	0	0	0	0	2	2	0	0	0	0	4	2.00	N	5.50	Y	3.50
	LO24	13	10	3	8	62	17	0	7	0	0	120	LO24	5	3	8	2	20	79	0	2	1	0	120	4.30	N	5.27	Y	0.97
	LO25	7	16	7	22	42	0	0	2	1	0	97	LO25	11	11	10	20	26	17	0	0	1	1	97	3.81	N	3.95	N	0.14
	LO26A	1	10	1	5	18	0	0	1	0	0	36	LO26A	0	1	1	6	7	20	0	0	1	0	36	3.83	N	5.26	Y	1.43
	LO26B	2	3	8	11	4	0	0	0	0	0	28	LO26B	1	1	0	2	3	20	0	0	0	1	28	3.43	N	5.41	Y	1.98
	LO26C	8	3	3	3	1	0	0	0	0	0	18	LO26C	0	2	3	2	7	4	0	0	0	0	18	2.22	N	4.44	N	2.22
	LO26D	4	8	1	4	0	0	0	0	0	0	17	LO26D	1	1	1	3	5	6	0	0	0	0	17	2.29	N	4.65	Y	2.35
	LO27A	9	9	3	10	14	0	0	0	0	0	45	LO27A	1	1	0	1	9	33	0	0	0	0	45	3.24	N	5.56	Y	2.31
	LO27B	11	5	10	22	18	1	0	0	0	0	67	LO27B	4	6	2	5	21	29	0	0	0	0	67	3.51	N	4.79	Y	1.28
	LO27C	5	4	2	4	0	0	0	0	0	0	15	LO27C	1	3	1	1	5	4	0	0	0	0	15	2.33	N	4.20	N	1.87
	LO27D	14	1	2	5	3	0	0	0	0	0	25	LO27D	1	2	0	4	11	7	0	0	0	0	25	2.28	N	4.72	Y	2.44
	LO28A	16	4	2	9	17	3	0	2	0	0	53	LO28A	3	2	0	2	13	27	1	2	2	1	53	3.31	N	5.15	Y	1.84
	LO28B	9	2	2	9	32	5	0	1	0	0	60	LO28B	1	4	3	7	25	13	3	0	3	1	60	4.15	N	4.70	Y	0.55
	LO28C	13	1	3	5	4	0	0	0	0	0	26	LO28C	0	0	3	4	8	11	0	0	0	0	26	2.46	N	5.04	Y	2.58
	LO28D	5	3	2	3	0	0	0	0	0	0	13	LO28D	1	1	0	3	5	3	0	0	0	0	13	2.23	N	4.46	N	2.23
	LO29A	19	7	10	21	27	16	0	0	1	0	101	LO29A	2	4	3	9	31	48	0	4	0	0	101	3.78	N	5.13	Y	1.35
	LO29B	14	7	3	3	30	6	0	0	0	0	63	LO29B	4	5	4	5	19	26	0	0	0	0	63	3.73	N	4.71	Y	0.98
	LO29C	5	4	0	2	13	1	0	0	0	0	25	LO29C	1	0	2	3	10	5	1	0	3	0	25	3.68	N	4.71	Y	1.03
	LO29D	9	2	1	5	5	0	0	0	0	0	22	LO29D	3	0	1	3	10	4	0	0	0	1	22	2.77	N	4.38	N	1.61
	LO30A	9	5	2	7	21	0	0	0	0	0	44	LO30A	1	0	0	1	2	40	0	0	0	0	44	3.59	N	5.80	Y	2.20
	LO30B	2	4	2	1	11	0	0	0	0	0	20	LO30B	0	1	0	0	7	12	0	0	0	0	20	3.75	N	5.45	Y	1.70
	LO30C	13	12	1	9	8	0	0	0	0	0	43	LO30C	2	1	4	7	12	16	0	0	0	1	43	2.70	N	4.76	Y	2.06
	LO30D	8	10	4	8	2	0	0	0	0	0	32	LO30D	5	2	3	6	9	7	0	0	0	0	32	2.56	N	4.03	N	1.47
	LO31	3	4	3	5	2	0	0	0	0	0	17	LO31	0	0	0	0	3	14	0	0	0	0	17	2.94	N	5.82	Y	2.88
	LO32	12	5	2	8	8	0	0	1	0	0	36	LO32	2	1	4	4	14	10	0	0	1	0	36	2.86	N	4.63	Y	1.77
	LO33	7	5	9	7	1	0	0	0	0	0	29	LO33	0	1	2	6	10	7	1	0	1	1	29	2.66	N	4.77	Y	2.11
	LO34	7	18	21	27	21	0	0	2	0	0	96	LO34	10	2	7	12	13	52	0	0	0	0	96	3.39	N	4.79	Y	1.40
	LO35	5	7	3	1	0	0	0	0	0	0	16	LO35	1	1	2	1	5	6	0	0	0	0	16	2.00	N	4.63	Y	2.63
	LO36	13	13	15	16	39	11	0	0	1	0	108	LO36	5	3	2	2	25	70	0	0	0	1	108	3.82	N	5.33	Y	1.50
	LO37	1	0	0	0	1	0	0	0	0	0	2	LO37	0	0	0	0	1	1	0	0	0	0	2	3.00	N	5.50	Y	2.50
	LO38	13	7	13	24	12	0	0	2	0	0	71	LO38	11	1	1	9	18	31	0	0	0	0	71	3.22	N	4.62	Y	1.40
	LO39	29	7	6	26	23	0	0	2	1	0	94	LO39	8	5	13	22	29	9	6	0	2	0	94	3.08	N	4.00	N	0.92

Unit III	LO40	11	1	2	5	2	0	0	0	0	0	21	LO40	0	1	1	1	5	12	0	1	0	0	21	2.33	N	5.30	Y	2.97
	LO41	1	4	3	2	0	0	0	0	0	0	10	LO41	0	2	4	2	1	1	0	0	0	0	10	2.60	N	3.50	N	0.90
	LO42	0	2	0	1	2	0	0	0	0	0	5	LO42	0	0	1	0	1	3	0	0	0	0	5	3.60	N	5.20	Y	1.60
	LO43	16	10	15	19	4	0	0	0	0	0	64	LO43	0	4	12	10	24	14	0	0	0	0	64	2.77	N	4.50	Y	1.73
	LO44A	0	10	3	3	0	0	0	0	0	0	16	LO44A	0	1	1	0	6	8	0	0	0	0	16	2.56	N	5.19	Y	2.63
	LO44B	0	0	0	0	0	0	0	0	0	0	0	LO44B	0	0	0	0	0	0	0	0	0	0	0	--	N/A	--	N/A	--
	LO44C	8	6	4	4	5	0	0	0	0	0	27	LO44C	1	1	3	7	7	8	0	0	0	0	27	2.70	N	4.56	Y	1.85
	LO45A	1	0	0	1	1	0	0	0	0	0	3	LO45A	1	0	0	0	0	2	0	0	0	0	3	3.33	N	4.33	N	1.00
	LO45B	12	1	3	8	12	2	0	0	0	0	38	LO45B	2	2	3	5	7	19	0	0	0	0	38	3.34	N	4.84	Y	1.50
	LO45C	16	8	6	10	6	0	0	0	0	0	46	LO45C	2	2	6	9	10	17	0	0	0	0	46	2.61	N	4.61	Y	2.00
	LO46A	5	6	5	3	0	0	0	0	0	0	19	LO46A	0	2	3	4	3	1	2	0	4	0	19	2.32	N	3.85	N	1.53
	LO46B	12	6	5	3	9	0	0	0	0	0	35	LO46B	3	4	6	11	8	2	1	0	0	0	35	2.74	N	3.68	N	0.93
	LO46C	12	1	3	6	4	0	0	0	0	0	26	LO46C	0	3	1	8	6	7	0	0	1	0	26	2.58	N	4.52	Y	1.94
	LO47A	3	3	3	0	1	0	0	0	0	0	10	LO47A	0	0	1	2	4	3	0	0	0	0	10	2.30	N	4.90	Y	2.60
	LO47B	23	15	9	13	16	1	0	0	0	0	77	LO47B	20	0	2	5	6	29	2	5	8	0	77	2.83	N	4.03	N	1.20
	LO47C	5	3	0	4	1	0	0	0	0	0	13	LO47C	0	0	1	1	4	3	0	2	2	0	13	2.46	N	5.00	Y	2.54
	LO48A	0	1	1	0	1	0	0	0	0	0	3	LO48A	0	0	0	0	1	2	0	0	0	0	3	3.33	N	5.67	Y	2.33
	LO48B	4	8	3	4	4	0	0	0	0	0	23	LO48B	0	4	1	5	6	6	0	1	0	0	23	2.83	N	4.41	N	1.58
	LO48C	2	2	2	0	0	0	0	0	0	0	6	LO48C	0	0	3	0	1	2	0	0	0	0	6	2.00	N	4.33	N	2.33
	LO49	5	11	22	16	6	0	0	0	0	0	60	LO49	1	7	8	7	15	21	1	0	0	0	60	3.12	N	4.54	Y	1.43
	LO50	3	7	2	2	2	0	0	0	0	0	16	LO50	2	3	5	3	2	1	0	0	0	0	16	2.56	N	3.19	N	0.63
	LO51	17	7	12	12	4	0	0	0	0	0	52	LO51	4	5	18	13	6	6	0	0	0	0	52	2.60	N	3.58	N	0.98
	LO52	5	10	7	7	3	0	0	0	1	0	33	LO52	2	2	5	3	12	9	0	0	0	0	33	2.78	N	4.45	N	1.67
	LO53	12	11	10	12	15	0	0	0	0	0	60	LO53	1	4	10	4	26	15	0	0	0	0	60	3.12	N	4.58	Y	1.47
	LO54	7	6	7	4	2	0	0	0	0	0	26	LO54	9	1	1	4	6	5	0	0	0	0	26	2.54	N	3.46	N	0.92
	LO55	5	16	9	5	5	0	0	0	0	0	40	LO55	4	2	1	2	9	20	1	1	0	0	40	2.73	N	4.84	Y	2.12
	LO56	13	9	4	9	4	0	0	0	0	0	39	LO56	0	4	9	10	5	0	4	1	6	0	39	2.54	N	3.57	N	1.03
Unit IV	LO57	6	6	0	0	0	0	0	0	0	0	12	LO57	0	0	0	0	6	4	0	1	1	0	12	1.50	N	5.40	Y	3.90
	LO58	5	7	0	1	0	0	0	0	0	0	13	LO58	0	2	1	1	3	6	0	0	0	0	13	1.77	N	4.77	Y	3.00
	LO59A	10	5	4	2	0	0	0	0	0	0	21	LO59A	0	0	1	1	3	14	1	1	0	0	21	1.90	N	5.58	Y	3.67
	LO59B	8	9	6	14	2	0	0	0	0	0	39	LO59B	1	1	5	5	7	19	0	0	0	1	39	2.82	N	4.92	Y	2.10
	LO60A	8	7	8	5	0	0	0	0	0	0	28	LO60A	1	0	0	1	4	22	0	0	0	0	28	2.36	N	5.61	Y	3.25
	LO60B	8	7	8	11	1	0	0	0	0	0	35	LO60B	2	1	1	4	13	14	0	0	0	0	35	2.71	N	4.91	Y	2.20
	LO60C	1	2	8	0	0	0	0	0	0	0	11	LO60C	0	0	0	3	2	2	1	0	3	0	11	2.64	N	4.86	Y	2.22
	LO61A	1	0	0	1	0	0	0	0	0	0	2	LO61A	0	0	1	0	0	1	0	0	0	0	2	2.50	N	4.50	Y	2.00
	LO61B	6	3	1	1	0	0	0	0	0	0	11	LO61B	0	0	1	0	3	6	0	0	0	1	11	1.73	N	5.40	Y	3.67
	LO61C	2	8	6	2	0	0	0	0	0	0	18	LO61C	1	0	3	0	6	8	0	0	0	0	18	2.44	N	4.89	Y	2.44
	LO62A	11	1	1	3	0	0	0	0	0	0	16	LO62A	0	0	0	2	6	6	1	0	1	0	16	1.75	N	5.29	Y	3.54
	LO62B	4	10	11	13	2	0	0	0	0	0	40	LO62B	0	0	0	1	6	31	0	0	2	0	40	2.98	N	5.79	Y	2.81
	LO62C	1	5	2	0	0	0	0	0	0	0	8	LO62C	1	0	0	1	3	1	0	0	2	0	8	2.13	N	4.33	N	2.21
	LO63A	0	0	0	0	0	0	0	0	0	0	0	LO63A	0	0	0	0	0	0	0	0	0	0	0	--	N/A	--	N/A	--
	LO63B	1	5	3	3	0	0	0	0	0	0	12	LO63B	0	2	1	2	4	3	0	0	0	0	12	2.67	N	4.42	N	1.75
	LO63C	3	10	7	2	0	0	0	0	0	0	22	LO63C	0	6	1	2	10	3	0	0	0	0	22	2.36	N	4.14	N	1.77
	LO64	9	7	2	3	2	0	0	0	0	0	23	LO64	1	2	0	3	1	2	1	5	8	0	23	2.22	N	3.78	N	1.56
	LO65	9	8	4	1	0	0	0	0	0	0	22	LO65	2	1	8	4	3	4	0	0	0	0	22	1.86	N	3.77	N	1.91
	LO66	18	15	18	7	1	0	0	0	0	0	59	LO66	1	0	0	4	16	38	0	0	0	0	59	2.29	N	5.51	Y	3.22
	LO67	7	15	16	12	0	0	0	0	0	0	50	LO67	0	3	4	6	29	7	0	1	0	0	50	2.66	N	4.67	Y	2.01
	LO68	6	11	8	7	3	0	0	0	0	0	35	LO68	1	2	2	5	14	11	0	0	0	0	35	2.71	N	4.77	Y	2.06
	LO69	5	10	4	10	0	0	0	0	0	0	29	LO69	2	0	3	9	4	11	0	0	0	0	29	2.66	N	4.59	Y	1.93
	LO70	21	21	17	9	4	0	0	0	0	0	72	LO70	5	0	1	10	11	28	0	11	6	0	72	2.36	N	4.93	Y	2.57
	LO71	27	1	4	4	7	1	0	0	0	0	44	LO71	0	0	0	1	5	38	0	0	0	0	44	2.23	N	5.84	Y	3.61
	LO72	13	1	1	0	0	0	0	0	0	0	15	LO72	0	1	0	1	1	12	0	0	0	0	15	1.20	N	5.53	Y	4.33
	LO73	10	13	9	7	1	0	0	0	0	0	40	LO73	2	1	2	6	15	14	0	0	0	0	40	2.40	N	4.83	Y	2.43
	LO74	6	14	16	10	7	1	0	0	0	0	54	LO74	1	7	0	2	17	26	0	0	0	1	54	3.02	N	4.98	Y	1.96
	LO75	6	7	1	0	0	0	0	0	0	0	14	LO75	0	0	0	2	5	5	1	1	0	0	14	1.64	N	5.25	Y	3.61
	LO76	8	9	3	3	0	0	0	0	0	0	23	LO76	1	0	1	2	14	4	0	1	0	0	23	2.04	N	4.82	Y	2.77
	Mean:	9.03	7.75	7.56	7.74	8.77	4.8							3.03	2.81	4.21	5.46	9.73	14.7						2.68		4.71		2.03
	Standard Error:	0.56	0.44	0.64	0.58	0.86	0.22							0.26	0.19	0.36	0.45	0.67	1.24						0.05		0.05		0.07
																							Standard Deviation:	0.59		0.59		0.80
																										Count:	**83**	

