8-3-2023

Cumulative Advantage and Student Performance in General Chemistry: An Empirical Study at a Midwest Tier 1 University (2016-2019)

Nicole Barvian
University of Missouri-St. Louis, ncbh3r@umsystem.edu

Cumulative Advantage and Student Performance in General Chemistry: An Empirical Study at a Midwest Tier 1 University (2016-2019)

Nicole C. Barvian
B.S. Chemistry; B. A. Biology, Regis University, Denver 1991
M.Ed. Secondary Education, University of Missouri-St. Louis, 2023

Dissertation
The Graduate School at the University of Missouri-St. Louis
in partial fulfillment of the requirements for the degree
Doctor of Education with an emphasis in Educational Practice

December
2023

Dissertation Committee

Keith W. Miller, Ph.D.
Chairperson

Charles R. Granger, Ph.D.

Helene Sherman, Ed.D.

Abstract

General chemistry (GC) plays a crucial role in the academic and career outcomes of students interested in postsecondary science, engineering and medicine. As a whole, chemistry education research literature has reported student attrition rates averaging from 25% to 35% for nearly a century. Student attrition rates can be attributed to a number of factors, including high school preparation (e.g. Tai et al. 2006), opportunity to learn (Carroll, 1989) and accumulated time and practice in a domain (e.g. Ericsson et al. 1993). The factors, in turn, are affected by mechanisms of social stratification, in particular, the social mechanism of cumulative advantage (Merton, 1968, 1988). In the context of education, cumulative advantage is evident in the form of achievement gaps, which result from the accumulation of knowledge by one group over another across K-12 education (e.g. DiPrete and Eirich, 2006). Empirical performance data was analyzed using exploratory data analysis, null hypothesis significance testing and modeling to infer whether a process of cumulative advantage was present. Results of the analyses support the hypothesis that differential high school preparation could be, at least partially responsible for GC student outcomes. The implications for practice are discussed and broad suggestions for promoting equity are described.

Key Words: General chemistry, introductory college chemistry, educational research, chemistry education research, cognitive science, achievement gaps, cumulative advantage.

Dedications

I wish to thank my committee members, Keith W. Miller, Charles R. Granger and Helene Sherman for their help and support. Without their input, I would not have been able to produce the best version of this work. Likewise, the data analysis was made possible by an anonymous donor who graciously shared their deidentified data.

It is unlikely I would have ever majored in chemistry without the support of my college mentors, Professor Surendra Mahapatro and the late Father William T. Miller, S. J. of Regis University. Neither ever suggested I did not belong in chemistry, for which I will be forever grateful.

I am grateful for the love and support of my family, particularly my husband Mark. It was he who ran our household and organized our own college students thus allowing me to focus on this work. Last, but not least, Melissa Clarke and Roxane McWilliams provided me with invaluable support. If either one grew weary of listening to my musings, they at least never said so.

Again, my humble thanks are extended to each and every one of you.

Table of Contents

Chapter 1: Introduction

Background

One of the first foundational courses taken by science and engineering majors is an introductory chemistry course. Chemistry is also mandatory for any student applying to medical school, regardless of major. The course is known as general chemistry or introductory college chemistry, depending on the institution. Traditionally, all enrolled students have been expected to perform as though they were chemistry majors. The term general chemistry (GC) was preferentially used here, thought to be more descriptive of that practice.

For the past century, chemistry faculty have reported GC attrition rates ranging from 25% - 35% (e.g. Cornog & Stoddard, 1925; Hovey & Krohn, 1958; Martin, 1942; McFate & Olmstead, 1999; Ye et al., 2016). Student attrition in this context is defined as the DFWI rate, the combined number of student withdrawals, course incompletes, and grades of D or F. Koch and Drake (2018) reported an average DFWI rate of 29.4% across 31 institutions (p. 1). Of those students, only 25.9% remained enrolled in the same institution a year later (p. 4).

Student performance has traditionally been attributed to natural student aptitude rather than acquired knowledge and subsequent practice. For example, Cornog and Stoddard (1925) described “eliminating” 25% to 33% of enrolled GC students using aptitude testing (p. 707). Ralph and Lewis (2018) used SAT-mathematics scores as a proxy aptitude test and concluded that students scoring in the lowest quartile had “low math aptitude” (p. 870) with poor chances of passing GC.

Students believed to be unsuited for GC have often been described in stereotypical ways. Martin (1942) suggested low performing students were suffering from “hypothyroidism or another abnormal physiological condition” (p. 277). Kotnik (1974) described underprepared students as having “psychological problems...[and] even an extremely antisocial outlook” (p.

165). Spencer (1996) argued that Black and Latino students naturally received the lowest chemistry scores based on mathematical aptitude (p. 1152).

It is unclear what percentage of high schools ever offered chemistry to their students. Early comments in the chemistry education literature suggest many high schools did not off the course (e.g. Hovey & Krohn, 1958; Martin, 1942). Even for students with access to high school chemistry, the quality of the course was questioned (e.g. Coley, 1973; Kotnik, 1974). Recently, the National Science Board (2018) indicated that only about three-quarters of American high schools offer chemistry (p. 148).

A schism developed between high school chemistry teachers and postsecondary chemistry faculty which was described as a "cold war" (Rosen, 1956, p. 322). This arguably made the transition from high school chemistry to GC even more difficult for a student. Streitberger (1977) remarked that "...there is little evidence that high school chemistry is doing its part to insure (*sic*) that students are informed" (p. 1977).

Statement of the Problem

Arguably, many of the issues plaguing chemistry education are related to traditional beliefs and practices. Bowen and Cooper (2022) described "embedded harmful and inequitable practices" based on "tradition and assumptions" (p. 185) within chemistry education. One such practice, the "weeding-out" of students from GC has been described as problematic since at least 1953 (Barr et al. 2010, p. 53). Weston et al. (2019) described weed-out courses as being intentionally structured towards "getting rid of a high proportion of students" (p. 214).

According to Ferrare and Miller (2019) a significant portion, 26% of chemistry faculty, admitted to believing that chemistry ability was innate rather than acquired (p. 127). Notably, this view *was not* held by biology and physics instructors in the same study. These beliefs may have been responsible for the tepid response to the discovery that students did not learn chemistry conceptually by solving problems (e.g. Nurrenbern and Pickering, 1987; Pickering, 1990; Sawrey, 1990).

As noted by Ryu et al. (2021) relevant chemistry education research is published outside of the two major chemistry education journals (p. 3626). For example, only one study (e.g. Tai et al. 2006) regarding the importance of high school coursework on later GC performance was published in the *Journal of Chemical Education*. Most other studies (e.g. Tai et al. 2005; Tai and Sadler; 2007; Sadler et al. 2014) were published in more select science education journals.

Conceptual Framework

In 2012, the National Research Council (NRC) issued a report describing discipline-based education research (DBER) in science and engineering. The report went on to state that:

> A long-term goal of DBER is to understand how people learn science and engineering in order to improve learning and teaching. Research that advances this goal must be grounded in an understanding of what it means to develop expertise in a discipline and the challenges inherent in developing that expertise. (p. 189).

What it means to "develop expertise" naturally differs by subject. It would be ill-founded to presume that just because a subject fell under the umbrella of science, technology, engineering and mathematics (STEM), the same pedagogical concerns were shared by each.

As chemistry is a mature science, students cannot construct their own understanding of the subject. Since chemistry occurs unseen and is described by models based on empirical data the student must rely heavily on what is learned in school. Scerri (2003) described learning science as "reaching a position where...enough of the shared, and temporarily accepted, store of knowledge" (p. 472) had been obtained by a student. Consequently, it is reasonable to assume students with more high school preparation will be better prepared for GC.

For example, students entering GC can have no GC experience or up to two years of experience in high school. Mathematics preparation is also a factor for GC performance because of the deliberate emphasis placed on quantitative problem solving (e.g. Stowe et al., 2021; Shah et al., 2022). In general, students who have taken AP chemistry and calculus in high school tend to perform better in GC. These students usually have facile algebra skills, a better conceptual understanding of chemistry, and more stoichiometry practice (Tai, et al., 2005; Tai et al., 2006; Sadler & Tai 2007).

The acquisition of expertise in a domain is a well-known principle underlying cognitive learning theories (e.g. Ericsson et al. 1993; Glaser & Chi, 1988). Ericsson et al. (1993) proposed that the performance level in a complex domain such as chemistry was a "monotonic function of the amount of practice" (p. 367). They supported this claim with empirical evidence, suggesting that the idiom "practice makes perfect" could apply to chemistry education.

Multiple learning theories have considered time an important variable for learning. For example, Bloom (1968) proposed the idea that mastery of a subject depended on both the time needed to learn the subject and the available time given to the student. Bloom later demonstrated that students who were taught using a tutor outperformed students learning by conventional classroom methods (Bloom, 1984). The Carroll (1989) model of learning included a variable known as the "opportunity to learn" dictated by the time allowed for learning in a formal school setting.

Purpose

The idea that students do not have equal access to educational opportunities and resources appears to be self-evident. But recent chemistry education research studies tend to overlook high school preparation. However, such a stance can appear disingenuous, unless one believes students are born with a knowledge of chemistry. Stale arguments about nature versus nurture hardly seem applicable when considering a subject which is already inaccessible to about 25% of high school students (National Science Board, 2018, p. 148).

Educational resources, much like any other valued commodity, such as wages and healthcare, are stratified according to social status (e.g. Cheng, 2014; Hasl et al. 2022). Dornbusch et al. (1996) described the social structure of education as holding idealized beliefs in meritocracy. They described education as being simultaneously plagued by “social stratification, status attainment, credentialism, and the emphasis on ability differences” (p. 401).

The sociological theory of resource stratification and inequitable distribution is known as cumulative advantage (CA), after Merton (1968, 1988). The idea of cumulative advantage can be understood in terms of simple colloquialisms such as “the rich get richer” and “success breeds success” to name a few. Bowen et al. (2005) described CA in higher education as:

> “The accumulation of (often small) advantages and disadvantages over the course of the first 18 years of life that leads to massive preparation differences by the time of college application” (p. 225).

Cumulative advantage is a social mechanism, defined by Rigney (2010) using a schematic proposed by Hedstrom and Swedberg (1998, p. 9):

$$I \rightarrow M \rightarrow O$$

Where an input (I) is transformed by a social mechanism (M) into an output (O). DiPrete and Eirich (2006) described cumulative advantage as “a general mechanism for inequality across any temporal process...in which a favorable relative position becomes a resource that produces further relative gains” (p. 271).

Mathematically, the mechanism of cumulative advantage can be expressed as Equation 1.1, where the level of some resource Y_{it} depends on a previous level of that resource,

$Y_{i(t-1)}(1+\Upsilon)$ when ϒ > 0, Equation 1.1.

$$Y_{it} = Y_{i(t-1)}(1+\Upsilon) \qquad \text{Equation 1.1}$$

DiPrete and Eirich (2006) suggested that "...evidence for CA processes can take several forms" (p. 274) relating to the variables represented by Y_{it} and $Y_{i(t-1)}(1+\Upsilon)$. Three types of evidence in particular was suggested: (1) distributional analysis, (2) mathematical modeling, and (3) increasing variance between individual performance measures over time (p. 274-276).

For example, DiPrete and Eirich (2006) argued that CA processes produce skewed variable distributions mathematically characterized by power-law functions (p. 274). They also recommended the "estimation and testing" of mathematical models (p. 275) based on Equation 1.1 derived by Allison et al. (1982). Lastly, DiPrete and Eirich (2006) argued that growing inequality as a function of time should be demonstrable since the concept of CA relies on that assumption (p. 275).

A CA process is presumed to be present based on a preponderance of the evidence such as that described above. However, it is important to note that the full characterization of a CA process is considerably more complex. For example, Baumert et al. (2012) described the use of both latent growth curve analysis and quasi-simplex modeling to determine whether a CA process affected reading and mathematics development in elementary school (p. 1347). The importance of preliminary data justifying a need for more lengthy and time consuming analyses cannot be overstated.

Empirical data generated by students enrolled in GC at a medium-size Midwestern suburban Tier 1 research university was analyzed for evidence of CA. The variable $Y_{i(t-1)}(1+\Upsilon)$ served as a proxy measure of high school preparation; the independent variable represented by the Exam 1 score. The absolute knowledge gain (AKG) was defined as the sum of the last three course exams, and served as a proxy for the variable Y_{it} or the dependent variable.

Research Questions

Research questions were posed to answer the specific questions associated with the conceptual framework of CA as previously described.

R1: What did the shape and symmetry of the distribution of Exam 1 scores suggest when plotted as a histogram?

R2: How strong was the correlation between the proxy measure of high school preparation and the proxy measure of the dependent variable, AKG?

R3: Based on the presumed relationship between Exam 1 and AKG, to what extent could a predictive relationship be modeled?

R4: To what degree could the continuity or change in student rank from Exam 1 to AKG be demonstrated graphically?