2017: IRQs with SAE

Unit	LO#	Pre-R PL Frequencies									Record Error	Total	LO#	Post-R PL Frequencies									Record Error	Total	Mean Pre-R PL	LO Achieved? (Mean PL ≥ 4.50)	Mean Post-R PL	LO Achieved? (Mean PL ≥ 4.50)	Mean R
		I3	C1	I1	I2	C2	C3	P1	P3	P2				I3	C1	I1	I2	C2	C3	P1	P3	P2							
Unit I	LO1	8	21	8	13	3	0	0	0	0	0	53	LO1	1	4	2	3	16	27	0	0	0	0	53	2.66	N	5.08	Y	2.42
	LO2	17	7	5	5	4	0	0	0	0	0	38	LO2	1	8	4	3	11	9	2	0	0	0	38	2.26	N	4.17	N	1.90
	LO3	7	11	4	9	2	0	0	0	0	0	33	LO3	0	3	4	5	12	9	0	0	0	0	33	2.64	N	4.61	Y	1.97
	LO4	1	26	9	4	0	0	0	0	0	0	40	LO4	0	2	13	14	6	2	2	0	1	0	40	2.40	N	3.81	N	1.41
	LO5	1	9	7	0	0	0	0	0	0	0	17	LO5	2	1	8	3	1	1	1	0	0	0	17	2.35	N	3.19	N	0.83
	LO6	2	21	31	9	8	0	0	0	0	0	71	LO6	4	6	16	23	9	11	1	1	0	0	71	3.00	N	3.87	N	0.87
	LO7A	1	42	12	6	8	0	0	0	0	0	69	LO7A	0	9	13	14	25	8	0	0	0	0	69	2.68	N	4.14	N	1.46
	LO7B	2	17	4	2	2	0	0	0	0	0	27	LO7B	1	0	4	5	11	5	0	1	0	0	27	2.44	N	4.54	Y	2.09
	LO7C	12	9	9	3	1	0	0	0	0	0	34	LO7C	1	4	8	3	10	8	0	0	0	0	34	2.18	N	4.21	N	2.03
	LO7D	1	8	1	0	0	0	0	0	0	0	10	LO7D	0	2	2	1	4	0	1	0	0	0	10	2.00	N	3.78	N	1.78
	LO8A	5	12	8	4	2	0	0	0	0	0	31	LO8A	0	1	0	1	4	25	0	0	0	0	31	2.55	N	5.68	Y	3.13
	LO8B	9	12	19	5	5	0	0	0	0	0	50	LO8B	0	2	6	9	16	15	0	1	1	0	50	2.70	N	4.75	Y	2.05
	LO8C	13	9	2	4	1	0	0	0	0	0	29	LO8C	0	1	2	0	13	13	0	0	0	0	29	2.00	N	5.21	Y	3.21
	LO8D	8	9	1	1	1	0	0	0	0	0	20	LO8D	1	2	4	2	6	5	0	0	0	0	20	1.90	N	4.25	N	2.35
	LO9A	27	13	22	14	2	1	0	0	0	0	79	LO9A	4	0	12	12	17	34	0	0	0	0	79	2.42	N	4.77	Y	2.35
	LO9B	4	20	18	5	16	1	0	1	0	0	65	LO9B	1	6	8	5	16	27	1	0	1	0	65	3.19	N	4.75	Y	1.56
	LO9C	18	15	6	1	1	0	0	0	0	0	41	LO9C	2	6	3	5	14	10	0	1	0	0	41	1.83	N	4.33	N	2.50
	LO9D	11	6	2	2	1	0	0	0	0	0	22	LO9D	2	5	0	4	4	2	2	1	2	0	22	1.91	N	3.53	N	1.62
	LO10A	6	14	12	5	3	2	0	0	0	0	42	LO10A	2	2	7	6	8	8	5	1	2	1	42	2.79	N	4.21	N	1.43
	LO10B	12	21	36	24	16	5	0	0	0	0	114	LO10B	8	5	17	10	25	38	6	0	4	1	114	3.23	N	4.49	N	1.26
	LO10C	15	16	7	4	2	0	0	0	0	0	44	LO10C	3	6	3	6	14	8	2	1	1	0	44	2.14	N	4.15	N	2.01
	LO10D	18	16	6	2	5	0	0	0	0	0	47	LO10D	2	10	4	6	16	6	1	0	2	0	47	2.15	N	3.95	N	1.81
	LO11A	11	7	6	5	5	1	0	0	0	0	35	LO11A	2	0	1	1	9	22	0	0	0	0	35	2.69	N	5.31	Y	2.63
	LO11B	10	14	12	8	13	3	0	2	0	0	62	LO11B	2	7	7	9	11	22	1	1	2	0	62	3.15	N	4.48	N	1.33
	LO11C	17	22	5	2	2	0	0	0	0	0	48	LO11C	2	1	5	6	18	15	0	0	0	1	48	1.96	N	4.74	Y	2.79
	LO11D	23	9	3	5	1	0	0	0	0	0	41	LO11D	0	5	7	5	17	7	0	0	0	0	41	1.83	N	4.34	N	2.51
	LO12	7	4	1	2	2	0	0	0	0	0	16	LO12	0	2	4	1	1	8	0	0	0	0	16	2.25	N	4.56	Y	2.31
	LO13	11	4	1	1	1	0	0	0	0	0	18	LO13	1	1	9	4	1	2	0	0	0	0	18	1.72	N	3.50	N	1.78
	LO14	9	14	5	1	2	0	0	0	0	0	31	LO14	1	3	1	3	10	13	0	0	0	0	31	2.13	N	4.84	Y	2.71
	LO15	6	16	9	6	10	0	0	1	1	0	49	LO15	6	4	4	9	11	14	0	0	0	1	49	2.96	N	4.19	N	1.23
	LO16	3	17	7	6	3	0	0	0	0	0	36	LO16	2	1	5	6	10	12	0	0	0	0	36	2.69	N	4.58	Y	1.89
	LO17	19	13	28	13	2	0	0	0	0	0	75	LO17	1	6	2	3	17	45	0	1	0	0	75	2.55	N	5.22	Y	2.67
	LO18	2	12	1	0	0	0	0	0	0	0	15	LO18	0	2	7	3	2	1	0	0	0	0	15	1.93	N	3.53	N	1.60
	LO19	13	30	47	22	8	0	0	0	0	0	120	LO19	4	10	25	14	34	28	0	3	1	1	120	2.85	N	4.29	N	1.44
	LO20	19	27	37	18	9	0	0	0	0	0	110	LO20	2	11	30	15	29	16	2	0	4	1	110	2.74	N	4.03	N	1.29
Unit II	LO21	3	18	15	5	4	0	0	1	0	0	46	LO21	1	2	8	4	13	17	0	0	0	1	46	2.76	N	4.71	Y	1.96
	LO22	0	11	11	5	11	0	0	0	0	0	38	LO22	1	5	9	4	7	12	0	0	0	0	38	3.42	N	4.24	N	0.82
	LO23	0	2	0	0	0	0	0	0	0	0	2	LO23	0	0	0	0	1	1	0	0	0	0	2	2.00	N	5.50	Y	3.50
	LO24	12	24	13	16	59	28	0	1	0	0	153	LO24	3	3	2	7	29	102	2	1	3	1	153	4.12	N	5.48	Y	1.36
	LO25	10	22	4	18	45	0	0	1	1	0	101	LO25	8	15	15	18	23	21	0	0	0	1	101	3.67	N	3.96	N	0.29
	LO26A	1	11	4	4	13	0	1	0	1	0	35	LO26A	1	6	1	1	7	19	0	0	0	0	35	3.52	N	4.83	Y	1.31
	LO26B	4	5	2	1	1	0	0	0	0	0	13	LO26B	0	0	6	0	2	5	0	0	0	0	13	2.23	N	4.46	N	2.23
	LO26C	12	5	6	3	1	0	0	0	0	0	27	LO26C	0	6	1	3	9	6	0	0	0	2	27	2.11	N	4.32	N	2.21
	LO26D	5	3	1	0	0	0	0	0	0	0	9	LO26D	0	0	1	2	3	1	0	0	1	1	9	1.56	N	4.57	Y	3.02
	LO27A	10	15	5	6	7	0	0	0	0	0	43	LO27A	0	4	0	2	8	29	0	0	0	0	43	2.65	N	5.35	Y	2.70
	LO27B	3	17	12	10	16	0	0	0	0	0	58	LO27B	1	8	3	6	15	25	0	0	0	0	58	3.33	N	4.74	Y	1.41
	LO27C	6	7	0	2	2	0	0	0	0	0	17	LO27C	1	2	3	0	6	3	0	0	0	2	17	2.24	N	4.13	N	1.90
	LO27D	6	7	3	1	0	0	0	0	0	0	17	LO27D	0	0	3	4	6	4	0	0	0	0	17	1.94	N	4.65	Y	2.71
	LO28A	16	9	3	9	12	1	0	1	0	0	51	LO28A	2	3	3	2	13	25	1	0	1	1	51	2.90	N	5.00	Y	2.10
	LO28B	8	10	8	15	24	3	0	0	1	0	69	LO28B	5	5	12	14	17	14	2	0	0	0	69	3.68	N	4.12	N	0.44
	LO28C	11	10	0	2	6	0	0	0	0	0	29	LO28C	1	1	6	3	5	13	0	0	0	0	29	2.38	N	4.69	Y	2.31
	LO28D	3	7	5	2	1	0	1	1	0	0	20	LO28D	1	2	3	2	3	9	0	0	0	0	20	2.50	N	4.55	Y	2.05
	LO29A	11	10	8	10	34	27	0	2	1	0	103	LO29A	1	3	5	5	30	55	1	2	1	0	103	4.27	N	5.27	Y	1.00
	LO29B	11	18	5	7	27	9	0	0	0	0	77	LO29B	4	8	6	10	13	35	0	0	0	1	77	3.62	N	4.64	Y	1.02
	LO29C	5	14	2	2	8	0	0	0	0	0	31	LO29C	0	2	5	5	9	5	1	0	4	0	31	2.81	N	4.38	N	1.58
	LO29D	5	8	1	7	6	0	0	0	0	0	27	LO29D	1	1	0	6	10	9	0	0	0	0	27	3.04	N	4.85	Y	1.81
	LO30A	5	7	3	11	18	0	0	0	0	0	44	LO30A	1	1	0	0	4	37	0	0	0	1	44	3.68	N	5.70	Y	2.02
	LO30B	4	5	1	1	2	0	0	0	0	0	13	LO30B	1	1	1	3	4	3	0	0	0	0	13	2.38	N	4.31	N	1.92
	LO30C	9	16	4	9	11	0	0	0	0	0	49	LO30C	2	6	14	7	14	6	0	0	0	0	49	2.94	N	3.88	N	0.94
	LO30D	4	11	6	7	3	0	0	0	0	0	31	LO30D	0	2	7	6	8	7	0	0	0	1	31	2.81	N	4.37	N	1.56
	LO31	3	9	2	0	1	0	0	0	0	0	15	LO31	0	1	1	0	5	7	0	0	0	1	15	2.13	N	5.14	Y	3.01
	LO32	4	6	1	3	8	0	0	0	0	0	22	LO32	0	1	1	2	6	12	0	0	0	0	22	3.23	N	5.23	Y	2.00
	LO33	6	11	9	5	1	0	0	0	0	0	32	LO33	1	3	5	3	10	9	0	0	0	1	32	2.50	N	4.45	N	1.95
	LO34	5	28	20	29	24	0	0	1	0	0	107	LO34	11	3	8	18	21	46	0	0	0	0	107	3.37	N	4.62	Y	1.25
	LO35	7	2	16	7	0	0	0	0	0	0	32	LO35	1	3	2	4	12	10	0	0	0	0	32	2.72	N	4.66	Y	1.94
	LO36	13	14	20	11	30	16	0	0	0	0	104	LO36	3	5	3	5	26	62	0	0	0	0	104	3.76	N	5.23	Y	1.47
	LO37	2	0	0	0	0	0	0	0	0	0	2	LO37	0	0	0	0	0	1	1	0	0	0	2	1.00	N	6.00	Y	5.00
	LO38	7	24	16	22	14	0	0	0	0	0	83	LO38	7	4	7	9	35	20	0	0	0	1	83	3.14	N	4.48	N	1.33
	LO39	22	22	21	43	23	0	0	0	1	0	132	LO39	6	15	26	23	34	15	8	0	4	1	132	3.18	N	3.92	N	0.74

Unit III	LO40	2	8	0	3	0	0	0	0	0	0	13	LO40	0	1	1	0	4	7	0	0	0	0	13	2.31	N	5.15	Y	2.85
	LO41	1	5	3	1	0	0	0	0	0	0	10	LO41	0	3	4	3	0	0	0	0	0	0	10	2.40	N	3.00	N	0.60
	LO42	5	3	3	4	0	0	0	0	0	0	15	LO42	0	2	0	0	5	8	0	0	0	0	15	2.40	N	5.13	Y	2.73
	LO43	10	21	14	12	6	0	0	0	0	0	63	LO43	3	4	15	5	23	13	0	0	0	0	63	2.73	N	4.27	N	1.54
	LO44A	1	4	3	1	0	0	0	0	0	0	9	LO44A	0	2	0	0	3	4	0	0	0	0	9	2.44	N	4.78	Y	2.33
	LO44B	1	0	0	0	1	0	0	0	0	0	2	LO44B	0	0	1	0	1	0	0	0	0	0	2	3.00	N	4.00	N	1.00
	LO44C	3	10	3	5	1	0	0	0	0	0	22	LO44C	1	3	5	2	8	3	0	0	0	0	22	2.59	N	4.00	N	1.41
	LO45A	11	4	7	5	0	0	0	0	0	0	27	LO45A	0	0	0	0	5	22	0	0	0	0	27	2.22	N	5.81	Y	3.59
	LO45B	13	13	8	10	6	2	0	0	0	0	52	LO45B	5	3	3	3	17	21	0	0	0	0	52	2.79	N	4.67	Y	1.88
	LO45C	6	11	6	8	7	0	0	0	0	0	38	LO45C	0	3	8	6	12	9	0	0	0	0	38	2.97	N	4.42	N	1.45
	LO46A	1	0	0	1	1	0	0	0	0	0	3	LO46A	0	1	0	1	0	1	0	0	0	0	3	3.33	N	4.00	N	0.67
	LO46B	4	14	6	6	6	3	0	0	0	0	39	LO46B	0	9	6	1	11	6	2	2	1	1	39	3.13	N	3.97	N	0.84
	LO46C	11	6	0	2	1	0	0	0	0	0	20	LO46C	0	0	5	1	8	6	0	0	0	0	20	1.80	N	4.75	Y	2.95
	LO47A	3	6	5	1	0	0	0	0	0	0	15	LO47A	0	6	3	2	2	2	0	0	0	0	15	2.27	N	3.40	N	1.13
	LO47B	4	26	8	13	16	3	0	0	0	0	70	LO47B	1	6	1	0	20	41	1	0	0	0	70	3.29	N	5.25	Y	1.96
	LO47C	10	3	0	2	2	0	0	0	0	0	17	LO47C	0	3	1	1	4	8	0	0	0	0	17	2.00	N	4.76	Y	2.76
	LO48A	1	3	0	0	0	0	0	0	0	0	4	LO48A	0	0	1	0	3	0	0	0	0	0	4	1.75	N	4.50	Y	2.75
	LO48B	2	10	6	4	2	0	0	0	0	0	24	LO48B	3	6	2	1	9	2	1	0	0	0	24	2.75	N	3.57	N	0.82
	LO48C	2	0	0	0	0	0	0	0	0	0	2	LO48C	0	0	0	0	0	2	0	0	0	0	2	1.00	N	6.00	Y	5.00
	LO49	5	22	23	17	8	0	0	0	0	0	75	LO49	1	16	7	7	27	12	1	2	2	0	75	3.01	N	4.13	N	1.12
	LO50	6	16	3	0	0	0	0	0	0	0	25	LO50	2	5	11	3	2	2	0	0	0	0	25	1.88	N	3.16	N	1.28
	LO51	7	14	12	12	4	0	0	0	0	0	49	LO51	1	10	15	16	4	1	1	0	0	1	49	2.84	N	3.32	N	0.48
	LO52	1	9	5	4	1	0	0	0	2	0	22	LO52	3	2	5	0	7	5	0	0	0	0	22	2.75	N	3.95	N	1.20
	LO53	6	28	10	9	11	0	0	0	0	0	64	LO53	1	9	9	3	24	18	0	0	0	0	64	2.86	N	4.47	N	1.61
	LO54	4	17	5	5	3	0	0	0	1	0	35	LO54	2	5	3	7	10	8	0	0	0	0	35	2.59	N	4.20	N	1.61
	LO55	3	18	19	3	6	0	0	0	0	0	49	LO55	1	2	3	4	10	29	0	0	0	0	49	2.82	N	5.18	Y	2.37
	LO56	11	16	5	2	5	0	0	1	0	0	40	LO56	0	5	15	5	6	1	7	0	1	0	40	2.33	N	3.47	N	1.14
Unit IV	LO57	2	1	1	0	0	0	0	0	0	0	4	LO57	0	0	0	1	2	1	0	0	0	0	4	1.75	N	5.00	Y	3.25
	LO58	5	12	3	1	1	0	0	0	0	0	22	LO58	1	4	2	3	5	7	0	0	0	0	22	2.14	N	4.27	N	2.14
	LO59A	16	3	0	2	0	0	0	0	0	0	21	LO59A	1	1	3	3	1	9	0	2	1	0	21	1.43	N	4.61	Y	3.18
	LO59B	4	26	8	2	2	0	0	0	0	0	42	LO59B	0	1	2	5	8	26	0	0	0	0	42	2.33	N	5.33	Y	3.00
	LO60A	1	9	10	2	0	0	0	0	0	0	22	LO60A	1	0	0	1	7	13	0	0	0	0	22	2.59	N	5.36	Y	2.77
	LO60B	21	17	5	6	3	0	0	0	0	0	52	LO60B	9	0	3	6	9	24	0	0	1	0	52	2.10	N	4.53	Y	2.43
	LO60C	1	1	1	0	0	0	0	0	0	0	3	LO60C	0	0	0	0	0	3	0	0	0	0	3	2.00	N	6.00	Y	4.00
	LO61A	1	0	0	0	0	0	0	0	0	0	1	LO61A	0	0	0	1	0	0	0	0	0	0	1	1.00	N	4.00	N	3.00
	LO61B	2	8	1	1	1	0	0	0	0	0	13	LO61B	0	1	0	0	6	6	0	0	0	0	13	2.31	N	5.23	Y	2.92
	LO61C	1	3	0	0	0	0	0	0	0	0	4	LO61C	0	1	0	2	1	0	0	0	0	0	4	1.75	N	3.75	N	2.00
	LO62A	5	2	0	2	1	0	0	0	0	0	10	LO62A	1	0	0	2	3	4	0	0	0	0	10	2.20	N	4.80	Y	2.60
	LO62B	12	24	12	15	3	0	0	0	0	0	66	LO62B	2	1	0	2	9	48	0	2	1	1	66	2.59	N	5.56	Y	2.97
	LO62C	1	13	2	0	0	0	0	0	0	0	16	LO62C	0	0	0	4	7	5	0	0	0	0	16	2.06	N	5.06	Y	3.00
	LO63A	0	0	0	0	0	0	0	0	0	0	0	LO63A	0	0	0	0	0	0	0	0	0	0	0	--	N/A	--	N/A	--
	LO63B	3	14	1	1	1	0	0	0	0	0	20	LO63B	0	4	1	1	4	10	0	0	0	0	20	2.15	N	4.75	Y	2.60
	LO63C	4	13	5	1	0	0	0	0	0	0	23	LO63C	0	5	0	1	10	6	0	0	1	0	23	2.13	N	4.55	Y	2.42
	LO64	5	12	0	0	1	0	0	0	0	0	18	LO64	0	1	1	2	4	1	2	1	6	0	18	1.89	N	4.33	N	2.44
	LO65	8	7	4	1	0	0	0	0	0	0	20	LO65	1	1	6	8	3	1	0	0	0	0	20	1.90	N	3.70	N	1.80
	LO66	12	22	8	4	0	0	0	0	0	0	46	LO66	1	1	0	3	11	30	0	0	0	0	46	2.09	N	5.43	Y	3.35
	LO67	12	13	7	4	0	0	0	0	0	0	36	LO67	1	5	1	3	13	9	1	2	1	0	36	2.08	N	4.53	Y	2.45
	LO68	4	20	9	2	1	0	0	0	0	0	36	LO68	0	5	2	0	17	12	0	0	0	0	36	2.33	N	4.81	Y	2.47
	LO69	7	6	5	3	0	0	0	0	0	0	21	LO69	1	0	5	7	1	7	0	0	0	0	21	2.19	N	4.33	N	2.14
	LO70	16	33	14	16	4	0	0	0	0	0	83	LO70	2	2	3	1	25	48	1	0	1	0	83	2.51	N	5.33	Y	2.83
	LO71	20	17	8	8	2	1	0	0	0	0	56	LO71	1	1	3	0	9	42	0	0	0	0	56	2.25	N	5.52	Y	3.27
	LO72	12	7	1	0	1	0	0	0	0	0	21	LO72	1	0	5	0	2	13	0	0	0	0	21	1.62	N	4.95	Y	3.33
	LO73	1	24	13	5	1	0	0	0	0	0	44	LO73	0	4	4	7	16	13	0	0	0	0	44	2.57	N	4.68	Y	2.11
	LO74	16	16	9	11	2	0	0	0	0	0	54	LO74	0	3	0	1	13	37	0	0	0	0	54	2.39	N	5.50	Y	3.11
	LO75	6	10	2	2	0	0	0	0	0	0	20	LO75	0	0	2	4	3	11	0	0	0	0	20	2.00	N	5.15	Y	3.15
	LO76	4	21	2	2	1	0	0	0	0	0	30	LO76	1	4	3	4	15	3	0	0	0	0	30	2.17	N	4.23	N	2.07
	Mean:	7.52	13	8.36	6.65	7.15	6.63							2.3	4.06	5.84	5.24	10.7	14.8						2.49		4.57		2.07
	Standard Error:	0.52	0.71	0.73	0.59	0.8	0.34							0.18	0.29	0.49	0.42	0.73	1.37						0.05		0.06		0.08
																								Standard Deviation:	0.60		0.63		0.86
																											Count:	67	