Definition of Terms

- Chemistry Faculty: The term describes a faculty member tenured in the *chemistry department* rather than the college of education. The approach to educational research can, as a result, differ in philosophy and scope.
- Cognitive Learning Theory: An umbrella term for theories of learning originating from cognitive science. The entirety of the seminal research forms the basis of Ericsson et al. (1993), which is then elaborated. In order to understand the conceptual framework used here, one must presume that prior knowledge and accumulated practice are critical factors for the rate of new learning.
- Cumulative Advantage (CA): The general mechanism of resource stratification and subsequent distribution of resources based social status proposed by Merton (1968,

1988). Closely related to CA is the term Matthew effect, intended to describe CA on an individual level. The current convention is to use CA in all cases; exceptions are noted where applicable.

- Domain Specific Knowledge: Knowledge specific to a particular domain, such as the mole concept in chemistry. Simonsmeier et al. (2021) described domain specific knowledge as the "central component of competence, academic achievement, expertise, and similar cognitive learning outcomes" (p. 2 and references therein).
- Exploratory Data Analysis (EDA): Data analysis technique attributed to Tukey (e.g. 1962 and 1966) EDA employs mostly graphical techniques to reveal what form the underlying model might take according to the Engineering Statistical Handbook (NIST/SEMATECH, 2023).
- Mathematical Modeling: The construction of a deductive model in the form of a mathematical model to approximate a relationship or simulate the differences between variables and their behavior.
- Null Hypothesis Significance Testing (NHST): A method of statistical inference based on the rejection of a null hypothesis at a predetermined significance level.
- Social Stratification: Social stratification describes the inequitable access to resources based on social status, such as race, gender and socioeconomic status (e.g. Harris et al. 2020). For example, Kerckhoff and Glennie (1998) described the social stratification which affected high school status, curriculum rigor, achievement test scores and postsecondary educational attainment.
- Weed-Out Course: In a blog post, Zimmerman (2022) described weed-out courses as ones in which "universities let students sink or swim" and where a "certain fraction of students are expected to fail."

Procedures

Deidentified empirical student performance measures were analyzed using the conceptual framework of CA as described by DiPrete and Eirich (2006, p 274-275). They described a number of characteristic trends, patterns and variable relationships which could be detected or inferred from empirical data. The methods used to detect the expected characteristics were based on three different epistemological approaches: (1) traditional null hypothesis significance testing (NHST), (2) exploratory data analysis (e.g. Tukey, 1962, 1968) and (3) model development (e.g. de Mast et al. 2023; Mazur, 2006). These methods are described in further depth in Chapter 2 and their applications in Chapter 3.

Significance

The Next Generation Science Standards (National Research Council, 2012) include eight Science and Engineering Practices (SEPs). Their importance in the framework was described as follows:

> Students cannot fully understand scientific and engineering ideas without engaging in the practices of inquiry and the discourse by which such ideas are developed and refined. At the same time, they cannot learn or show competence in practices except in the context of specific content. (NRC Framework, 2012, p. 218)

The SEPs were designed to capture the different types of inquiry and data analysis procedures employed by scientists and engineers. In Appendix F (2013), the NRC stated that “engaging in scientific investigation requires not only skill but also knowledge that is specific to each practice” (NRC Framework, 2012, p. 30).

For instance, scientists and engineers frequently use causal inference models to understand the possible cause and effect relationships under their purview. Inference models are based on variables which can be identified in one or more of the following ways: exploratory data analysis, experiential knowledge, or derived from theory (Mast et al., 2023, p. 53). Analytical methods employed include graphical analysis, mathematical modeling and the use of existing conceptual frameworks. These methods are not to be confused with the null hypothesis significance testing which is frequently the sole method by which relationships are probed (e.g. Spencer, 1996, Ralph & Lewis, 2018).

Cohen (1995) remarked that "An error in elementary logic made frequently by NHST proponents...is the thoughtless, usually implicit, conclusion that if H_0 is rejected, then the theory is established" (p. 999). For example, a correlation between SAT-mathematics scores and GC performance is just that. Unless the variables have been fully explicated within a theoretical construct comprised of multiple perspectives, numerous threats to validity jeopardize any conclusions drawn (Mack et al. 2019, p. 404).

The novelty of the approach taken to analyze empirical student performance data was the utilization of the well-known conceptual framework of cumulative advantage. Since cumulative advantage can be used to explicate social stratification mechanisms in general, it was applied to differential high school preparation. Furthermore, the use of the conceptual framework was justified based on established theories of learning (e.g. Ericsson et al. 1993) and empirical evidence supporting conclusions about GC education (e.g. Tai et al. 2006; Sadler & Tai, 2007). All available analytical tools, exploratory analysis, NHST, mathematical modeling and graphical modeling were used to infer that students did not enter GC with the same degree of preparation.

Assumptions

- The conceptual framework of this study assumes knowledge is acquired, has intrinsic value, and is subject to social stratification.

- In accordance with Ericsson et al. (1993) and references therein, competency in a domain such as chemistry occurs over the course of years rather than weeks.
- It was presumed that the majority of students who enroll in GC do so with the intention of completing and passing the course.
- Laboratory activities tend to be highly structured and confirmatory in nature, and so are not considered relevant to the current discussion of chemistry education.

Limitations

- The findings of this study are confined to the institution from which the data was obtained. Researchers at other institutions are encouraged to characterize their own students' data similarly.
- The data collected and analyzed is representative of GC prior to the COVID-19 pandemic and school closures in March 2020.
- Data representing N = 409 students was analyzed to understand the distribution of the maximum number of students. However, the absolute knowledge gain could only be calculated for N = 343 students, since all four exam scores were required in that analysis.
- The data only represented students who remained enrolled in the course; the number of withdrawals was unknown.
- The use of mathematical models related to cumulative advantage inferred that CA explained all cases of poor student performance, an unlikely scenario. However, the error from that assumption was considered smaller than the error from the presumption that some 25% of students are unable to grasp the subject of chemistry.

Delimitations

- While some archived chemistry performance related data was available as far back as 2012, changes made to ACS chemistry exams, high school AP chemistry courses, and state chemistry standards made measures prior to 2016 potentially irrelevant to the present day.
- Student data was used as is, in the form of absolute measurement scores. Consequently, there is insufficient data available from which to draw conclusions concerning DFW rates or grading practices.
- The data used was deliberately chosen as a more suitable indicator of chemistry preparation than ACT or SAT scores and high school grade point average.

Organization

The organization of the dissertation follows a standard five-chapter format. Chapter one is an overview of the study placed within a theoretical framework. The research questions, null hypotheses, methodology and significance of the study are briefly described here. Chapter two contains the literature review and Chapter three describes the research design, and proposed methodology. Chapter four describes the results of the study and Chapter five considers the implications of the results and provides suggestions for further research.

Summary of Chapter One

Historically, chemistry ability has been understood in terms of individual inherited capacity rather than the consequence of knowledge and skill acquisition. The social stratification of GC education, specifically the inequality associated with course availability has remained largely hidden. The social mechanism of CA was used as a conceptual framework with which to interpret empirical student performance data. Several available analysis methods were used to probe, model and predict the likely effects of CA of educational attainment in GC.

Chapter 2: Review of the Literature

An internet search using the terms "most difficult college major" finds chemistry often listed as number one, or at least among the top five, based on a typical grade point average of about 2.8 (e.g. Conlin, 2023). Low grade point averages are indicative of the persistent struggle of students to meet GC faculty expectations (Weston et al., 2019). Chemistry faculty seem especially prone to beliefs that chemistry ability is inherited. According to Ferrare and Miller (2019) over one-quarter of chemistry faculty admitted believing in genetically determined chemistry ability, a belief not shared by their fellow biology and physics educators (p. 125).

These beliefs may persist as a remnant of constructivist philosophy which was introduced into chemistry education research in the late 1980s (Bodner, 1986; Herron & Greenbowe, 1986). Herron proposed that GC ability was the consequence of the transition to Piaget's formal operational stage (Herron, 1975 and 1978). The future of students presumed to be at the concrete operational stage was considered grim. Bodner (1986) suggesting that "Teaching and learning are not synonymous; we can teach, and teach well, without having the students learn" (p. 873).

Conversely, cognitive scientists proposed that the science of learning could inform the practice of education (Anderson et al. 1998). The fields of cognitive science and educational psychology generally agree that knowledge is acquired and accumulates over time (Chi & Glaser, 1988; Ericsson et al. 1993; Sweller et al. 1998). Since GC educators have traditionally complained about underprepared students (Herron & Greenbowe, 1986; Hovey & Krohn, 1958; Kotnik, 1974) it should be worthwhile to consider the time involved in learning chemistry.

The understanding that learning chemistry might take more than one year in high school could account for the achievement gaps and differential performance between different GC student groups (e.g. Harris et al., 2020; Ryu et al., 2021). Harris et al. (2020) calculated achievement gaps for over 25,000 students, reporting values ranging from 0.12 to 0.54 on a 4.0 scale for various student groups (p. 1). For students from minoritized groups, these gaps represented the effects of educational inequality. In other words, all students do not enter GC with similar experiences.

Ryu et al. (2021) suggested that many of the students who experienced educational inequality go on to experience educational inequity as well (p. 3621). Meaning that all students do not have the same access to educational opportunities. They noted that "...academic and psychological characteristics of historically minoritized students are not causes of the achievement and retention gaps" (3627). Rather, they suggested researchers look for root causes of the academic and psychological differences they ascribed to factors at the K-12 level.

Search of the Literature

Two chemistry education research journals, the *Journal of Chemical Education* (JCE) and *Chemistry Education Research and Practice* (CERP) were consulted. *JCE* in particular was used as a historical source since the journal has been published since 1924. However, as noted by Ryu et al. (2021), specific information regarding student attrition rates needed to be obtained elsewhere. A comprehensive historical review of chemistry education research was found in *Chemical Reviews* (Cooper & Stowe, 2018) which provided a list of other relevant articles.

Other sources included the *American Educational Research Journal, Journal of Research in Science Teaching, Journal of Educational Research, Research in Higher Education Journal, Science Advances and Science Educator.* Since physics and chemistry together comprise the physical sciences, literature from that venue was considered as well. Sources included *Physical Review of Physics Education* and *Contemporary Physics*. Much of the seminal work in cognitive science and educational psychology used physics as the focal point of empirical studies (e.g. Chi et al. 1981; Larkin et al., 1980). Much of that work has been collected in compilations such as

Cognitive Skills and Their Acquisition, *Carnegie Mellon Symposia on Cognition Series* edited by John R. Anderson (1981) and *The Nature of Expertise*, edited by Chi, et al. (1988).

This literature review also focused heavily on cognitive and educational psychology which included *Annual Review of Psychology*, *Educational Psychologist,* and *Educational Psychology Review*. Articles related to the sociological aspects of education were found in *American Journal of Sociology, Annual Review of Sociology, Journal of Political Economy* and *Sociological Inquiry.* In addition, primary sources regarding research and research designs were consulted including *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Shadish et al., 2002) and an earlier edition by Cook and Campbell (1979). Sources describing nonparametric statistics referred to *Nonparametric Statistics for the Behavioral Sciences* (Siegel & Castellan Jr., 1988), were also consulted.

Background

Student ability is often framed within one of two conceptual frameworks in the chemistry education research literature. Ferrare and Miller (2019) described a cognitive frame based on perceived student value, the “individual ability frame” (p. 123). Individual student ability was described as “the main driver of success in a context of objectivity with little to no mention of external constraints” (p. 122).

Bensimon (2005) described a “deficit cognitive frame” (p. 102) which described student deficiencies as dictating GC performance. The deficit cognitive frame was associated with the use of characteristic stereotypes related to different cultures and socioeconomic status. Stereotypical beliefs lead to discourse revolving around strategies meant to “fix” deficiencies in student affect, study skills or cognition (p. 103). Despite the efforts of well-intentioned educators, deficits continued to be used as a way to justify differential student performance.

The focus of chemistry education research on the presumed deficiencies of GC students is evidenced by the description of countless diagnostic tests described in the literature. A small sampling of diagnostic tests are listed in Table 2.1. Notably, this stance is at odds with scientific training, which features the concept of Occam's razor (Duignan, 2023). Occam's razor suggests that the simplest answer is generally the correct one. Arguably, then, student performance in GC should be considered a consequence of high school preparation, or lack thereof.

Table 2.1

Diagnostic Exam Performance Variables

Diagnostic Exam	**Example**
Chemistry Aptitude and Training	Cornog and Stoddard (1925)
Toledo Chemistry Placement Exam	Hovey and Krohn (1958, 1963)
ACT Scores	Coley (1973)
Piagetian Tasks	Albanese et al. (1976)
SAT Scores	Ozsogomonyan and Loftus (1979)
Classroom Test of Scientific Reasoning	Lawson (1978)
Logical Reasoning Ability	Bunce and Hutchinson (1993)
Fullerton Chemistry Placement Exam	McFate and Olmstead III (1999)
Student Pre-Semester Assessment	Wagner et al. (2002)
Alternative Conceptions Inventory	Mulford and Robinson (2002)
Chemistry Self-Concept Inventory	Bauer (2005)
First Semester Exam	Mills et al. (2009)
Group Assessment of Logical Thinking	Bird (2010)
Neural Network Diagnostic Algebra Test	Cooper and Pearson (2012)
California Chemistry Diagnostic Test	Legg et al. (2001)
Domain Specific Mindset	Santos et al. (2021)
Mathematics Automaticity	Shelton et al. (2021)
Social-Psychological Interventions	Wang et al. (2021)

Multiple studies have suggested that high school status and curriculum rigor are predictive of later GC performance. For example, a direct causal relationship between rigorous mathematics and educational attainment in general has been demonstrated (e.g., Irvin et al., 2017; Long et al., 2012). This in turn, suggests that accessibility to advanced mathematics coursework is a factor for later educational performance (e.g., An, 2022; Morgan et al., 2018; Sadler & Sonnert, 2018).