2018: SRQs with SAE

				Pre-R PL Frequencies												Post-R PL Frequencies															
Unit	Exam	Year	LO#	I3	C1	I1	I2	C2	C3	P1	P3	P2	Record Error	Total	LO#	I3	C1	I1	I2	C2	C3	P1	P3	P2	Record Error	Total	Mean Pre-R PL	LO Achieved? (Mean PL ≥ 4.50)	Mean Post-R PL	LO Achieved? (Mean PL ≥ 4.50)	Mean R
Unit I	UI Lecture	16/17	LO19	17	18	47	27	8	0	0	0	0	1	118	LO19	1	6	2	2	19	34	0	0	0	1	65	2.92	N	5.09	Y	2.17
		16/17	LO20	5	8	47	35	5	0	0	0	0	0	100	LO20	2	9	20	15	16	1	0	0	0	2	65	3.27	N	3.59	N	0.32
		16/17	LO10B	6	15	21	24	23	4	0	0	0	2	95	LO10B	4	5	5	9	17	24	0	0	0	1	65	3.59	N	4.59	Y	1.00
		16/17	LO6	1	10	25	21	9	0	0	0	0	1	67	LO6	2	10	11	14	18	9	0	0	0	1	65	3.41	N	3.98	N	0.58
		16	LO16	4	1	7	12	5	0	0	0	0	1	30	LO16	2	3	3	3	22	31	0	0	0	1	65	3.45	N	5.08	Y	1.63
		17	LO11B	1	1	5	8	9	1	0	1	0	0	26	LO11B	0	1	2	3	15	43	0	0	0	1	65	4.04	N	5.52	Y	1.48
	UI Lab	17	LO9A	11	3	12	8	2	1	0	0	0	0	37	LO9A	5	3	1	5	10	40	0	0	0	1	65	2.73	N	5.06	Y	2.33
		16/17	LO17	19	9	25	16	5	0	0	0	0	0	74	LO17	4	1	1	2	13	42	0	0	1	1	65	2.72	N	5.30	Y	2.59
		16	LO1	0	12	13	4	2	0	0	0	0	0	31	LO1	2	13	10	4	19	13	0	0	2	2	65	2.87	N	4.05	N	1.18
		16/17	LO10B	5	8	21	13	12	4	0	1	0	0	64	LO10B	0	0	2	4	14	28	4	1	11	1	65	3.49	N	5.42	Y	1.92
		16/17	LO9B	7	4	33	7	10	1	0	1	1	0	64	LO9B	2	7	6	5	15	27	0	0	0	2	64	3.19	N	4.69	Y	1.50
		17	LO20	3	14	9	1	6	0	0	0	0	0	33	LO20	0	13	1	3	19	5	10	1	8	5	65	2.79	N	4.05	N	1.26
	Mid-Unit II Lecture	16/17	LO19	7	25	18	11	7	0	0	0	0	0	68	LO19	1	10	30	11	10	2	0	0	0	1	65	2.79	N	3.39	N	0.60
		16	LO8B	1	12	14	8	1	0	0	0	0	0	36	LO8B	0	7	16	9	12	20	0	0	0	1	65	2.89	N	4.34	N	1.45
		16/17	LO11C	24	20	11	4	1	0	0	0	0	0	60	LO11C	4	7	3	9	25	17	0	0	0	1	66	1.97	N	4.46	N	2.49
		17	LO11D	16	5	3	3	1	0	0	0	0	0	28	LO11D	2	14	7	6	21	14	0	0	0	1	65	1.86	N	4.13	N	2.27
		16/17	LO10D	18	20	5	2	4	0	0	0	0	0	49	LO10D	1	11	7	7	26	12	0	0	0	1	65	2.06	N	4.28	N	2.22
		16/17	LO4	2	23	17	7	0	0	0	0	0	0	49	LO4	0	10	18	7	17	12	0	0	0	1	65	2.59	N	4.05	N	1.46
	Mid-Unit II Lab	16/17	LO9C	27	20	13	8	3	0	0	0	0	0	71	LO9C	3	4	13	8	17	15	2	0	1	2	65	2.15	N	4.28	N	2.13
		16/17	LO17	15	12	11	9	1	0	0	0	0	0	48	LO17	0	10	5	1	16	25	2	1	4	1	65	2.35	N	4.72	Y	2.37
		17	LO10C	10	10	3	2	1	0	0	0	0	0	26	LO10C	2	12	3	3	24	16	2	0	2	1	65	2.00	N	4.38	N	2.38
		17	LO10B	2	7	12	2	2	0	0	0	0	0	25	LO10B	1	5	15	6	14	12	8	0	3	1	65	2.80	N	4.19	N	1.39
		16	LO7C	18	5	1	1	0	0	0	0	0	0	25	LO7C	1	15	13	2	20	11	2	0	0	1	65	1.40	N	3.94	N	2.54
		17	LO10D	8	7	5	1	3	0	0	0	0	0	24	LO10D	0	14	12	4	26	5	2	0	1	1	65	2.33	N	3.93	N	1.60
Unit II	UII Lecture	16/17	LO39	40	0	10	48	4	0	0	2	1	0	105	LO39	4	9	19	14	14	4	0	0	0	1	65	2.76	N	3.58	N	0.81
		16/17	LO34	2	20	28	27	27	0	0	0	0	0	104	LO34	1	3	1	4	20	35	0	0	0	1	65	3.55	N	5.25	Y	1.70
		16/17	LO25	6	15	5	13	25	0	0	0	0	0	64	LO25	4	10	26	7	12	3	0	0	0	3	65	3.56	N	3.35	N	-0.21
		16/17	LO27B	4	15	17	19	23	0	0	0	0	0	78	LO27B	0	7	2	1	25	28	0	0	0	2	65	3.54	N	5.03	Y	1.49
		16/17	LO24	1	7	0	1	43	24	0	0	0	0	76	LO24	1	2	0	2	14	45	0	0	0	1	65	4.97	Y	5.52	Y	0.54
		16/17	LO22	2	12	21	14	17	0	0	0	0	0	66	LO22	5	7	16	12	15	9	0	0	0	1	65	3.48	N	3.81	N	0.33
	UII Lab	16/17	LO24	8	15	6	9	39	13	0	7	0	0	97	LO24	0	0	0	0	8	56	0	0	0	1	65	4.06	N	5.88	Y	1.82
		16/17	LO28B	11	7	9	17	36	8	0	1	1	0	90	LO28B	2	7	13	6	17	4	7	1	7	1	65	3.95	N	3.84	N	-0.12
		16/17	LO28A	17	4	4	15	26	4	0	3	0	0	73	LO28A	1	7	3	0	14	38	0	0	1	1	65	3.59	N	5.11	Y	1.53
		17	LO29A	3	1	0	2	14	14	0	0	1	0	35	LO29A	0	4	6	5	20	11	6	6	6	1	65	4.91	Y	4.61	Y	-0.30
		17	LO25	4	3	0	10	14	0	0	1	1	0	33	LO25	8	2	11	10	18	14	0	1	0	1	65	3.87	N	4.11	N	0.24
		16/17	LO36	7	4	2	5	28	15	0	0	1	0	62	LO36	1	2	1	1	13	46	0	0	0	1	65	4.44	N	5.52	Y	1.07
	Unit III Lecture	16/17	LO38	16	14	18	33	2	0	0	0	0	0	83	LO38	11	3	8	15	14	12	0	0	0	2	65	2.89	N	3.86	N	0.97
		16/17	LO29B	16	22	6	8	21	0	0	0	0	0	73	LO29B	2	10	20	11	14	6	0	0	0	2	65	2.95	N	3.68	N	0.74
		16/17	LO36	6	16	4	14	20	0	0	0	0	0	60	LO36	1	6	3	8	14	31	0	0	0	2	65	3.43	N	4.92	Y	1.49
		16/17	LO30C	7	19	2	10	12	0	0	0	0	0	50	LO30C	0	18	8	10	19	8	0	0	0	2	65	3.02	N	3.86	N	0.84
		16/17	LO30D	8	16	8	14	4	0	0	0	0	0	50	LO30D	2	7	11	12	19	11	0	0	0	3	65	2.80	N	4.16	N	1.36
		17	LO39	1	9	3	4	5	0	0	0	0	0	22	LO39	0	12	37	4	7	3	0	0	0	2	65	3.14	N	3.24	N	0.10
	Unit III Lab	16/17	LO29A	18	9	8	16	13	2	0	0	0	0	66	LO29A	1	9	17	3	11	21	1	0	0	2	65	3.05	N	4.24	N	1.20
		16/17	LO24	12	9	7	7	23	4	0	0	0	0	62	LO24	1	4	3	2	13	13	4	14	9	2	65	3.52	N	4.69	Y	1.18
		16/17	LO36	9	7	29	7	16	3	0	0	0	0	71	LO36	0	3	4	2	27	26	0	0	0	3	65	3.32	N	5.11	Y	1.79
		17	LO39	4	9	10	2	3	0	0	0	0	0	28	LO39	0	15	14	4	8	1	14	0	6	3	65	2.68	N	3.19	N	0.51
		16/17	LO33	11	13	18	11	1	0	0	0	0	0	54	LO33	1	10	8	2	22	16	4	0	0	2	65	2.59	N	4.39	N	1.80
		17	LO35	2	1	16	7	0	0	0	0	0	0	26	LO35	3	14	13	7	9	15	1	0	0	3	65	3.08	N	3.82	N	0.74
Unit III	Unit IV Lecture	16/17	LO43	19	19	22	19	7	0	0	0	0	0	86	LO43	1	15	6	2	26	13	0	0	0	2	65	2.72	N	4.21	N	1.49
		16/17	LO49	4	21	21	23	11	0	0	0	0	0	80	LO49	1	12	15	13	15	7	0	0	0	2	65	3.20	N	3.79	N	0.59