In a highly cited paper, Tai et al. (2005) investigated the influence of high school coursework on GC performance. Freshman science and engineering students, *N* = 1531 from 12 institutions of higher learning, reported their previous high school preparation. Nearly 52% of the students reported taking calculus, while 44% reported taking either AP or honors chemistry in high school (p. 1000). The better prepared students out-performed their peers with less experience. The multiple regression model constructed predicted that the last high school mathematics grade and SAT-Mathematics scores were significant predictors of GC performance.

A second multiple regression model reported by Tai et al. (2006) further explicated the importance of high school chemistry topics for later GC performance. Facile stoichiometry skills were a strong predictor of GC performance, with “recurring topic” versus “none at all” having standardized β-coefficients of .10 and −.08 respectively (p. 1707). The model also suggested that certain subjects, such as gas laws or nuclear reactions, had little to no value for later GC performance (p. 1707).

In both regression models, reported by Tai et al. (2005 and 2006), completing AP calculus in high school was the greatest single predictor of later GC performance. This was true whether a student had gone on to take the AP exam or not. In fact, mathematics was so important that the standardized β-coefficient was large enough to negate factors related to race, ethnicity, and parent education level (2006, p. 1707).

Tai et al. (2005) summarized the importance of mathematics to GC performance thusly:

> Although the advanced topics of calculus are not directly utilized in GC, calculus builds students’ facility with algebraic functions, graph interpretation (including slope), mental computation, and calculation of rates of change. It appears that the virtually automatic

facility with such mathematical skills that a successful calculus background bestows on students removes many impediments to understanding the quantitative aspects of chemistry that other students must endure (p. 1003-1004).

Students who had only taken algebra or trigonometry in high school did not fare as well in college as students who had taken calculus (e.g., Tai & Sadler, 2007; Maltese & Tai, 2011). Other researchers suggested the multiple reviews of algebra leading up to calculus explained performance (e.g., White & Mitchelmore, 1996). Still more studies suggested that the number of years of high mathematics influenced mathematical fluency and procedural task automation (e.g., Shelton et al., 2021; Powell et al., 2020).

However, according to the National Science Board (2018), only 76% and 56% of high schools offered chemistry and calculus, respectively. These values dropped to 65% and 33% in schools where greater than 75% of the enrolled students were Black or Latino (National Science Board, 2018, p. 148). The National Science Board also reported that, in general, only about 19% of students finish high school having taken calculus, the majority from the highest quintile in socioeconomic status (2018, p. 62).

Many rural students attend small, underfunded schools and lack access to advanced coursework as well (e.g., Irvin et al., 2017; Wells et al., 2019). According to Irvin et al. (2017) rural students were least likely to have access to advanced mathematics coursework; only 12% reportedly have taken calculus (p. 478). Orfield (2009) wrote that, for many students "there is no way to get the right preparation in their school regardless of their personal talent and motivation" (p. 4).

Theoretical Framework

In postsecondary chemistry education, chemistry faculty are trained first as scientists. Versed in the scientific method researchers should understand good research practice and include appropriate study controls. Unfortunately, many chemistry education research studies do not account for variables such as high school preparation (e.g. Mack et al. 2019). From a research point of view, the failure to equate students prior to the introduction of an educational intervention is a significant threat to the validity of the study results (Shadish et al. 2002).

Chemistry is not a subject which can be intuited from everyday life and so must be learned within the classroom (e.g. Fredrick and Walberg, 1980). Students cannot construct their own understanding of chemistry, according to philosopher and chemist Eric Scerri (2003). Scerri noted that chemistry was a “mature science” (p. 470), meaning:

> The process of learning science, unlike any other field, involves reaching a position where the student has understood enough of the shared, and temporarily accepted, store of knowledge so that he or she can communicate with others and even make contributions to the general scientific consensus. (p. 472)

How much of that useful knowledge is known by any particular GC student is not entirely clear. Assessment practices such as grading on a curve or norm referencing can further obfuscate differences in student preparation (e.g. Bowen & Cooper, 2022; Weston et al. 2019). Sadler and Tai (2007) reported their surprise at discovering that a significant number of students enrolled in GC had already successfully completed AP chemistry in high school. They suggested that:

> The population of students who have taken AP courses in high school and retake introductory courses has been largely neglected by researchers. Researchers have rarely structured past studies to reveal the degree to which AP courses bestow an added benefit upon students who take them. (p. 3)

Traditionally, measures intended to help predict postsecondary student performance have been used as proxy intelligence tests by chemistry faculty (e.g. Andrews & Andrews, 1979;

Ozsogomonyan & Loftus, 1979). For example, Ralph and Lewis (2018 and 2019) reported using SAT-mathematics scores as proxy intelligence tests. They concluded that “at-risk” chemistry students were those who had scored in the lowest SAT-mathematics quartile and had “low math aptitude” (Ralph & Lewis, 2019, p. 570).

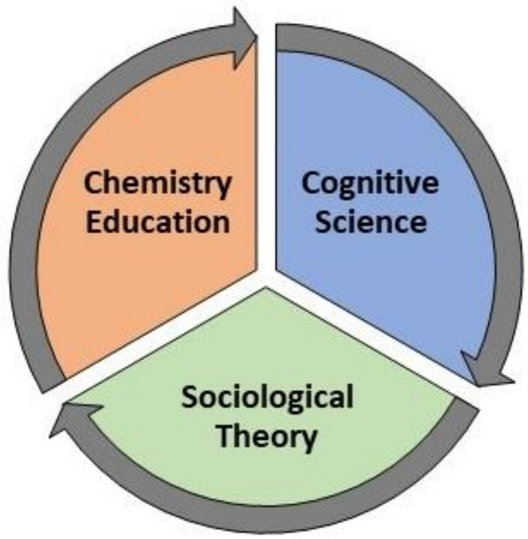
The use of the term aptitude, according to the *Washington Post* (1993) suggested individuals had “a natural talent” or “capacity” (Merriam-Webster, 2023). In 1990, a commission appointed by the College Board deemed the SAT an achievement rather than an aptitude test. As a result, the word was stricken from the title of the test in 1994 (Jordan and Achenbach, 1994).

These issues raised concerns about research based on “personal empiricism” or personal opinion in chemistry education research (Cooper & Stowe, 2018). Bowen and Cooper (2022) suggested that “there are still practices in chemistry education that arguably do not embody informed practice and have remained entrenched despite evidence of their harm” (p. 185). In a previous editorial, grading on a curve was described as “educational malpractice” (Cooper & Klymkowsky, 2020).

The conceptual framework of this research was tripartite in nature, as shown in the schematic diagram in Figure 2.1. Three closely related aspects to GC education were considered: (1) modern theories of learning based on cognitive science (e.g. NRC, 2012; Ericsson et al., 1993), (2) historical and contemporary descriptions of student sorting described in the chemistry education literature (e.g. Cornog & Stoddard, 1925; Hovey & Krohn, 1958; Spencer, 1996; Ralph & Lewis, 2018), and (3) the social repercussions of educational resource stratification and educational attainment (e.g. Boliver & Schindler, 2021; Dornbusch et al. 1996).

Figure 2.1

Tripartite Conceptual Framework



Cognitive Science

In 2012, the National Research Council (NRC) issued a report defining discipline-based education research (DBER) in science and engineering. The conceptual framework used to understand DBER was based on the concept of "expert-novice differences" (p. 58) specifically citing a seminal publication by Ericsson et al. (1993). In doing so, the NRC endorsed cognitive learning theory within the context of postsecondary science education. Furthermore, they agreed with Ericsson et al. that expertise lay on a continuum, which could span ten years or more between novice and expert status (p. 366).

The expert-novice framework referred to by the NRC was based on decades of work describing cognition, expertise and performance. Glaser and Chi (1988) summarized the paradigm by characterizing experts as possessing the following:

1. Large stores of organized domain-specific knowledge in long-term memory.
2. A deep conceptual understanding of the domain and associated problem types.

3. Automated procedural skills which allowed tasks to be performed without thought.
4. The ability to employ metacognitive skills, such as repeatedly checking their work. (Glaser & Chi, 1988, p. xvii-xx).

Novice problem solvers, on the other hand, were said to rely on approaches to problem-solving involving trial and error (Chi et al., 1981; Davenport et al. 2008; Ericsson, 2017). Novices are believed to benefit from guidance when planning and executing multi-step problems (e.g., Davenport et al. 2008; Chi et al. 1981; Nakhleh & Mitchell, 1993). Novice problem solvers, usually overwhelmed with information, often cannot employ metacognitive strategies, making them appear sloppy and disorganized (e.g. Davenport et al. 2008; Talanquer, 2006).

Ericsson et al. (1993) further elaborated on the expert-novice paradigm by recognizing that “the effect of practice on performance is larger than earlier believed possible” (p. 363). They suggested that the factors of “zeal and labor” had been minimized in favor of “genetic factors” (p. 364). They proposed and provided empirical data to support the claim that individual performance was a “monotonic function of the amount of accumulated practice” (p. 367). In other words, while the magnitude of acquired knowledge at any given time might fluctuate, on average practice improved performance (e.g. Ericsson & Charness, 1994, Ericsson & Kintsch, 1995).

The time needed to acquire expertise was predicted to be at least ten years based on previous work by Simon and Chase (1973). Simon and Chase described chess masters as taking ten or more years of intense practice to achieve their status (p. 402). In addition to the time needed for practice, Ericsson (2017) also suggested time was needed for “...cognitive, perceptual, physiological, neurological and anatomical changes...” (p. 1).

Ericsson et al. (1993) also specified the parameters of what was considered “deliberate” or useful practice. Deliberate practice, as described, placed responsibility for student progress on both the instructor and learner. For example, the instructor was responsible for tailoring tasks based on preexisting learner knowledge and for providing immediate feedback. The learner was expected to consciously engage in practice even at times when it was not “inherently motivating” (p. 368 - 369). In addition, there were daily limits to the effectiveness of practice. Beyond a time limit of two to four hours a day, the risk of “reduced benefits” became problematic (p. 370).

General Chemistry Performance

As described, learning a complex subject such as chemistry required the acquisition, processing, encoding and organization of new knowledge into memory. Chemistry students needed to acquire the domain-specific declarative knowledge, the “what” of chemistry and the “how” of solving chemistry problems (e.g. Anderson, 1982, 1987; de Jong & Ferguson-Hessler, 1996). Extensive practice of procedural skills were required to automate skilled tasks such as stoichiometry.

The temporal aspects to GC education included (1) time engaged in the classroom, (2) time spent restructuring prior knowledge (e.g., Chi et al., 1981; Glaser & Chi, 1988), (3) time needed for attendant neurological changes (e.g., Brod et al., 2013), (4) time needed to develop procedural skills (e.g., Anderson et al., 2019) and (5) time needed to automate procedural skills (e.g., Ericsson et al., 1993). Considered within that context, a student with a second year of chemistry in high were expected to transition better into GC.

At issue was not whether students could develop the cognitive skills needed for GC performance. Instead, a crucial issue was whether students had the time to do so, particularly if practice was limited to two to four hours a day (Ericsson et al. 1993). The time spent engaged with the course itself could be problematic, since GC is typically a 5-credit hour course. A typical student can spend three hours in lecture, three hours in recitation, and three hours in the laboratory every week. When the extra time needed to complete homework, study, and work was added in, mental exhaustion arguably would become a very real threat.

Just how much time might be needed to overcome weak science and mathematics preparation in high school was suggested in a study by Shah et al. (2020). They reported that such students benefitted from supplemental mathematics support in the first semester of GC. However, when the students matriculated back into GC in the second semester, gaps in mathematics preparation were even more pronounced (p. 1822).

It is worth noting that the emphasis placed on stoichiometry and quantitative problem-solving ability is based on traditional institutional practices. From the earliest accounts, chemistry educators have used mathematical background as a way to sort students (Barr et al., 2010; Bowen & Cooper, 2022). Shah et al. (2022) noted that quantitative problem solving was the “gold standard” (p. 22) for determining GC student worth. Ralph et al. (2022) described the focus on problem solving as marginalizing “students based on their access to pre-college math preparation [and misrepresenting] the intellectual work of chemistry” (p. 1870).

Sociological Considerations

Robert K. Merton, a sociologist at Columbia University first defined the characteristic roles, norms and mores which defined science as a social institution (1942/1973). The goal of science, according to Merton, was the “extension of knowledge” certified through empirical evidence and logical consistency (p. 270). Merton described four scientific imperatives, one of which was universalism. Universalism requires a scientist objectively consider “truth-claims... to preestablished impersonal criteria” (p. 270).