		16/17	LO51	17	15	13	22	6	0	0	0	0	0	73	LO51	3	12	28	10	8	2	0	0	0	2	65	2.79	N	3.22	N	0.43
		16/17	LO53	6	15	5	10	20	0	0	0	0	0	56	LO53	0	3	2	3	22	33	0	0	0	2	65	3.41	N	5.27	Y	1.86
		16/17	LO45C	10	15	10	16	7	0	0	0	0	0	58	LO45C	7	11	8	10	23	4	0	0	0	2	65	2.91	N	3.68	N	0.77
		16/17	LO47B	3	13	3	9	22	4	0	0	0	0	54	LO47B	0	2	0	0	11	50	0	0	0	2	65	3.85	N	5.70	Y	1.85
	Unit IV Lab	16/17	LO47B	24	28	14	17	10	0	0	0	0	0	93	LO47B	8	7	7	10	11	19	0	0	1	2	65	2.58	N	4.06	N	1.48
		16/17	LO45B	24	13	11	16	9	1	0	0	0	0	74	LO45B	1	8	6	5	17	25	1	0	0	2	65	2.68	N	4.68	Y	2.00
		16/17	LO55	5	29	25	8	4	0	0	0	0	0	71	LO55	0	4	2	2	12	43	0	0	0	2	65	2.68	N	5.40	Y	2.72
		16/17	LO53	12	24	15	11	6	0	0	0	0	0	68	LO53	0	11	20	3	22	6	0	0	0	3	65	2.63	N	3.87	N	1.24
		16/17	LO49	6	12	24	10	3	0	0	0	0	0	55	LO49	1	6	4	1	20	27	0	2	0	4	65	2.85	N	4.93	Y	2.08
		16/17	LO46B	16	13	9	8	10	1	0	0	0	0	57	LO46B	5	2	24	14	6	8	1	1	2	2	65	2.75	N	3.64	N	0.89
Unit IV	Post-Unit IV Lecture	16/17	LO70	7	36	20	15	5	0	0	0	0	0	83	LO70	5	3	1	9	16	29	0	0	0	2	65	2.70	N	4.83	Y	2.13
		17	LO71	12	11	4	5	2	1	0	0	0	0	35	LO71	3	0	2	1	11	46	0	0	0	2	65	2.34	N	5.46	Y	3.12
		16/17	LO62B	5	18	18	23	3	0	0	0	0	0	67	LO62B	3	0	1	1	7	51	0	0	0	2	65	3.01	N	5.57	Y	2.56
		16/17	LO60B	12	14	13	15	4	0	0	0	0	0	58	LO60B	3	5	8	12	17	18	0	0	0	2	65	2.74	N	4.41	N	1.67
		17	LO76	4	17	2	2	1	0	0	0	0	0	26	LO76	0	14	11	8	25	5	0	0	0	2	65	2.19	N	3.94	N	1.74
		16	LO67	1	8	8	7	0	0	0	0	0	0	24	LO67	2	14	6	8	27	6	0	0	0	2	65	2.88	N	3.98	N	1.11
	Post Unit IV Lab	16/17	LO70	30	18	11	10	3	0	0	0	0	0	72	LO70	2	2	2	1	17	34	1	0	4	2	65	2.14	N	5.26	Y	3.12
		16/17	LO74	17	16	21	11	5	0	0	0	0	0	70	LO74	0	16	9	2	19	17	0	0	0	2	65	2.59	N	4.19	N	1.60
		16	LO66	7	9	15	3	1	0	0	0	0	0	35	LO66	5	10	0	7	18	23	0	0	0	2	65	2.49	N	4.46	Y	1.97
		17	LO62B	9	13	4	3	2	0	0	0	0	0	31	LO62B	2	3	0	0	13	43	1	0	1	2	65	2.23	N	5.43	Y	3.20
		16/17	LO73	9	23	13	11	0	0	0	0	0	0	56	LO73	2	7	7	6	16	25	0	0	0	2	65	2.46	N	4.62	Y	2.15
		16/17	LO67	18	14	11	7	0	0	0	0	0	0	50	LO67	1	7	9	1	8	3	11	7	15	3	65	2.14	N	3.59	N	1.45
			Mean:	10.1	13.1	13.3	11.5	10.3	5.83							2.75	7.87	9.37	6.07	16.4	20.2						2.97		4.44		1.47
			Standard Error:	0.94	0.85	1.15	1.03	1.18	0.48							0.26	0.55	0.95	0.5	0.63	1.72						0.08		0.08		0.09
																										Standard Deviation:	0.66		0.69		0.80
																													Count:	31	

APPENDIX D

Items from the Motivated Strategies for Learning Questionnaire (MSLQ)

Although not included in the administered questionnaire, negatively worded items have been identified here with "(REVERSED)" at the end of the question item. Question items are also presented according to their respective scales, whereas in the administration of the MSLQ the questions would be presented in numerical order. The following Likert-Type scale is used to collect a response for each item:

(not at all true of me) 1 2 3 4 5 6 7 (very true of me)

Part A: Motivation Scales

Value Component: Intrinsic Goal Orientation

1. In a class like this, I prefer course material that really challenges me so I can learn new things.
16. In a class like this, I prefer course material that arouses my curiosity, even if it is difficult to learn.
22. The most satisfying thing for me in this course is trying to understand the content as thoroughly as possible.
24. When I have the opportunity in this class, I choose course assignments that I can learn from even if they don't guarantee a good grade.

Value Component: Extrinsic Goal Orientation

7. Getting a good grade in this class is the most satisfying thing for me right now.
11. The most important thing for me right now is improving my overall grade point average, so my main concern in this class is getting a good grade.
13. If I can, I want to get better grades in this class than most of the other students.
30. I want to do well in this class because it is important to show my ability to my family, friends, employer, or others.

Value Component: Task Value

4. I think I will be able to use what I learn in this course in other courses.
10. It is important for me to learn the course material in this class.
17. I am very interested in the content area of this course.
23. I think the course material in this class is useful for me to learn.
26. I like the subject matter of this course.
27. Understanding the subject matter of this course is very important to me.

Expectancy Component: Control Beliefs

2. If I study in appropriate ways, then I will be able to learn the material in this course.
9. It is my own fault if I don't learn the material in this course.
18. If I try hard enough, then I will understand the course material.
25. If I don't understand the course material, it is because I didn't try hard enough.

Expectancy Component: Self-Efficacy for Learning and Performance

5. I believe I will receive an excellent grade in this course.
6. I'm certain I can understand the most difficult material presented in the readings for this course.

12. I'm confident I can learn the basic concepts taught in this course.
15. I'm confident I can understand the most complex material presented by the instructor in this course.
20. I'm confident I can do an excellent job on the assignments and tests in this course.
21. I expect to do well in this class.
29. I'm certain I can master the skills being taught in this class.
31. Considering the difficulty of this course, the teacher, and my skills, I think I will do well in this class.

Affective Component: Test Anxiety
3. When I take a test I think about how poorly I am doing compared to other students.
8. When I take a test I think about items on other parts of the test I can't answer.
14. When I take tests I think of the consequences of failing.
19. I have an uneasy, upset feeling when I take an exam.
28. I feel my heart beating fast when I take an exam.

Part B: Learning Strategies Scales

Cognitive and Metacognitive Strategies: Rehearsal
39. When I study for this class, I practice saying the material to myself over and over.
46. When studying for this course, I read my class notes and course readings over and over again.
59. I memorize key works to remind me of important concepts in this class.
72. I make lists of important items for this course and memorize the lists.

Cognitive and Metacognitive Strategies: Elaboration
53. When I study for this class, I pull together information from different sources, such as lectures, readings, and discussions.
62. I try to relate ideas in this subject to those in other courses whenever possible.
64. When reading for this class, I try to relate the material to what I already know.
67. When I study for this course, I write brief summaries of the main ideas from the readings and my class notes.
69. I try to understand the material in this class by making connections between the readings and the concepts from the lectures.
81. I try to apply ideas from course readings in other class activities such as lecture and discussion.

Cognitive and Metacognitive Strategies: Organization
32. When I study the readings for this course, I outline the material to help me organize my thoughts.
42. When I study for this course, I go through the readings and my class notes and try to find the most important ideas.
49. I make simple charts, diagrams, or tables to help me organize course material.
63. When I study for this course, I go over my class notes and make an outline of important concepts.

Cognitive and Metacognitive Strategies: Critical Thinking
38. I often find myself questioning things I hear or read in this course to decide if I find them interesting.

47. When a theory, interpretation, or conclusion is presented in class or in the readings, I try to decide if there is good supporting evidence.
51. I treat the course material as a starting point and try to develop my own ideas about it.
66. I try to play around with ideas of my own related to what I am learning in this course.
71. Whenever I read or hear an assertion or conclusion in this class, I think about possible alternatives.

Cognitive and Metacognitive Strategies: Metacognitive Self-Regulation

33. During class time I often miss important points because I'm thinking of other things. (REVERSED)
36. When reading for this course, I make up questions to help focus my reading.
41. When I become confused about something I'm reading for this class, I go back and try to figure it out.
44. If course readings are difficult to understand, I change the way I read the material.
54. Before I study new course material thoroughly, I often skim it to see how it is organized.
55. I ask myself questions to make sure I understand the material I have been studying in this class.
56. I try to change the way I study in order to fit the course requirements and the instructor's teaching style.
57. I often find that I have been reading for this class but don't know what it was all about. (REVERSED)
61. I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over when studying for this course.
76. When studying for this course I try to determine which concepts I don't understand well.
78. When I study for this class, I set goals for myself in order to direct my activities in each study period.
79. If I get confused taking notes in class, I make sure I sort it out afterwards.

Resource Management Strategies: Time and Study Environment

35. I usually study in a place where I can concentrate on my course work.
43. I make good use of my study time for this course.
52. I find it hard to stick to a study schedule. (REVERSED)
65. I have a regular place set aside for studying.
70. I make sure that I keep up with the weekly readings and assignments for this course.
73. I attend this class regularly.
77. I often find that I don't spend very much time on this course because of other activities. (REVERSED)
80. I rarely find time to review my notes or readings before the exam. (REVERSED)

Resource Management Strategies: Effort Regulation

37. I often feel so lazy or bored when I study for this class that I quit before I finish what I planned to do. (REVERSED)
48. I work hard to do well in this class even if I don't like what we are doing.
60. When course work is difficult, I either give up or only study the easy parts. (REVERSED)
74. Even when course materials are dull and uninteresting, I manage to keep working until I finish.

Resource Management Strategies: Peer Learning

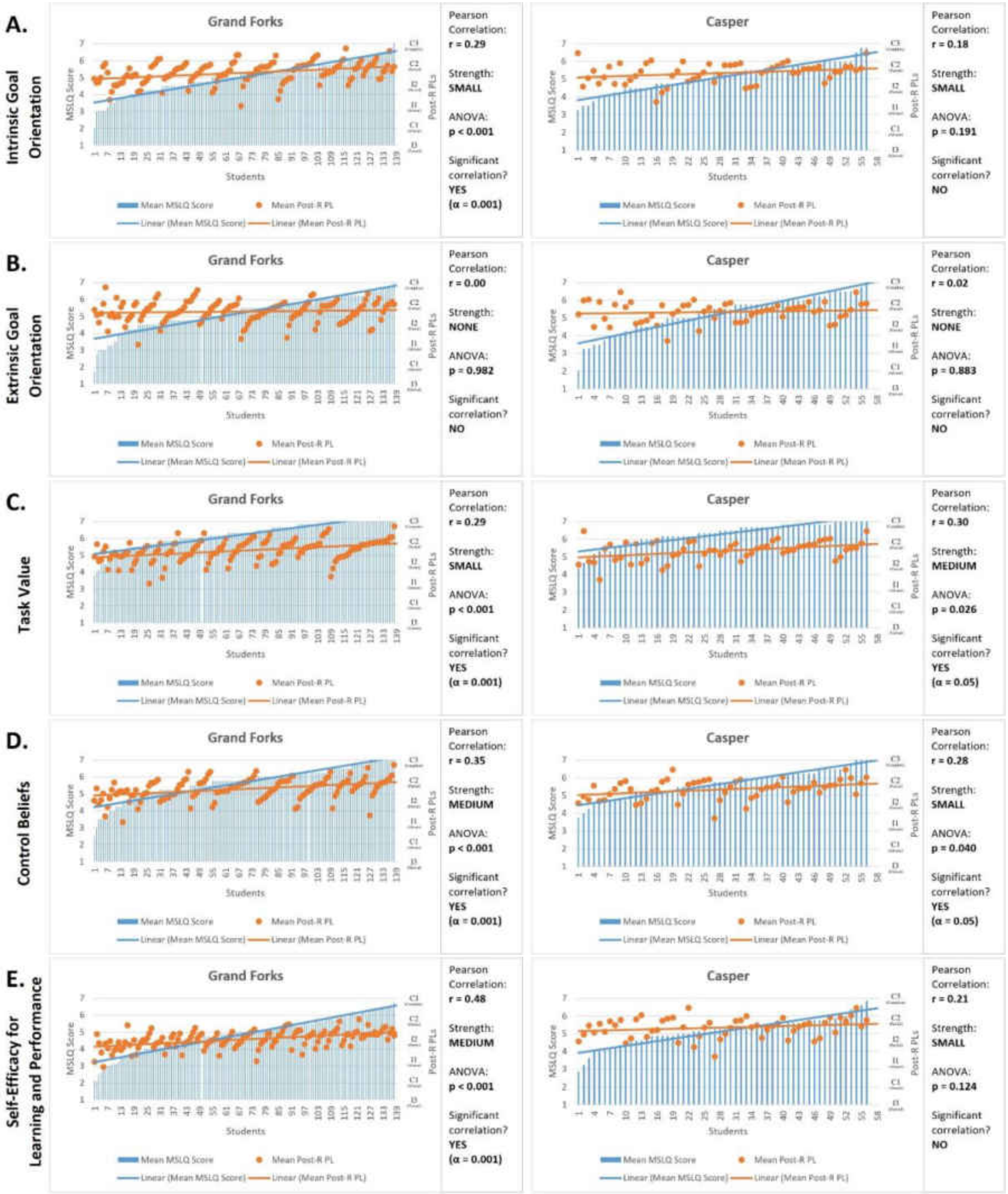
34. When studying for this course, I often try to explain the material to a classmate or friend.
45. I try to work with other students from this class to complete the course assignments.
50. When studying for this course, I often set aside time to discuss course material with a group of students from the class.

Resource Management Strategies: Help Seeking

40. Even if I have trouble learning the material in this class, I try to do the work on my own, without help from anyone. (REVERSED)
58. I ask the instructor to clarify concepts I don't understand well.
68. When I can't understand the material in this course, I ask another student in this class for help.
75. I try to identify students in this class whom I can ask for help if necessary.

APPENDIX E

Figure III-3. Mean MSLQ Scale Score Correlations vs. Post-R PLs.



A. Intrinsic Goal Orientation

Grand Forks: Pearson Correlation: **r = 0.29** Strength: **SMALL** ANOVA: **p < 0.001** Significant correlation? **YES (α = 0.001)**

Casper: Pearson Correlation: **r = 0.18** Strength: **SMALL** ANOVA: **p = 0.191** Significant correlation? **NO**

B. Extrinsic Goal Orientation

Grand Forks: Pearson Correlation: **r = 0.00** Strength: **NONE** ANOVA: **p = 0.982** Significant correlation? **NO**

Casper: Pearson Correlation: **r = 0.02** Strength: **NONE** ANOVA: **p = 0.883** Significant correlation? **NO**

C. Task Value

Grand Forks: Pearson Correlation: **r = 0.29** Strength: **SMALL** ANOVA: **p < 0.001** Significant correlation? **YES (α = 0.001)**

Casper: Pearson Correlation: **r = 0.30** Strength: **MEDIUM** ANOVA: **p = 0.026** Significant correlation? **YES (α = 0.05)**