Merton (1968) first proposed a mechanism for the stratification of resources and rewards in science research. The majority of resources were awarded to a select number of scientists based on status or prestige. The same phenomenon was described at the institutional level in prestigious research institutions. Institutions such as Harvard, Berkeley and the California Institute of Technology were described as attracting more graduate students, producing more Nobel laureates and receiving more research funding. The result was a self-reinforcing cycle of accumulating benefits and advantage (Merton, 1968, p. 62).

Merton (1988) later described science education as being “so organized as to put a premium on relatively early manifestations of ability − in a word, on precocity” (p. 613). Merton referred to the early sorting, choosing, and labeling of college students as being “gifted” or “having what it takes” in science as college freshman. The consequences of this sorting process favored one group of students while the other was “cut off from support and response” (p. 614).

Merton first used the term Mathew effect (1968) initially to distinguish between individual and population-based resource stratification, labeled as cumulative advantage. Hower, the ubiquitous nature of social resource stratification led to the adoption of the more general term cumulative advantage (CA). Merton (1988) summarized all aspects of cumulative advantage (CA) in the following summary:

> Individual self-selection and institutional social selection interact to affect successive probabilities of being variously located in the opportunity structure of science. And it is such unanticipated and unintended consequences of purposive social action - that tend to persist. (p. 615 - 616).

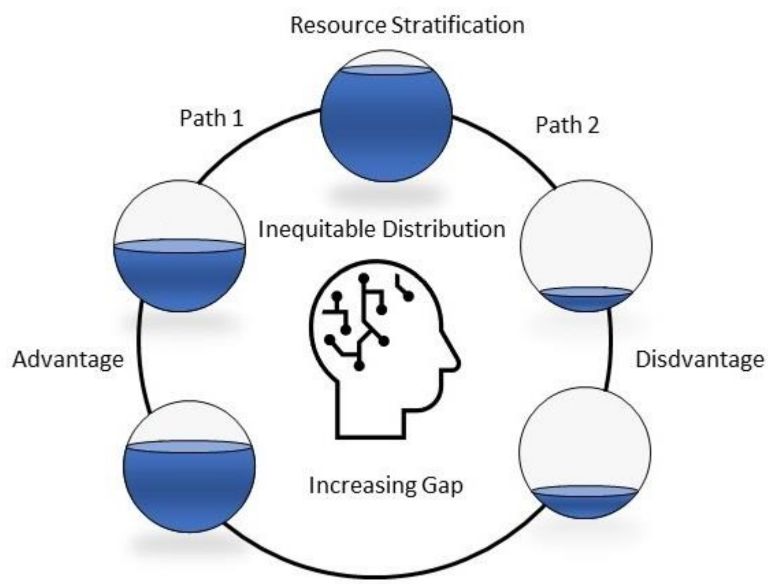
However, the term Matthew effect is sometimes found within the education literature after Walberg and Tsai (1983) who first used the term in that context. In a more recent review, DiPrete and Eirich (2006) expanded the scope of what was considered (CA). They called for more attention by researchers on the theory, mechanisms, and methodology of the phenomenon (p.

292). They used the term CA exclusively and that convention was practiced here as well. Only in the researcher has specifically used the term Matthew effect is the term substituted for CA.

The most basic description of CA describes resource stratification and accumulation of some valued resource by one individual or group. Further acquisition of the resource is then proportional to the initial amount. Over time, the gap between the two individuals or groups exacerbates the initial gap in resources, Figure 2.2. In terms of education, in general, valuable resources might include time, knowledge and cognitive growth. All of those factors can exacerbate known achievement gaps between individuals and groups (e.g. Bensimon, 2005; Ceci & Papierno, 2005; Reardon, 2011).

Figure 2.2

Schematic Representation of Cumulative Advantage as Knowledge Acquisition



Ericsson et al. (1993) suggested that “the highest levels of performance and achievement appear to require at least around 10 years of intense preparation” (p. 366). While GC students never approach expertise, the more experience a student has, the better. Consequently, students with more high school preparation are likely to perform better in GC.

The importance of time as a factor in learning has been recognized in other learning theories. For example, the Carroll Model (Carroll, 1989) and the Learning for Mastery Model proposed by Bloom (1968) each contained time dependent variables. Fredrick and Walberg (1980) described time as "invariably operating" on learning and student achievement (p. 193).

Carroll (1989) included a variable he described as the "opportunity to learn" (OTL), defined as "the amount of time allowed for learning...by a school schedule or program" (p. 26). Grodsky et al. (2008) suggested OTL was a critical factor in educational resource stratification. They based their viewpoint on the fact that students with more OTL generally receive lower scores of achievement tests such as the ACT and SAT (p. 388).

Data Analysis Considerations

The methodology used in this research was comprised in part of the practices used by scientists and engineers in industry. The importance of understanding the practical applications of science was reflected in *A Science Framework for K-12 Science Education* (National Research Council, 2012). The resulting Next Generation Science Standards included eight science and engineering practices (SEPs) meant to contextualize scientific inquiry in practical applications. Three of the eight SEPs were specifically applied in this study and are described in detail in Figure 2.3.

Developing and Using Models, SEP Practice 2, is a fundamental part of the physical sciences, since they are largely abstract. For example, chemistry reactions which occur on the macroscopic scale are modeled at the submicroscopic scale, as well as being described using symbolic representations. This aspect of chemistry has long been known to be confusing to students (Johnstone, 2006).

Figure 2.3

Example of NGSS Science and Engineering Practices Used in the Analysis of Empirical Data

Practice 2: Developing and Using Models	**Practice 4: Analyzing and Interpreting Data**	**Practice 5: Using Mathematics and Computational Thinking**
Models represent ideas and explanations through the use of diagrams, drawings, mathematical representations.	Investigations produce data in which trends and patterns can be identified using tabulation, graphical and visual interpretation, and statistical analysis.	Mathematics and computational thinking are tools for representing variables and relationships between them. These include statistical analysis and mathematical expressions.

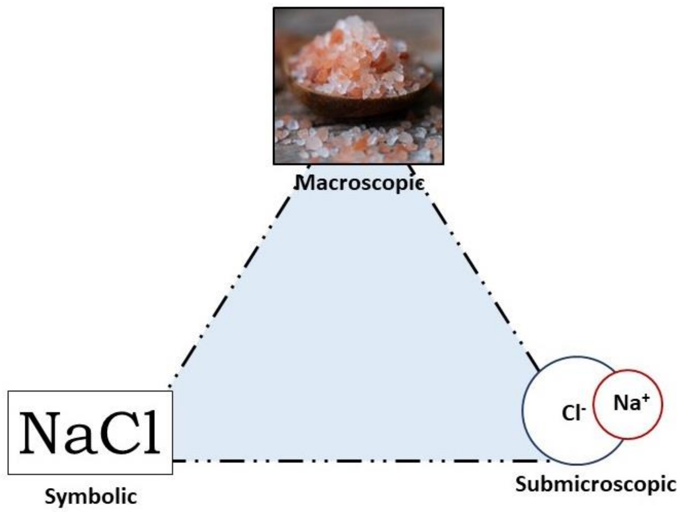
Note: Adapted from the National Research Council Framework for K-12 Science Education, 2013, Appendix F, p. 6-10.

The relationship between the three representations of chemistry was conceptualized as a model known as Johnstone's Triangle, after Alex H. Johnstone of the University of Glasgow (Johnstone, 2010, p. 24; Cooper and Stowe, p. 6058). While the reader does not necessarily need understand the concept, it is mentioned obliquely and so is provided for clarity. Figure 2.3 shows the "Triangle" for table salt, NaCl, as conceptualized by the author.

Students are required to understand that the macroscopic form of a compound is modeled at the submicroscopic level. The chemical reaction is communicated via the symbolic representation. Any individual working with chemical reactions must be able to move seamlessly between the three concepts. As an aside, note that the pinkish color to the salt shown is the result of trace impurities in the solid.

Figure 2.3

Johnstone's Triangle of Table Salt



The analysis of chemical phenomena using mathematical models is stoichiometry an example of which is shown in Equation 2.1. The reaction in question is the combustion of ethane gas to produce carbon dioxide and water.

$$2\ CH_2H_6(g) + 7\ O_2(g) \rightarrow 4\ CO_2(g) + 6\ H_2O(g) \qquad \text{Equation 2.1}$$

The equation demonstrates the relationships between variables and the numerical coefficients are balanced. Consequently, the equation as written can be used to test hypotheses such as:

H_0: The combustion of 4 moles of ethane produces 12 moles of water.

H_a: The combustion of 4 moles of ethane does not produce 12 moles of water.

The same is true for testing hypotheses relating to the products in the equation as well. For example, if only 2 moles of carbon dioxide were collected after the combustion of ethane, a reasonable hypothesis might be that only 1 mole of ethane had undergone combustion.

Stoichiometry is an extraordinarily successful mathematical model, since it is simple and can be used by anyone who understands the conventions associated with it. However, it is important to understand that the mathematical model appears deceptively simple. The model is not mere speculation, but a model based on a great deal of theory. Multiple theories comprise the conceptual framework of stoichiometry, such as the mole concept, the conservation of mass and the conservation of energy, to name a few.

Epistemological Considerations

DiPrete and Eirich (2006) described how the analysis of empirical data could suggest the presence of CA (p. 274-275). The methods they described, distributional analysis, mathematical modeling and growing inequality can differ epistemologically, leading to unnecessary confusion. For example, data analysis can involve one or more of the following approaches: (1) statistical null hypothesis significance testing, (2) exploratory data analysis (e.g. Tukey, 1962, 1968) and (3) model development (e.g. de Mast et al. 2023; Mazur, 2006). The author took no particular stance on the value of the various methods, and indeed, used all three during data analysis.

Null Hypothesis Significance Testing

When considering statistical analysis, the majority of researchers tend to think of null hypothesis significance testing (NHST). NHST includes parametric statistical analysis methods such as t-tests, analysis of variance and linear regression. For example, a *t*-test can be used to determine whether the null hypothesis, that there is no difference between the means of two groups, can be rejected. The problem according to Cohen (1994) was that analysis can devolve into “mechanical dichotomous decisions around a sacred .05 criterion” (Cohen, 1994, p. 997). At issue is the frequent lack of an alternative hypothesis, since rejected the null does not prove the opposite of the question initially posed. Tukey (1969) wrote that if the "real question cannot be answered by a correlation coefficient it can be fatal to insist on using [one]...whether or not some other question appears to be answered” (p. 84)

The standardized and normed intelligence and aptitude tests popularized by Stanford psychologist Lewis Terman (1877-1956) were based on NHST. Terman believed that intelligence (1) was hereditary, (2) normally distributed in the population, and (3) a linear function (Hawks, 2015). This view assumed that parametric statistics, based on a presumed normal distribution, could determine student intelligence and predict future outcomes. Consequently, the Stanford-Binet Intelligence Scale, or test, was normed against scores hand-picked for conformity by Terman and his collaborator, Merrill (p. 2).

This same principle is used when grading tests scores on a curve, which also presumes a normal distribution, requiring that performance measures be forced into a normal distribution . According to Fendler and Irfan (2008):

> The bell curve takes diversity and reduces it to a simple and comprehensible average; it takes a statistical probability and converts it to an expectation that can inform policy; it takes random historical occurrences and imposes patterns of relations on them. (p. 76).

A recent review by Bowen and Cooper (2022) reviewed the use of grading on a curve in chemistry.

Mills et al. (2009) described creating a diagnostic measure of probable student success on a standardized curve. This was accomplished by first transforming numerical data into Z-scores, Equation 2.2.

$$Z = \frac{X-\mu}{\sigma} \qquad \text{Equation 2.2}$$

Mills et al. (2009) converted 667 GC first exam scores and total course performance scores less first exam scores, into Z-scores. They reported a Pearson correlation coefficient of $r = 0.81$, $p = 0.05$, presumed, between the two variables. Using the correlation curve they were able calculate how far from the mean a particular score was. The probability of a student passing GC when their first exam score was at one standard deviation below the norm was predicted to be 22% (p. 741).

The success rate for correctly predicting GC course failure was reported as to be 75.2% (p. 740). The predictive ability was likely influenced by the content of first exam score, which was heavily weighted towards stoichiometry (p. 739). However, the ability to generalize the tests to other institutions was limited, since natural differences between instructors, student background and other factors likely could not be captured with a single set of variables (p. 742).

Exploratory Data Analysis

The *Engineering Statistics Handbook* describes the use of exploratory data analysis (EDA) as "an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical)" (NIST/SEMATECH, 2023, 1.1.1). The purpose of EDA is to allow the data to reveal what form the underlying model might take. Examples of graphical methods of analysis include histograms, probability plots, residual plots, quantile-quantile plots and scatter plots, to name a few.

Tukey is generally credited with this approach, likening the researcher to a detective using techniques with the necessary flexibility (e.g. 1962 and 1966). EDA has been described by some as "a widely celebrated paradigm shift in analysis" (Rodgers, 2010, p. 4). According to Tufte (2001) "Tukey...opened up the field, as his brilliant technical contributions made it clear that the study of statistical graphics was intellectually respectable and not just about pie charts" (p. i).

Cleveland and McGill (1985) called graphs “powerful tools both for analyzing scientific data and for communicating quantitative information” (p. 828).