D. Control Beliefs

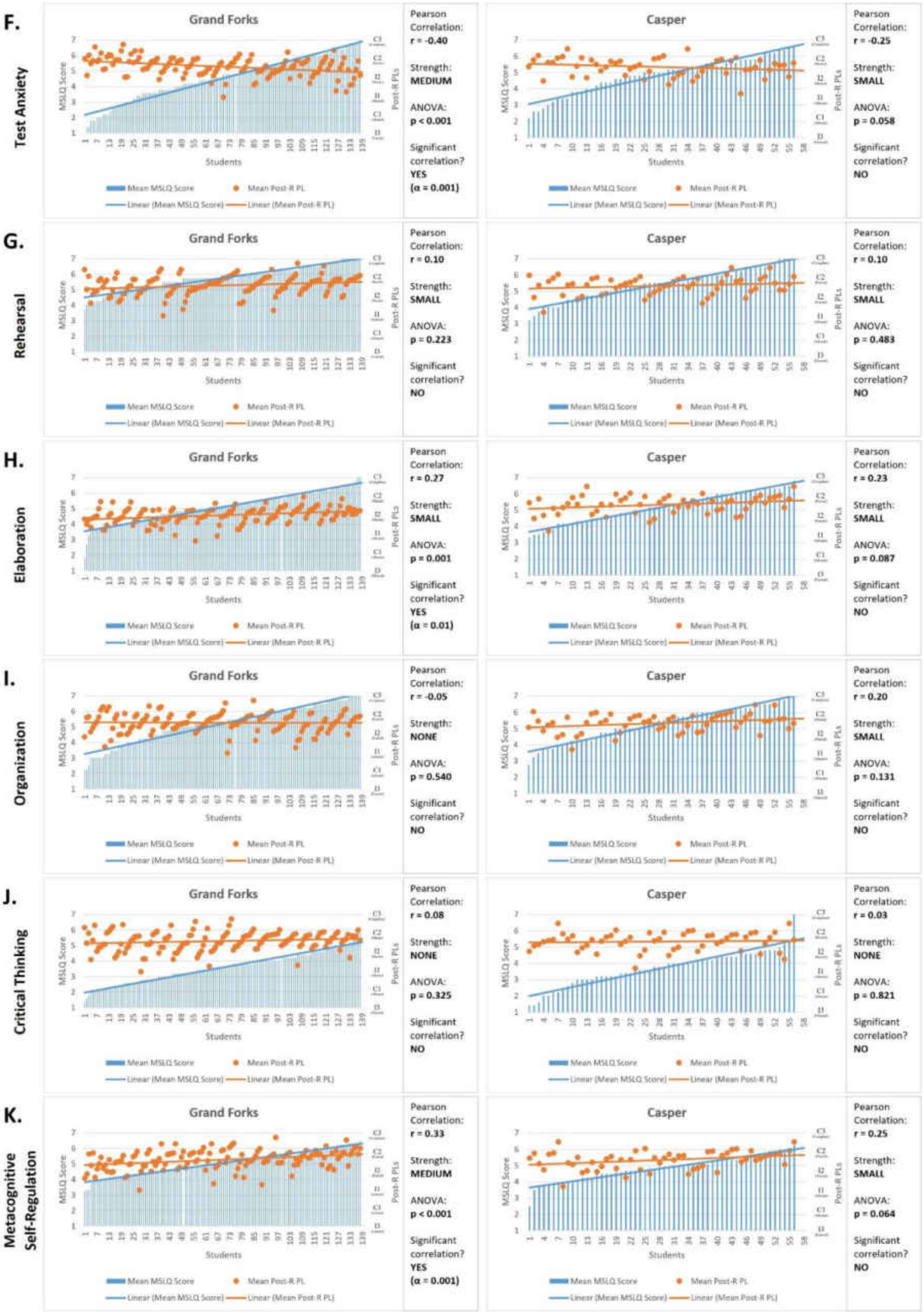
Grand Forks: Pearson Correlation: **r = 0.35** Strength: **MEDIUM** ANOVA: **p < 0.001** Significant correlation? **YES (α = 0.001)**

Casper: Pearson Correlation: **r = 0.28** Strength: **SMALL** ANOVA: **p = 0.040** Significant correlation? **YES (α = 0.05)**

E. Self-Efficacy for Learning and Performance

Grand Forks: Pearson Correlation: **r = 0.48** Strength: **MEDIUM** ANOVA: **p < 0.001** Significant correlation? **YES (α = 0.001)**

Casper: Pearson Correlation: **r = 0.21** Strength: **SMALL** ANOVA: **p = 0.124** Significant correlation? **NO**



F. Test Anxiety

Grand Forks: Pearson Correlation: **r = -0.40**; Strength: **MEDIUM**; ANOVA: **p < 0.001**; Significant correlation? **YES (α = 0.001)**

Casper: Pearson Correlation: **r = -0.25**; Strength: **SMALL**; ANOVA: **p = 0.058**; Significant correlation? **NO**

G. Rehearsal

Grand Forks: Pearson Correlation: **r = 0.10**; Strength: **SMALL**; ANOVA: **p = 0.223**; Significant correlation? **NO**

Casper: Pearson Correlation: **r = 0.10**; Strength: **SMALL**; ANOVA: **p = 0.483**; Significant correlation? **NO**

H. Elaboration

Grand Forks: Pearson Correlation: **r = 0.27**; Strength: **SMALL**; ANOVA: **p = 0.001**; Significant correlation? **YES (α = 0.01)**

Casper: Pearson Correlation: **r = 0.23**; Strength: **SMALL**; ANOVA: **p = 0.087**; Significant correlation? **NO**

I. Organization

Grand Forks: Pearson Correlation: **r = -0.05**; Strength: **NONE**; ANOVA: **p = 0.540**; Significant correlation? **NO**

Casper: Pearson Correlation: **r = 0.20**; Strength: **SMALL**; ANOVA: **p = 0.131**; Significant correlation? **NO**

J. Critical Thinking

Grand Forks: Pearson Correlation: **r = 0.08**; Strength: **NONE**; ANOVA: **p = 0.325**; Significant correlation? **NO**

Casper: Pearson Correlation: **r = 0.03**; Strength: **NONE**; ANOVA: **p = 0.821**; Significant correlation? **NO**

K. Metacognitive Self-Regulation

Grand Forks: Pearson Correlation: **r = 0.33**; Strength: **MEDIUM**; ANOVA: **p < 0.001**; Significant correlation? **YES (α = 0.001)**

Casper: Pearson Correlation: **r = 0.25**; Strength: **SMALL**; ANOVA: **p = 0.064**; Significant correlation? **NO**

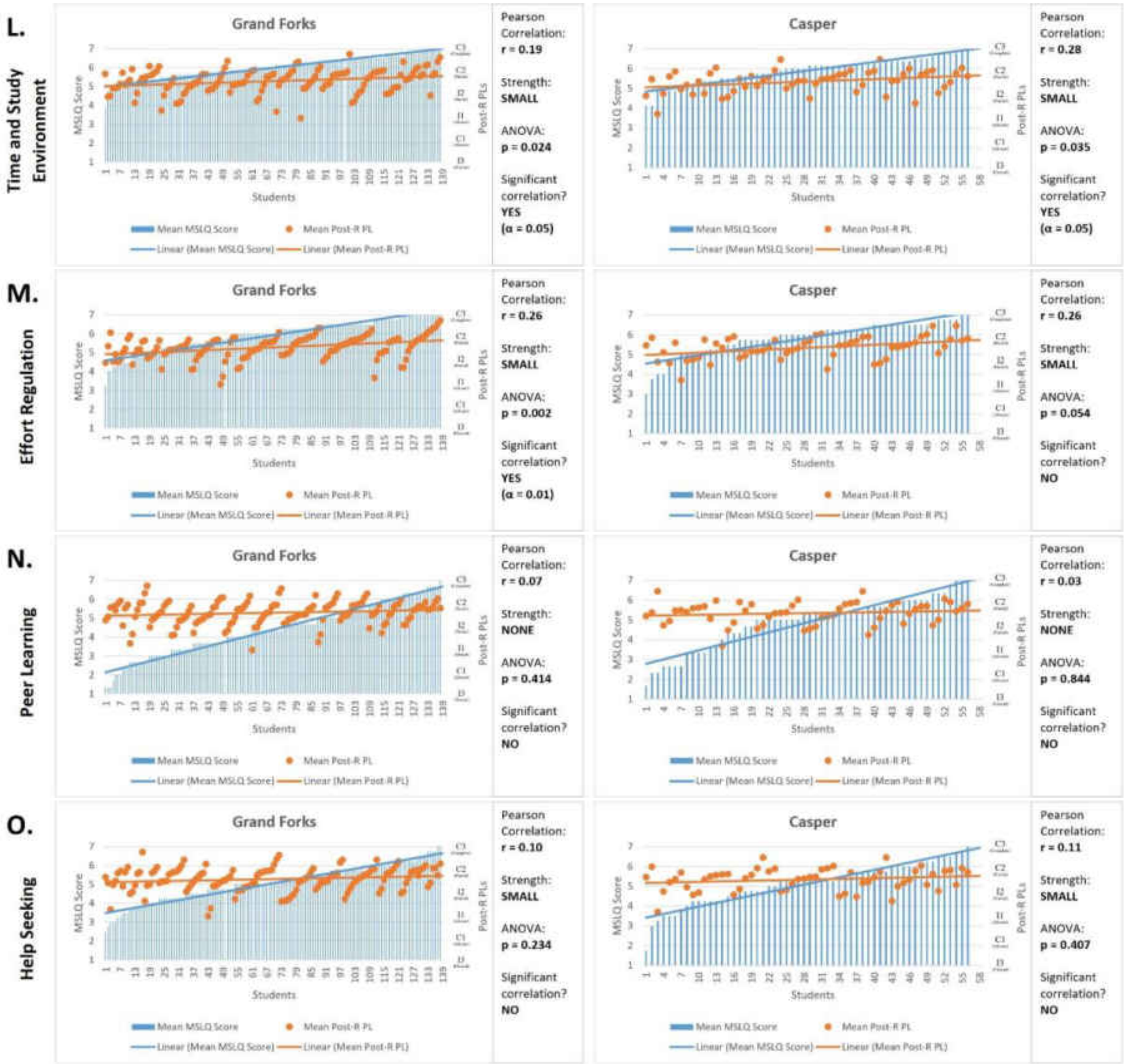
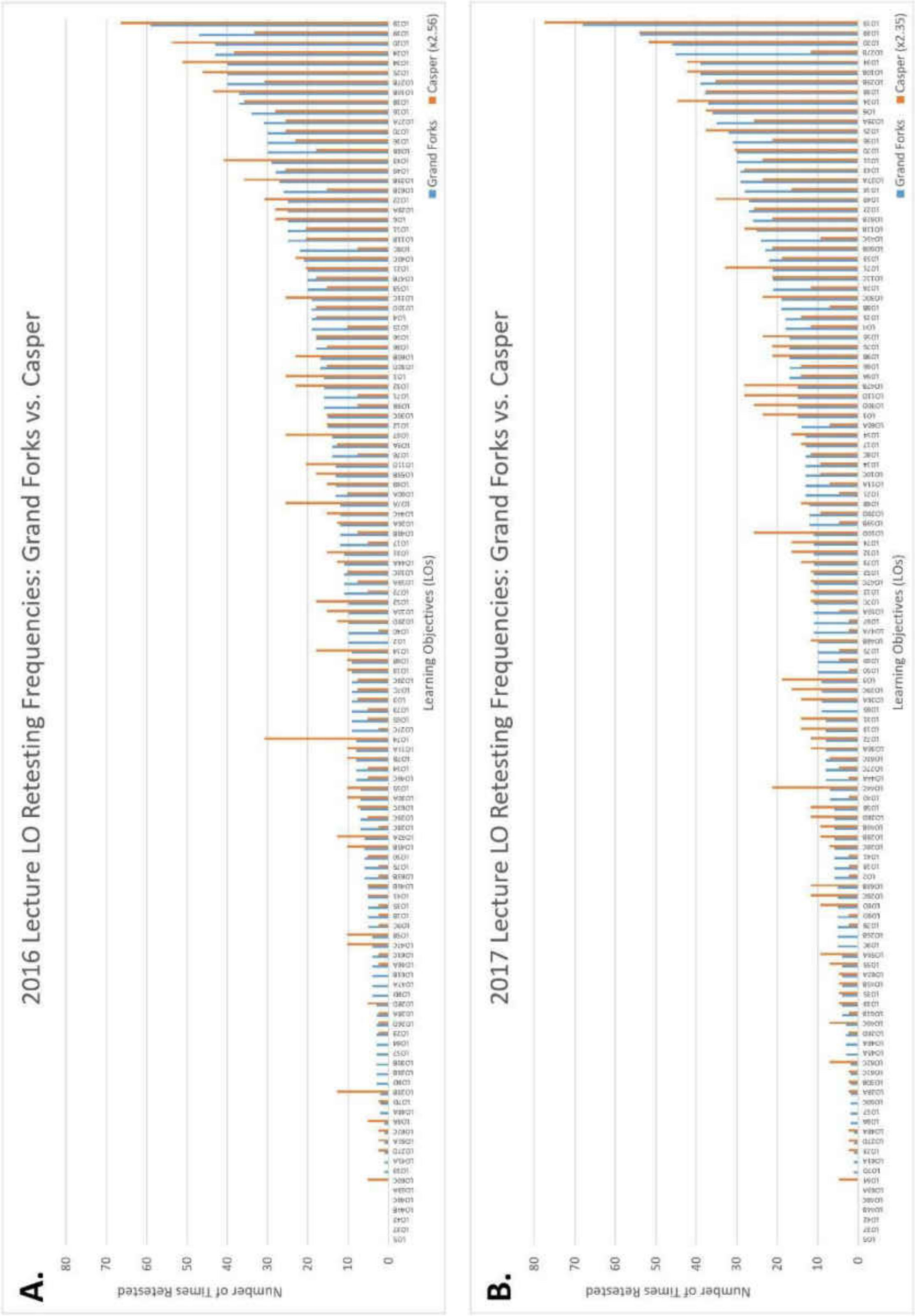