Arguably, graphical techniques may be better appreciated in manufacturing presumed to operate under the proverb time is money. For example, learning curves were first described by Wright (1936) who sought to improve the costs associated with building airplanes. Hirschmann (1964) described “profiting from the learning curve” in operations and supply management. In production, learning curve models based on “cognitive and motor element” ratios have also been described (Peltokorpi & Jaber, 2022). However, learning curves are still used in education despite their wide applicability in other fields (e.g. Kuhfeld & Soland, 2021).

In 2015, Trafimow and Marks (2015), new editors of *Basic and Applied Social Psychology* banned “p-values, t-values, F-values” and statements describing “significance”(p. 1). They stated that “A 95% confidence interval means... if an infinite number of samples were taken and confidence intervals computed, 95% of the confidence intervals would capture the population parameter” (p. 1). They suggested authors focus on descriptive statistics, effect sizes, data distribution and testing large enough sample sizes (Trafimow and Marks, 2015).

Mathematical Modeling

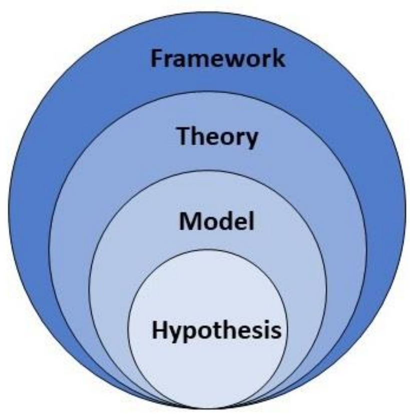
Mathematical modeling is used extensively in the physical sciences and engineering, although they are also employed in economics, psychology, sociology and education (e.g. Bothner et al. 2010; Cheng, 2014; DiPrete and Eirich, 2006; Walberg and Tsai, 1983). For example, Hasl et al. (2022) translated “wage dynamics into empirical models” (p. 5). Hasl was part of a multidisciplinary research team who considered theoretical aspects related to cumulative advantage and educational attainment.

Where mathematical models fall within an inquiry framework is shown in Figure 2.4, which shows the relationship between the elements within a conceptual framework, theory, models and hypothesis testing. Note that models are intermediary to applied theory and conventional hypothesis testing (Bäckman, & Edling, 1999; Carpiano & Daley, 2006, Kileen, 1999; Saltelli & Puy, 2023). That fact, according to Saltelli (2019) has kept them out of the ongoing "statistics crisis" (Saltelli, 2019, p. 1). Saltelli and Puy (2023) suggested that:

> Models are good at mapping one set of meanings or information onto another set, e.g., moving from assumptions to inferences without losing sight, in the correct use, of all the conditionalities involved in this transposition. (p. 2)

Figure 2.4

Relationship Between Elements of Inquiry



Note: Inspired by relationships described by Carpiano and Daley (2006).

Analysis Using Cumulative Advantage

Demonstrating that students enter GC with varying amounts of prior knowledge and preparation was of paramount interest. A novel analysis strategy was pursued based on the conceptual framework of CA. The methodology by which empirical performance data was analyzed an array of analysis techniques from the categories described. Which approach was used to answer a particular question depended on which epistemological approach matched the question the best.

In their review of CA, DiPrete and Eirich (2006) defined two forms of CA. One was described as “strict Mertonian CA” (p. 272) and described Merton’s original theory of the reward system in scientific research. They also described a “path-dependent” form of CA (p. 276) described by the relationship shown in Equation 2.3. The equation describes the growth of some variable *Y* as being dependent on the “entire history of whatever variables are in *X* and the history of the random shocks, ϵ...as long as $Y > 0$ (p. 276).

$$Y_{it} = Y_{i(t-1)}(1+\Upsilon) + \beta' X_{it} + \epsilon_{it} \qquad \text{Equation 2.3}$$

Both the strict Mertonian and path-dependent forms of CA meet all of the following observed characteristics described by DiPrete and Eirich (2006):

1. The growth rate in an outcome is a function of the current level.
2. Small advantages or disadvantages grow larger over time.
3. The process of CA can be exacerbated by status variables such as race, gender, and can persist over the lifetime.
4. Inequality grows over time at both the individual and population level (p. 280).

Holding all other factors constant, the key relationship to path-dependent CA is the relationship shown in equation 2.4:

$$Y_{it} = Y_{i(t-1)}(1+\Upsilon) \qquad \text{Equation 2.4}$$

Baumert et al. (2012) described path-dependent CA as being particularly suited to describing education. They noted that path-dependent CA was:

> Particularly likely in the context of formalized learning in schools, where the curricular structure of academic subjects tends to be cumulative, and prior knowledge tends to be the best predictor of the subsequent level of attainment. (p. 1348)

DiPrete and Eirich suggested that researchers consider empirical data for specific types of evidence relating to CA. Three particular analysis considerations were described, relating to distributional analysis, mathematical models and growing inequality (p. 274-275). The methods required inductive reasoning in the cases of distributional analysis and increasing variance. Deductive reasoning was also prescribed to probe the relationship between the variables Y_{it} and $(1+\Upsilon)Y_{i,\ t-1}$ in the form of mathematical and statistical modeling.

In the case of distributional analysis, DiPrete and Eirich were clear concerning what was expected visually. They noted that "CA-like processes produce right-skewed distributions on the outcome variable of interest" (p. 274). The mathematical model derived by Allison et al. (1982) was based on a discrete form of the Yule-Simon distribution (Simon, 1955). As a consequence, path-dependent CA is associated with the power-law distribution, as demonstrated by Newman (2005).

The discrete form of the power-law function is shown in Equation 2.5. According to Perc (2014) "not finding a power-law distribution or at least a related fat-tailed distribution" (p. 3) placed doubt about the presence of a CA process. The fat-tailed, or heavy-tailed distribution, alluded to by Perc simply describes a distribution with more outlying data points than expected (Bryson, 1974).

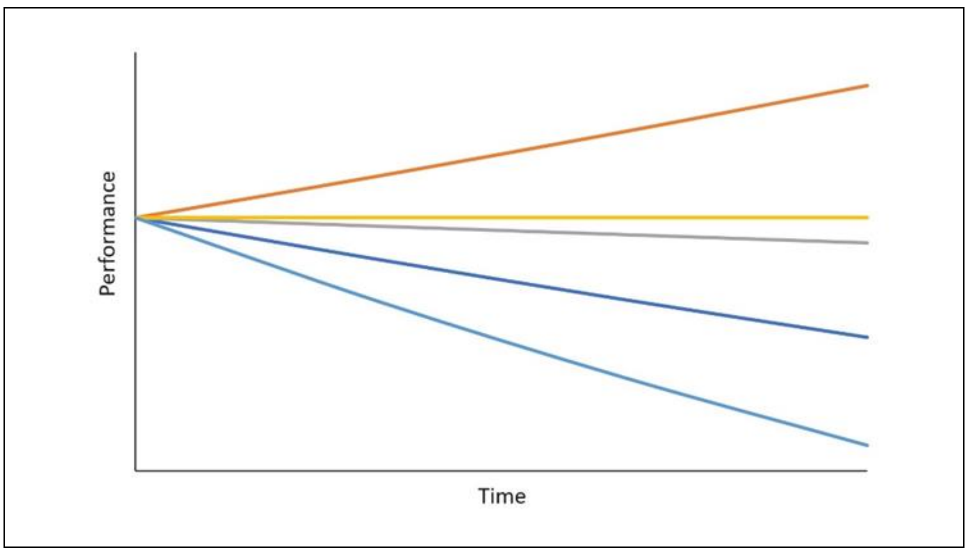
$$p(x) = Cx^{-\alpha} \qquad \text{Equation 2.5}$$

Increasing Variance or Inequality

EDA in practice, involves monitoring real data over time for signs of increasing variance or undesirable data patterns. In this case, increasing variance between student performance measures was expected to grow as predicted by fan-spread hypothesis attributed to Cook and Campbell (1979, p. 184) by Walberg and Tsai (1983). Walberg and Tsai described the “the increasing variation during the course of experience [that] leads to a fan-spread of points when outcomes are plotted against time” (p. 360). Not to be confused with the fan effect related to the ACT-R cognitive model (Anderson & Labiere, 1998), the fan-spread effect can be conceptualized by considering Figure 2.5.

Figure 2.5

Example of the Fan-Spread Effect or Increasing Variance in Learning Outcomes



Increasing variance can be the consequence of more than one or more factors alone or in combination. Possible factors include differences in learning rates, differences related to prior knowledge, or monotonic increases in performance (Ericsson et al. 1993). However, other sources of between-subject variance are also possible, such as inaccurate statistical assumptions regarding the distribution of error. However, in general, CA processes tend to lead to changes in rank performance over time (Kerckhoff & Glennie, 1998).

Mathematical Models

According to Equations 2.3 and 2.4, the variables Y_{it} and $(1+\Upsilon)Y_{i,\,t-1}$ are autocorrelated, that is, part of a consecutive series related by time (Shadish et al. 2002, p. 172). Consequently, the two variables will be correlated, a relationship which must be demonstrated, and the strength ascertained. The mathematical relationship between the two is important to understand. As suggested by Mazur (2006) "theories that appear to make similar predictions when stated in words can be more readily compared and evaluated when they are put into mathematical form" (p. 276).

The importance of mathematical modeling to better understand the relationship between two variables cannot be overstated. For example, a model demonstrating a relationship between Y_{it} and $(1+\Upsilon)Y_{i,\,t-1}$ could provide valuable insight. According to Novak (2022) proposing, "constructing and testing new theories" which are data driven are just as valuable as theory-oriented ones:

> Producing empirical evidence of a mathematical model that cannot support a theory in question does not imply that the theory is wrong. It only suggests that the theoretical statement is not confirmed (not to be confused with 'disconfirmed') or that the model builder made false background assumptions in the model design, e.g., assuming a linear relationship between variables instead of a quadratic one, or using tools/instruments that produce measurement errors. (p. 153).

Which was taken to mean that proposing and testing variable relationships should be a process in which one should readily participate.

Chemistry Education Research Dissertations

For completeness, chemistry education research dissertations and theses were searched for theoretical frameworks incorporating cognitive science and educational psychology. Between 2007 – 2021 only four chemistry education research dissertations referred to either. One of those four only obliquely referred to cognitive load theory, but references made to specific educational psychology concepts are clear.

Baluyut (2015) referenced Johnstone (1993) when referring to the “three levels of [chemical] representations, namely the macroscopic, the [sub]microscopic and the symbolic” (p. 22). This concept was credited to Johnstone (1982 and 1991) and is now known as Johnstone’s Triangle (NRC, 2012, p. 47). Johnstone believed the triangle was responsible for much of the overwhelm students experienced when learning chemistry. Johnstone illustrated the concept of working memory overload using a modest information processing model (e.g., Johnstone, 1997 and 2010) based on previous cognitive theory.

Three references: (1) Buis (2016), (2) Finney (2008), and (3) Mata (2019) all incorporated cognitive load theory into their theoretical frameworks. All cited references by Johnstone (e.g., Johnstone & El-Banna, 1986; Johnstone 1997) were related cognitive load while Finney focused on student competency. Buis and Finney also cited the seminal work on cognitive load theory and human cognitive architecture by Sweller et al. (1998), which described cognitive load in detail, along with human cognitive architecture.

Both Buis (p. 3) and Mata (p. 14) also incorporated Piagetian cognitive development theory into their theoretical frameworks. Pairing Piaget with work by Johnstone (e.g., 2006, p. 51) and Sweller was unusual, since neither had ever subscribed to Piaget’s developmental theories (e.g., Johnstone, 2006). In fact, Tricot and Sweller (2014) described Piaget’s stage theory (Piaget, 1972) as “bedeviling the field of cognitive development” (p. 272).

Only Finney (2008) specifically incorporated cognitive learning theory into the framework of her dissertation. Finney considered novice expert differences in the categorization of chemistry problems, citing seminal work by Larkin et al. (1980) and Chi et al. (1981). Finney also acknowledged that "Perhaps our lack of success in this area is directly related to our lack of understanding of novice problem solving in general chemistry" (p. 3).

Chapter 3: Methodology

Introduction

Multiple studies have suggested that high school preparation is predictive of later GC performance (Sadler and Tai, 2007; Tai et al. 2005; Tai et al. 2006). Models describing learning, such as the Carroll (1989) model considered time spent in the classroom as variable (Carroll, 1989; Grodsky. et al., 2008). Likewise, Ericsson et al. (1993) suggested the amount of accumulated practice guided by feedback was predictive of performance.

Differential student performance at the postsecondary level has been considered within the context of cumulative advantage, attributed to Merton (1968, 1988). For example, Bowen et al. (2005) described CA in higher education as:

> "The accumulation of (often small) advantages and disadvantages over the course of the first 18 years of life that leads to massive preparation differences by the time of college application" (p. 225).

Such appears to be the case for chemistry, a course in which students are known to enter with significant preparation gaps (e.g. Harris et al. 2020). DiPrete and Eirich (2006) in a review of CA, described the accumulation of an advantage such as knowledge or cognitive development as relying on past levels (p. 280). They indicated that evidence for CA could be determined in multiple ways (p. 274-275).

Research Paradigm and Design

A lengthy discussion of the epistemological approaches which could be used to analyze data was described in Chapter 2. All the data analysis approaches described fall under the post-positivist paradigm. Common practices in science and engineering presume hypotheses can only be falsified, and that rigor, validity, reliability, and objectivity are criteria which must be met (Guba & Lincoln, p. 112).

It is important to understand that the postpositivist paradigm is firmly grounded in scientific inquiry. However, there is flexibility in terms of just the methodology used, which is suitable for the question being asked (Wildemuth, 1993). Likewise, if multiple theories and approaches can be used to address a question, they can be utilized simultaneously (Carpiano & Daley, 2006). However, only so many approaches can be taken with numerical data, with mathematical models and statistical analysis predominating.

Research Design

Since a substantial amount of background discussion concerning methodology was undertaken in the previous chapter, it is not restated here. The data analyzed consisted of deidentified numerical performance measures generated by *N* = 409 students. The students were enrolled in GC at a medium-size suburban Tier 1 university located in the Midwest from fall 2016 to spring 2019. Since laboratory experiments at this level of chemistry are confirmatory in nature, only the lecture portion of the course was considered in terms of student performance.

The general demographic profile of the student body is shown in Table 3.1. The student body is more diverse than the state average, with almost half of first-time students receiving Pell grants. In general, the majority of the students were White, female, and part-time students. Approximately 94% of the matriculated students originated in one of five surrounding state counties (IPEDS EF Fall Enrollment Survey, 2020). However, finer detail concerning specific student demographics was not available.

The conceptual framework used for data analysis was CA as mathematically described by DiPrete and Eirich (2006). Path-dependent CA, mathematically modeled by Equation 3.1 was used to predict “the rate of future events as a positive function of previous events” (p. 275). The specific variables used are defined in Table 3.2.

$$Y_{it} = Y_{i(t-1)}(1+Y) \qquad \text{Equation 3.1}$$

Table 3.2

Definition of Study Variables

Variable	Symbol	Type	Definition
Exam 1 Scores	$Y_{i(t-1)}(1+Y)$	Independent	Proxy measure of prior high school experience.
Absolute Knowledge Gain (AKG)	Y_{it}	Dependent	Sum of the last three course exams; proxy measure of knowledge gain over the course of the semester.

The research questions posed arose naturally from the conceptual framework, domain specific content knowledge, cognitive learning theory and CA. The questions were a mixture of inductive and deductive analyses. Inductive analysis was used to understand the distribution of Exam 1 scores and whether increasing inequality or variance between students was evident over time. Deductive analysis was used to determine the relationship between the two defined variables. Once the relationship had been identified, mathematical models were also used to predict future events.

Research Questions:

R1: What did the shape and symmetry of the distribution of Exam 1 scores suggest when plotted as a histogram?

R2: How strong was the correlation between the proxy measure of high school preparation and the proxy measure of the dependent variable, AKG?

R3: Based on the presumed relationship between Exam 1 and AKG, to what extent could a predictive relationship be modeled?

R4: To what degree could the continuity or change in student rank from Exam 1 to AKG be demonstrated graphically?

Data Characterization

The research questions could not be addressed until the numerical data had been curated. Data curation was defined by the following procedures, which were necessary to determine whether the six individual cohorts could be combined into a single population:

1. Establishing whether the data was normally distributed.
2. Determining whether the cohorts could be combined into a single population.
3. Verifying whether the new population was normally distributed.

The answers to questions listed were necessary to establish what types of analysis parametric, or nonparametric could be used for further analyses. Once established, the hypotheses could be formalized.

Hypotheses

H_0 1: The distribution of first exam scores followed a normal distribution.

Hα 1: The distribution of first exam scores did not follow a normal distribution.

H_0 2: The population correlation coefficient, $\rho = 0$ the ranks did not covary.

Hα 2: The population correlation coefficient, $\rho \neq 0$ the ranks were assumed to covary.

H_0 3: Mathematical modeling of the relationship between Exam 1 scores and AKG could not be predicted based on the correlation parameter, ρ.

Hα 3: Mathematical modeling of the relationship between Exam 1 scores and AKG could be predicted based on the correlation parameter, ρ.

H_0 4: Graphical diagrams of student ranking showed no evidence of increasing variance.

Hα 4: Graphical diagrams of student ranking showed evidence of increasing variance.

Data Collection and Management

The data analyzed was obtained in the form of archived data obtained by the professor of record in the form of deidentified spreadsheets. The data was stored on a local computer and password protected. Copies of all data files and statistical test data are available upon demand. The data will be retained for five years past the point of the dissertation publication, at which time the data will be destroyed.

Permissions and Ethical Considerations

The research project was deemed non-human subjects research according to IRB #2065542 SL (Appendix 1). No identifying data remained in the spreadsheet and no contact could be made with any of the students. There were no known ethical conflicts associated with the study.

Instrumentation

Statistical analysis was performed using SAS® OnDemand for Academics (SAS.com) specifically the SAS® Studio application. Additional analyses were performed using Excel for Microsoft 365, 2010 edition, equipped with the Solver Add-in. An add-on for Excel, Power-User, was also used under a free academic license. Additional graphics were enhanced if necessary, using Microsoft Paint 11.2208.60 © 2022. The tables and figures were generated by the author, unless otherwise indicated.

Validity and Reliability

Validity describes how closely assumptions, statements of fact, measurements and conclusions relate to their actual meaning. Four types of validity were considered "construct validity, external validity, internal validity and statistical conclusion validity" as described by Shadish et al. (2002, p. 33). Archived data, analyzed in this study, can present multiple threats to validity in the case of causal relationships (Shadish et al., 2002). No claims of causal relationships were made, nullifying those concerns.

Cohort size and performance can vary depending on depending on whether the semester was in the fall or spring, consistent with observations made in other studies (e.g., Tashiro & Talanquer, 2021). In general, larger class sizes for the fall semester suggested those students were following traditional course sequences. However, those differences did not appear to be a threat to statistical validity. Hypothesis testing was performed using an *a priori* Type I error rate of $\alpha = 0.05$, corresponding to a 95% confidence limit.

Threats to internal study validity refer to the inferences made about the chosen variables. In this study, two proxy variables were assigned; Exam 1 scores for high school preparation and the sum of the last three exams as a proxy for learning gains. The use of these measures was not considered any more problematic than the use of SAT and ACT scores or high school GPA.

Finally, the external validity, or ability of this study to apply more generally to other situations was known to be limited. The population of interest in this study differs substantially from other research universities or contexts such as community colleges. Other researchers are encouraged to consider whether the proposals concerning the specific population described could be applied to their own student populations.

Statistical Analysis Plan

A detailed account of the analyses conducted is shown in Table 3.3 according to the order in which they were performed. While none of the statistical tests were considered out of the ordinary, references are provided for those wishing for more detail. The reference NIST/SEMATECH refers to the online version of the *Engineering Statistics Handbook*.

Table 3.3

Statistical Analysis Plan and Execution Order

Test	Test Name	Description	Reference
1	Descriptive Statistics	Mean, S.D., Median, 95% CI.	Price et al. (2023)
2	Skew and Kurtosis	Skew, Kurtosis.	NIST/SEMATECH 1.3.5.11
3	Measures of Location	Histogram	NIST/SEMATECH 1.3.5.1
4	Distribution Analysis	Q-Q plot, Probability plot.	NIST/SEMATECH 1.3.3
5	Analysis of Variation	Kruskal-Wallis ANOVA	NIST/SEMATECH 7.4.1
6	Goodness-of-fit Test	Kolmogorov-Smirnov Test	NIST/SEMATECH 1.3.5.16
7	Mathematical Model	Trendline	Excel
8	Correlation	Spearman rho	Siegel and Castellan (1988)
9	Nearest Neighbor Match	Spearman rho	Stuart (2010)
10	Mathematical Model	Trendline	Excel
11	Student Rank Change	Sankey Diagram	Power-User

Chapter 4: Results

Introduction

Empirical performance data generated by *N* = 409 GC students enrolled in the course from fall 2016 to spring 2019 was analyzed within the conceptual framework of CA. According to DiPrete and Eirich (2006) evidence suggestive of a CA process could be gathered from empirical data. Three types of evidence were considered relevant: (1) a skewed distribution of the variable affected by CA, (2) small scale modeling of the relationship between Y_{it} and $Y_{i(t-1)}(1+Y)$ and (3) evidence suggestive of growing inequality as a function of time.

Description of the Sample

Two slightly different GC student populations were analyzed for this study. The first sample was comprised of all students who had taken the first exam, *N* = 409. This number included students who had neither completed nor officially withdrawn from the course. However, since Exam 1 scores were the most suitable measure of high school preparation, the maximum amount of variability in this population was of particular interest.

The second student population analyzed was comprised of students who had taken all four course exams, *N* = 343 students. This data was described as the “matched” Exam 1 and AKG scores; AKG being the sum of the last three course exams. The number of data sets, *N* = 343 was 66 less students than the population initially described. Of this number, *N* = 6 student scores were excluded due to missing or incomplete data. The remaining *N* = 60 students had failed to complete the course for unknown reasons.

Research Questions

R1: What did the shape and symmetry of the distribution of Exam 1 scores suggest when plotted as a histogram?

R2: How strong was the correlation between the proxy measure of high school preparation and the proxy measure of the dependent variable, AKG?

R3: Based on the presumed relationship between Exam 1 and AKG, to what extent could a predictive relationship be modeled?

R4: To what degree could the continuity or change in student rank from Exam 1 to AKG be demonstrated graphically?

Hypotheses

H_0 1: The distribution of first exam scores followed a normal distribution.

Hα 1: The distribution of first exam scores did not follow a normal distribution.

H_0 2: The population correlation coefficient, $\rho = 0$ the ranks did not covary.

Hα 2: The population correlation coefficient, $\rho \neq 0$ the ranks were assumed to covary.

H_0 3: Mathematical modeling of the relationship between Exam 1 scores and AKG could not be predicted based on the correlation parameter, ρ.

Hα 3: Mathematical modeling of the relationship between Exam 1 scores and AKG could be predicted based on the correlation parameter, ρ.

H_0 4: Graphical diagrams of student ranking showed no evidence of increasing variance.

Hα 4: Graphical diagrams of student ranking showed evidence of increasing variance.

Preliminary Data Analysis

As described in Chapter 3, the intent was to analyze a single population of students rather than by cohort. To combine the cohorts, it was necessary to establish that between group variance did not exceed within group variance. Whether variance could be determined using a parametric or nonparametric hypothesis test depended on whether the data was normally distributed. The data was curated using descriptive statistics, measures of skew and kurtosis, distributional analysis and goodness-of fit tests.

An example analysis scheme is described in full for fall 2016 Exam 1 data, *N* = 85 students. An identical procedure was carried out for all cohort data corresponding to Exam 1 and matched Exam 1 and AKG data. However, that data was placed in Appendix 2 for the sake of brevity. An example of descriptive statistics consisting of the sample mean, standard deviation, median and the 95% confidence interval is provided in Table 4.1.

Table 4.1

Descriptive Statistics All First Exam Scores, N = 409

Group	N	Mean	Std Dev	Median	L 95% CL	U 95% CL
F 2016	85	197.60	41.94	206.10	188.56	206.64
S 2017	52	193.55	39.59	206.90	182.52	204.57
F 2017	90	191.73	45.62	203.90	182.17	201.28
S 2018	68	193.22	41.73	201.49	183.12	203.33
F 2018	57	199.86	41.19	209.45	188.93	210.79
S 2019	57	180.75	48.37	179.52	167.92	193.59

Example Skew and Kurtosis

Measures of skew and kurtosis are indicators of symmetry and data homogeneity (NIST/SEMATECH, 1.3.5.11). The skew indicates whether the distribution is shifted to the left or right of center. Kurtosis reflects the number of outlying data points present in the tails of a distribution. Examples of skew and kurtosis are found in Table 4.2.

Table 4.2

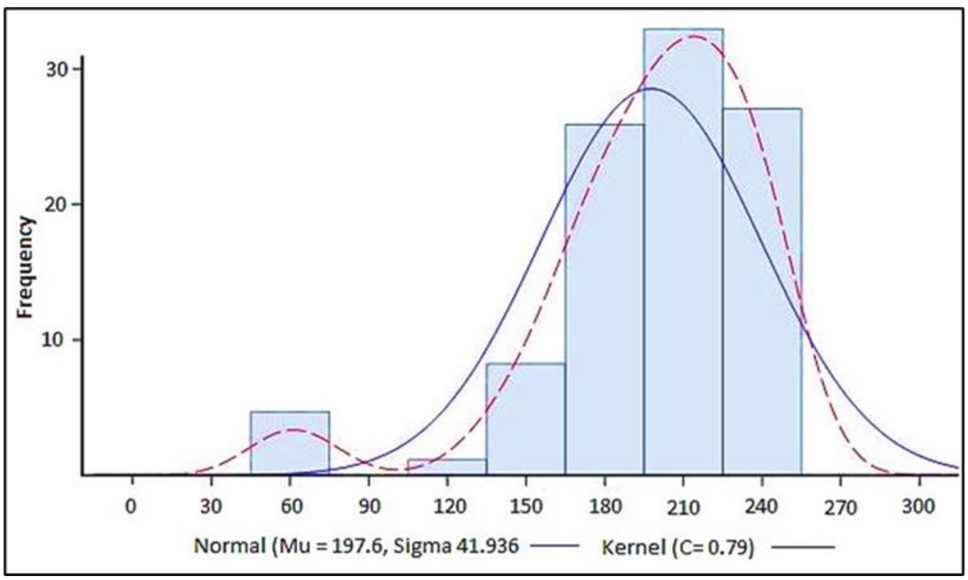
Skew and Kurtosis All First Exam Scores, N = 409

Group	N	Mean	Median	Skew	Kurtosis
F 2016	85	197.60	206.10	−1.57	3.25
S 2017	52	193.55	206.90	−1.10	0.44
F 2017	90	191.73	203.90	−1.31	1.18
S 2018	68	193.22	201.48	−0.80	0.05
F 2018	57	199.86	209.45	−1.62	3.55
S 2019	57	180.75	179.52	−0.55	−0.08

The consequences of skew and kurtosis can be understood visually by examining the histogram corresponding to the fall 2016 Exam 1 cohort, *N* = 85, Figure 4.1. The skew and kurtosis values of −1.37 and 3.25, respectively, are indicated in the histogram. The distribution, blue line, can be interpreted as shifted and squat in shape. The kernel density plot, the red dashed line, indicated a slightly bimodal distribution for the Exam 1 scores.