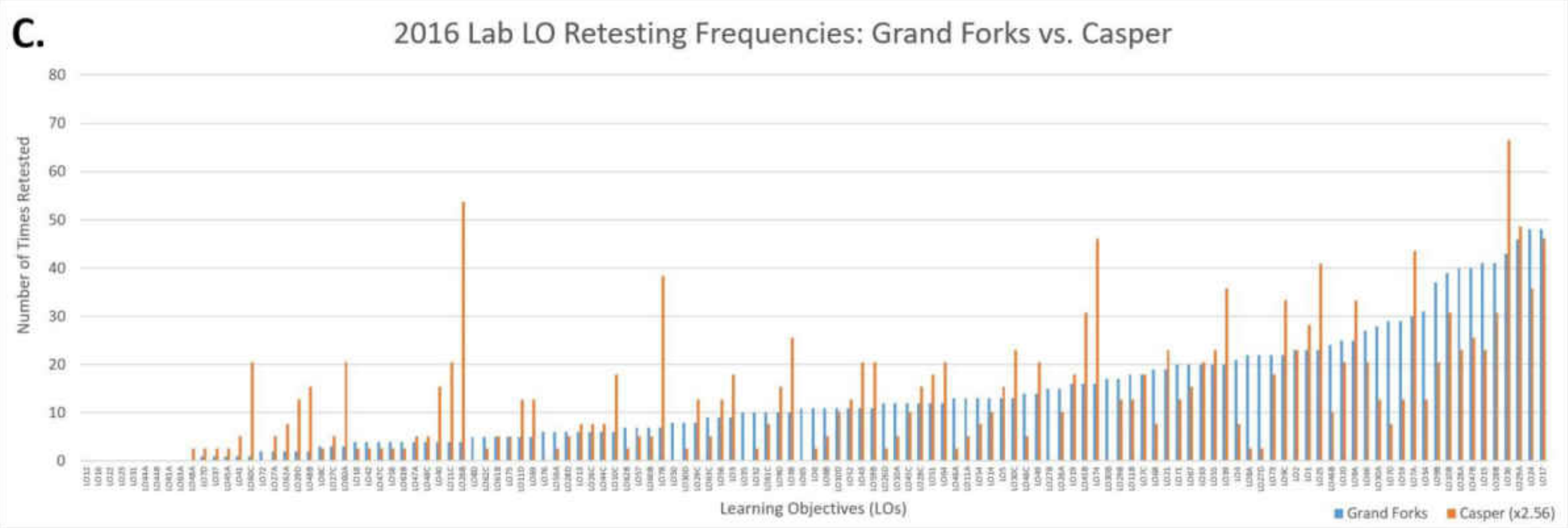
Figure III-3. Mean MSLQ Scale Score Correlations vs. Post-R PLs. Students' mean MSLQ scale scores were correlated to their respective mean post-R PLs (from their remediation questions) for each of the 15 MSLQ scales (A – O). All students from the 2016, 2017, and 2018 cohorts are individually represented within their respective Grand Forks or Casper populations for each MSLQ scale. Data is presented according to increasing mean MSLQ score then by increasing mean post-R PL. A regression analysis was completed for each data set. Resulting Pearson correlations (r) are presented with their interpreted strengths; r < 0.10 = no correlation, r ≥ 0.10 and ≤ 0.29 = small correlation, r ≥ 0.30 and ≤ 0.49 = medium correlation, and r ≥ 0.50 = strong correlation. ANOVA p values were used to determine significance of each Pearson correlation. Significant positive correlations resulted for both Grand Forks and Casper students with Task Value, Control Beliefs, and Time and Study Environment scales.

APPENDIX F

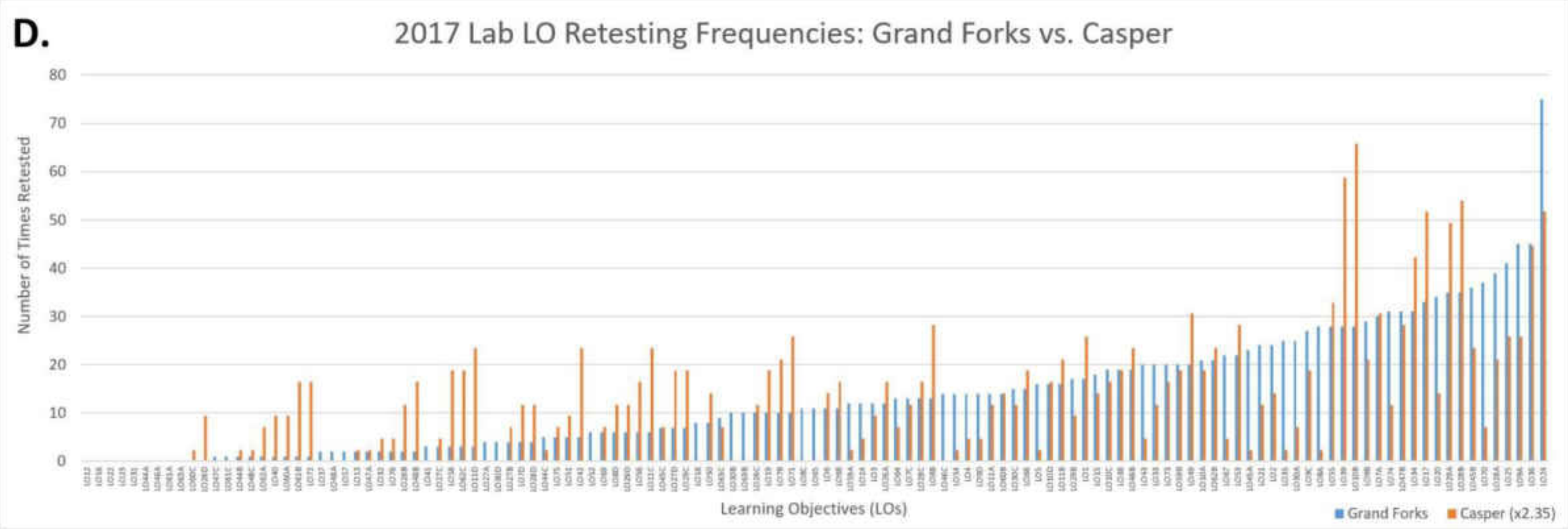
Figure III-8. Site- and Exam Type- Specific IRQ LO Retesting Frequencies.



C.


2016 Lab LO Retesting Frequencies: Grand Forks vs. Casper
Number of Times Retested
0
10
20
30
40
50
60
70
80
Learning Objectives (LOs)
Grand Forks
Casper (x2.56)

D.


2017 Lab LO Retesting Frequencies: Grand Forks vs. Casper
Number of Times Retested
0
10
20
30
40
50
60
70
80
Learning Objectives (LOs)
Grand Forks
Casper (x2.35)

E. (Tabular Summary of A-D)

	2016 Lecture (A)	2017 Lecture (B)	2016 Lab (C)	2017 Lab (D)
Correlation of Grand Forks to Casper students struggling with same LOs:	r = 0.89 (strong) p < 0.001 (significant)	r = 0.88 (strong) p < 0.001 (significant)	r = 0.63 (strong) p < 0.001 (significant)	r = 0.64 (strong) p < 0.001 (significant)
Number of LOs not retested by either population:	6 / 125 (4.8%)	6 / 125 (4.8%)	9 / 125 (7.2%)	9 / 125 (7.2%)
Number of LOs retested by only one population:	13 / 125 (10.4%)	11 / 125 (8.8%)	9 / 125 (7.2%)	17 / 125 (13.6%)
Number of LOs retested by both populations:	106 / 125 (84.8%)	108 / 125 (86.4%)	107 / 125 (85.6%)	99 / 125 (79.2%)
-Mean difference in retesting frequencies:	4.64 ± 3.91 times retested	5.03 ± 4.48 times retested	8.14 ± 7.78 times retested	9.41 ± 7.74 times retested
-Number of LOs retested by two-fold difference:	20 / 106 (18.9%)	29 / 108 (26.9%)	39 / 107 (36.4%)	42 / 99 (42.4%)
-Mean difference in retesting frequencies:	7.21 ± 5.10 times retested	7.40 ± 6.24 times retested	12.51 ± 9.40 times retested	13.90 ± 7.98 times retested

Figure III-8. Site- and Exam Type- Specific IRQ LO Retesting Frequencies. Retesting frequencies for each of the 125 course LOs are presented as those from Grand Forks vs. Casper for 2016 and 2017 lecture exams (A and B, respectively) as well as 2016 and 2017 lab exams (C and D, respectively). LO retesting frequencies for Casper populations were multiplied by 2.56 in 2016 and 2.35 in 2017 to equally compare respective Grand Forks vs. Casper populations and their resulting LO retesting frequencies. A summary of the findings are presented in tabular form (E). According to strong (r = 0.63 to 0.89), significant (p < 0.001) correlation analyses, Grand Forks and Casper populations were found to struggle with generally the same LOs in both lecture and lab settings. As such, most LOs were retested by both populations. However, significantly more LOs were retested between sites at very different frequencies for lab exams. This is likely due to having same lecture environments and exams but different (but comparable) lab environments and exams.