Figure 4.1

Example Histogram Exam 1 Scores Fall 2016, N = 85



Example Measures of Location

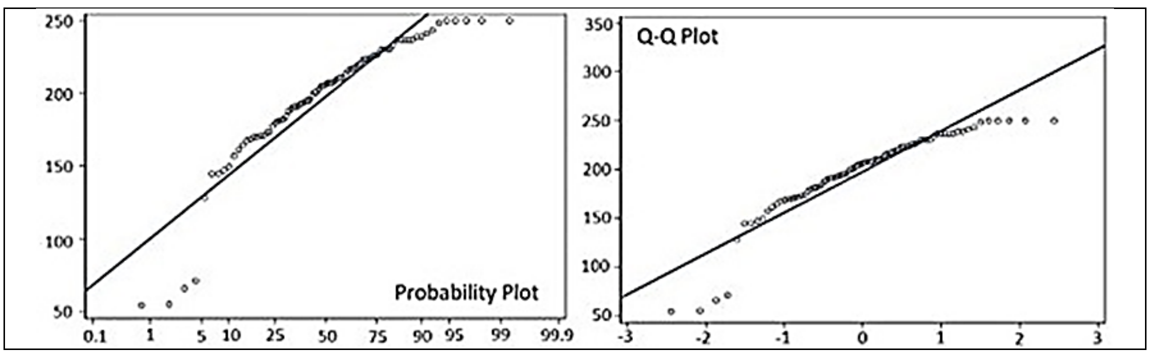
As already suggested by Figure 4.1, the histogram for fall 2016 Exam 1 scores were not normally distributed. The values of mu and sigma correspond to the mean and standard deviation of the population, respectively. If the data had been normally distributed, then the mean, 197.60 would have been equivalent to the calculated median which was reported as 206.10 in Table 4.1.

Example Distribution Analysis

Evidence for the non-normal distribution of Exam 1 scores is also evident in the probability and the quantile-quantile plot, Figure 4.2. In both plots, the distribution of the data points should be distributed and laid evenly across the 45 degree line. As demonstrated by both plots, there are a significant number of points which do not fall linearly along the plot as expected. The tail ends of the data in both plots show significant deviation from the expected distribution.

Figure 4.2

Probability and Q-Q Plot for *Exam 1 Scores Fall 2016, N = 85*



Example Kolmogorov-Smirnov Goodness-of-Fit Test

While all previous data has shown that the data was not normally distributed, a goodness-of-fit test should also be used to verify whether that assumption was true. When data is analyzed in a statistics program such as SAS® goodness-of-fit tests for normality are automatically generated. The test of interest for this data set was the Kolmogorov-Smirnov (KS) goodness-of-fit test, which is a hypothesis test:

H_0: The observed empirical distribution is consistent with the theorized normal distribution.

H_a: The observed empirical distribution is not consistent with the theorized normal distribution.

The KS goodness-of-fit test calculates the D-statistic, at the 95% confidence interval, Table 4.3:

Table 4.3

Kolmogorov-Smirnov Goodness-of-Fit Test Score

α	***D*-Statistic**	***Pr > D***
0.05	0.10655	0.018

Since 0.018 < 0.05 the null hypothesis, that the observed empirical distribution is consistent with a theorized normal distribution is rejected; the data is not considered normally distributed.

Example Kruskal-Wallis Analysis of Variance

The Kruskal-Wallis analysis of variance test is a nonparametric test which compares *k* samples under the following null hypothesis:

H_0: The data follow the same empirical distribution function.

Ha: The data do not follow the same empirical distribution function.

The test yields the *H*-statistic and a p-value which can be used to determine whether the null hypothesis can be rejected. Table 4.4 shows the Wilcoxan Scores for all Exam 1 scores, $N = 409$. The results of the Kruskal-Wallis Test are shown in Table 4.5, were $H(5) = 6.6472$, $p = 0.2482$. Since $p = 0.2482 > 0.05$ the null hypothesis, that the data follow the same empirical distribution function could not be rejected. Consequently, the six cohorts were combined into a single population.

Recall that all data was analyzed similarly, with results reported in Appendix 2. Table 4.5 also includes the Kruskal-Wallis one-way analysis of variance for the matched pair Exam 1 and

AKG, $H(5) = 4.2662$, $p = 0.5118$ and $H(5) = 6.0744$, $p = 0.2990$, respectively. Since 0.5118 and 0.2990 > 0.05, the null hypothesis that the data follow the same empirical distribution function could not be rejected. Consequently, the six cohorts were combined into a single population.

Table 4.4

Kruskal-Wallis One Way Analysis of Variance First Exam Scores By Group

Group	*N*	Sum of Scores	Expected H_0	Std Dev H_0	Mean Score
F2016	85	18380.5	17425	970	216.24
S2017	52	10561.5	10660	796	203.11
F2017	90	18327.0	18450	990	203.63
S2018	68	13891.0	13940	890	204.28
F2018	57	12830.0	11685	828	225.09
S2019	57	9855.0	11685	828	172.89

Table 4.5

Kruskal-Wallis One-Way Analysis of Variance

Group	N	Chi-Square	DF	Pr > D
Exam 1 Scores	409	6.6472	5	0.2482
Matched Exam 1	343	4.2662	5	0.5118
AKG	343	6.0744	5	0.2990

Comments on Nonparametric Testing

It has been noted often that real data is naturally skewed (e.g., Blanca et al., 2013; Yuan et al., 2017). For example, Walberg et al. (1984) described data related to “knowledge production and consumption” (p. 87) as being naturally skewed. Ericsson et al. (1993) demonstrated that that the relationship between performance and practice was a monotonic, nonlinear function (p. 367).

While parametric testing is often considered superior to nonparametric testing, the suitability of a particular test is actually dictated by the nature of the data. Arguments made about the robustness of parametric testing must take into account the potential error. For example, Leech (2002) noted that the Wilcoxon rank-sum test could be three to four times more powerful than a *t*-test under conditions of non-normality (p. 10 and references therein).

A meta-analysis by Cain et al. (2017) found that of the 1,567 univariate distributions examined, 74% did not meet the criteria for a normal distribution. Based on typical skew and kurtosis values, a simulation study suggested error rates were higher than expected. What would normally be a Type I error rate of 5% was found to be closer to 17% (p. 1716).

Many chemistry education research studies are presumed to lack goodness-of-fit data, since the results are not provided. In studies which referred to normal distributions, the work was perfunctory. For example, one research report simply stated that "All distributions were reasonably normal" based on a visual examination (Cracolice & Busby, 2015, p. 1793).

Another report described "significant negative skew, indicating the scores were more heavily distributed at the higher value" (Lewis & Lewis, 2007, p. 39). The skew suggested that some students were better prepared than others but that issue was not considered. It was presumed that since the study was focused on "identifying at risk students" (p. 39) the issue of normality was not pursued.

The performance data analyzed in this study was not normally distributed, and indeed, was not expected to be. Since statistical analysis is most accurate when the test dictated by the data is performed, assumptions of normality were not made. All subsequent statistical analysis tests performed used nonparametric techniques when hypothesis testing was required.

Results by Research Question

Research Question 1: What did the shape and symmetry of the distribution of Exam 1 scores suggest when plotted as a histogram?

It was reasoned that the first few weeks of a GC course are more likely to be spent reviewing high school chemistry. If true, the first exam score would be expected to be the highest of the four exams taken. For students who took all four course exams, *N* = 343, this appeared to be the case, at least based on the confidence intervals shown in Table 4.7.

Table 4.7

Comparison of Mean Exam Scores

Variable	*N*	Mean	Std Dev	L 95% CL	U 95% CL
Matched Exam 1	343	201.05	35.82	197.24	204.85
Matched Exam 2	343	151.79	51.53	146.31	157.26
Matched Exam 3	343	176.22	51.57	170.74	181.69
Matched Exam 4	343	146.17	58.09	140.01	152.34

DiPrete and Eirich (2006) suggested that "CA-like processes produce right skewed distributions on the outcome variable of interest" (p. 274). A histogram of Exam 1 scores, *N* = 409 was plotted, the distribution shown in Figure 4.3. This distribution was skewed left, with more students receiving Exam 1 scores at the higher end of the range of 35.82 – 250.00.

Normally distributed data is characterized by the majority of data points being grouped around the mean of the population. This results in values for the mean, median and mode being similar in value. This was emphatically not the case for the distribution of Exam 1 scores, shown in Figure 4.3. The values for the mean, median and mode were quite different, 193.03, 203.95 and 250.00, consistent with a skewed distribution.

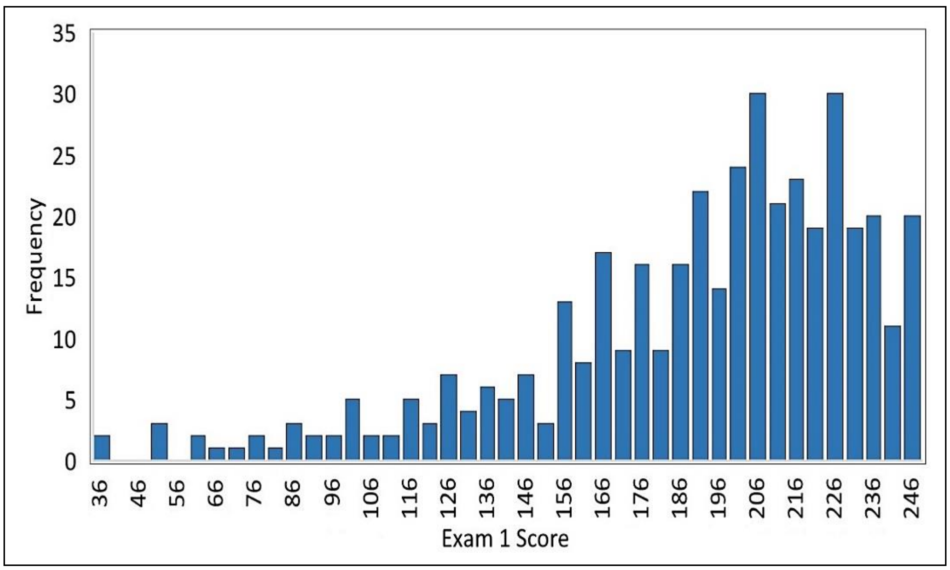
Cumulative distributions, power-law functions and path-dependent CA are all related concepts (DiPrete & Eirich, 2006, p. 274). Discrete power-law distributions law take the form of Equation 4.1:

$$p(x) = Cx^{-\alpha} \qquad \text{Equation 4.1}$$

Unlike normal distributions, distributions related to a power-law have erratic outlying data in the tails (Clauset et al. 2009, p. 1). Newman (2005) cautioned that “Few real-world distributions follow a power law over their entire range... not for smaller values of the variable being measured...[many] have a power-law tail” (p 330).

Figure 4.3

Distribution of Exam 1 Scores N = 409 (Bin Width = 5)



The significance of the power-law regards the interpretation by some that the presence of a CA-like process requires evidence of at least tailing data (e.g. Perc, 2014, p. 3). However, determining the exact form and value for the value of α in a power-law is not trivial (Clauset et al. 2009; Newman, 2005). However, trendlines can be fit to a histogram, and as mathematical models of the data, can be used comparatively. For example, a trendline fit to the upper 50% score distribution may take a different mathematical form than the 50%.

Consider Figure 4.4, which contains trendlines associated with the top and bottom 50% of the cumulative distribution frequency of Exam 1 scores. The top curve is fitted with an exponential trendline, while the bottom is fit with a power-law. The trendline equations were created with the actual data; the trendlines have simply been extrapolated over the full range of Exam 1 scores, Table 4.8. While no claim is made regarding the accuracy of the trendline fit, the fact that they are different is significant. The mathematical functions suggest students in the top 50% learn at an exponential rate, and are expected to gradually outpace the bottom 50%.

Figure 4.4

Cumulative Frequency Plots Top and Bottom 50% Exam 1, N = 409

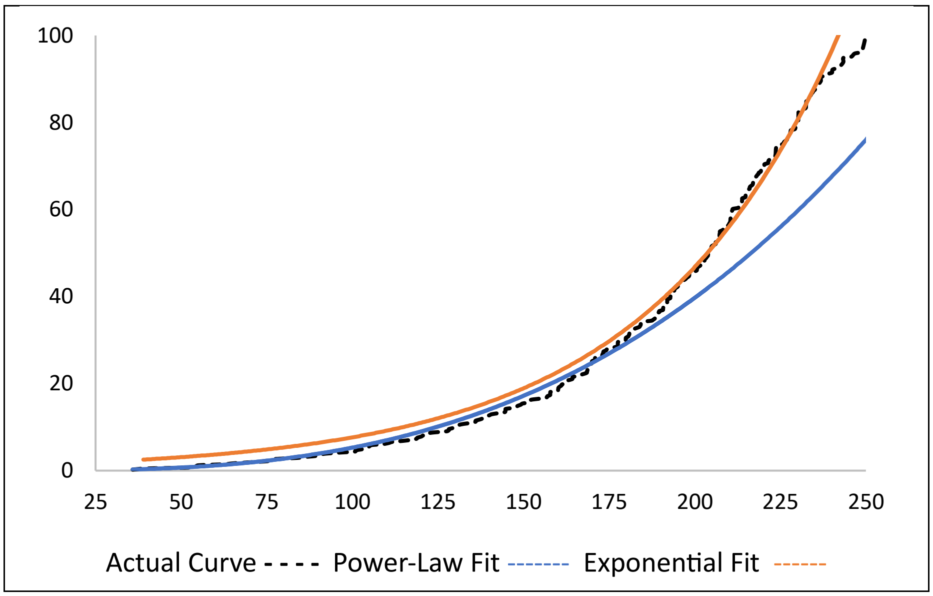


Table 4.8

Coefficient of Determination Values for Various Model Fits

Group	Fit	Equation	R^2
1	Exponential	$y = 1.2532e^{0.0181x}$	0.97
2	Power	$y = 8E\text{-}6x^{2.91}$	0.99

Research Question 2: How strong was the correlation between the proxy measure of high school preparation and the proxy measure of the dependent variable, AKG?

According to the Equation 4.2, Y_{it} and $Y_{i(t-1)}(1+\Upsilon)$ are autocorrelated, the same variable separated by time. Consequently, the two must be correlated, although to what extent is unknown. The correlation between the proxy measures, matched Exam 1 and AKG, *N* = 343, was calculated as rho (ρ), the nonparametric Spearman rank-order correlation coefficient.

$$Y_{it} \text{ and } Y_{i(t-1)}(1+\Upsilon) \qquad \text{Equation 4.2}$$

The null hypothesis for the Spearman correlation, ρ, is stated as:

H_0 The ranks do not covary between the two variables ρ = 0.

H_a: The ranks covary between two variables, ρ ≠ 0.

The calculated value of ρ is shown in Table 4.9; since $p < 0.0001 < 0.05$, the null hypothesis, that the ranks do not covary was rejected. Consequently, a correlation between the two measures was presumed, $\rho = .57002$, $p = .05$, a moderate correlation.

Table 4.9

Spearman Correlation Coefficient between Exam 1 and AKG, N = 343

Variables	α	Spearman ρ	p-Value
Exam 1 x AKG	0.05	.57002	< .0001

Research Question 3: Based on the presumed relationship between Exam 1 and AKG, to what extent could a predictive relationship be modeled?

While a correlational relationship between matched Exam 1 and AKG was established, the corresponding scatter plot, shown in Figure 4.5 is difficult to interpret. Table 4.9 shows the attempted fit with various trendline functions, all with mediocre coefficients of determination.

Figure 4.5

Scatter Plot of AKG as a Function of Exam 1 Scores

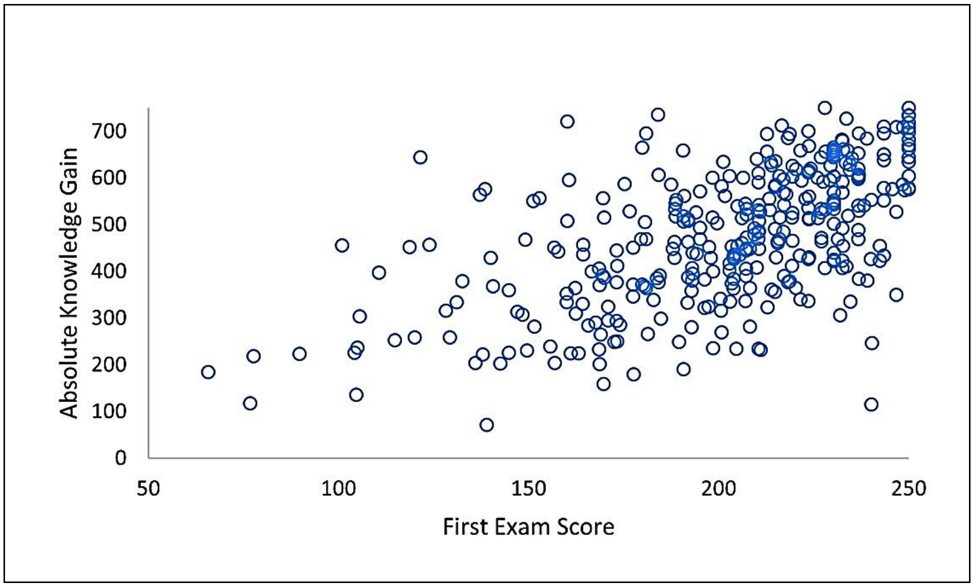


Table 4.9

Coefficient of Determination Values for Various Model Fits

Fit	Equation	R^2
Exponential	$y = 137.69e^{0.0059}$	.341
Power	$y = 2.4258x^{0.9873}$	.332
Linear	$y = 2.3127x + 8.3948$	.332

An idealized model of the relationship between Exam 1 and AKG was constructed using nearest neighbor matching to create ideal monotone pairs. The Spearman rank-order correlation for the ideal curve, ρ, is shown in Table 4.10. The p-value < 0.0001 < 0.05, indicated that the null hypothesis, that the ranks do not covary, was rejected. A nearly perfect monotonic model was created, $\rho = 0.99995$, $p = .05$ and overlaid as a black trendine on the former scatter-plot, Figure 4.5. The dashed red line, indicated that the best mathematical fit for the curve was given by an exponential function: y = 59.419e0.01x and $R^2 = .9961$.

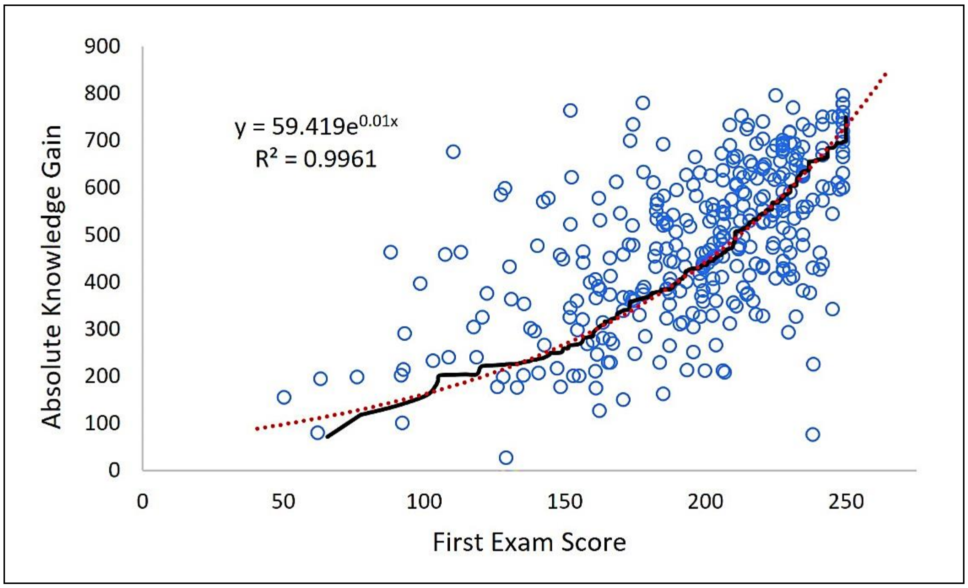
Table 4.10

Idealized Spearman Correlation Coefficient between Exam 1 and AKG, N = 343

Variables	α	Spearman ρ	p-Value
Exam 1 x AKG	0.05	0.99995	< .0001

Figure 4.5

Model of the Learning Curve Based on Idealized Monotonic Relationship



Although only a model, features consistent with known learning curve phenomena were present (Evans et al., 2018; Newell et al., 2001). For example, at the lower end of the curve, growth is lagged, followed by two plateaus before rising expoentially. The curve eventually ends in an uptick, corresponding to a ceiling effect for the highest performers.

The shape of the curve is consistent with what is known as an increasing returns learning curve model. Increasing return models are associated with dual phase learning, where both cognitive and motor skills contribute to increasing performance (e.g. Poltokorpi & Jaber, 2022). These models are also found in economics, business and models of productivity, where greater outputs are realized with smaller inputs. For example, Arthur (1996) in the the *Harvard Business Review*, explained increasing returns thusly:

> Increasing returns are the tendency for that which is ahead to get further ahead, for that which loses advantage to lose further advantage. They are mechanisms of positive feedback that operate within markets, businesses, and industries - to reinforce that which gains success or aggravate that which suffers loss. (p. 1)

A description which is closely related to the concept of cumulative advantage.

Research Question 4

The acquisition of new knowledge and skills as a function of time is formally a rate change, which varies by student. Differences in learning rates have been described in terms of the fan-spread hypothesis attributed to Cook and Campbell (1979, p. 184). The fan-spread is presumed to be the difference in the learning rates between students with different amounts of prior preparation. For example, Kerckhoff and Glennie (1998) found that CA was associated with changes in student ranking over time.

Demonstrating increasing variance as a function of time on an individual level simply requires a plot of performance as a function of time, or even a simple run time plot. However, displaying growing inequality for a large number of observations can be difficult to appreciate. The decision was made to display increasing inequality or variance as a function of time using a Sankey flow diagram (Kennedy & Sankey, 1898).

Sankey diagrams are used to demonstrate the flow from one state to the next, such as the flow of steam, their original use. Here the diagram was used to demonstrate changes to student rankings over time. The data was arranged as quartiles of matched Exam 1 scores flowing to quartiles of AKG values, Table 4.11.

Figure 4.11

Numerical Values Used to Construct the Sanky Diagram

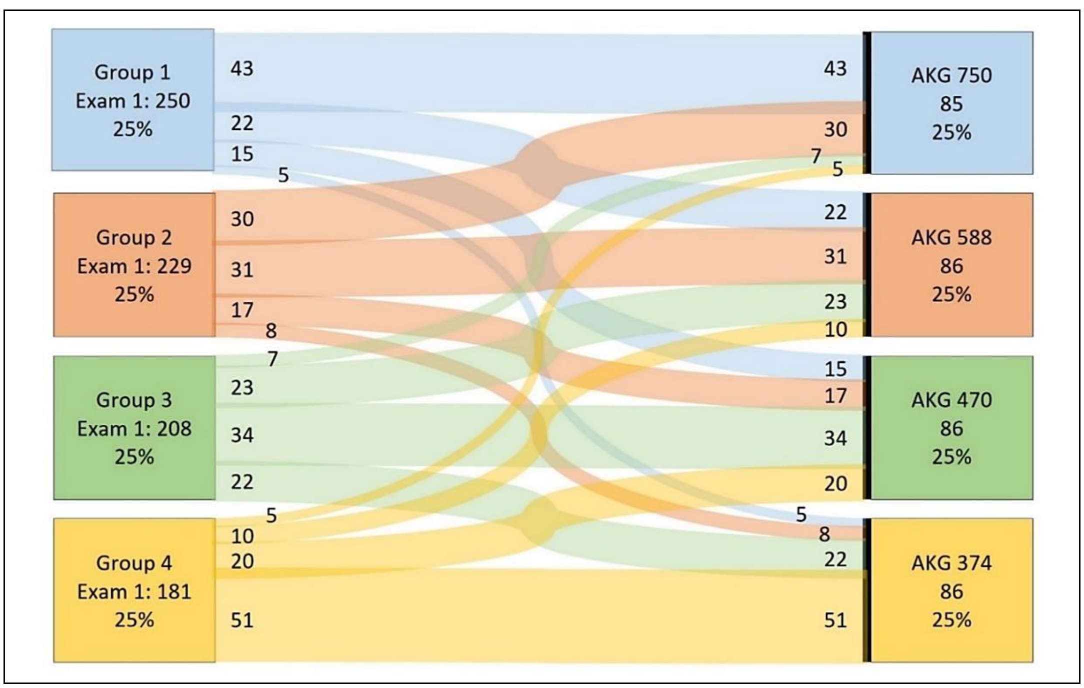
Exam 1 Score	*N*	AKG 750		AKG 588		AKG 470		AKG ≤ 374	
		No.	%	No.	%	No.	%	No.	%
230 - 250	85	43	51	22	26	15	17	5	6
209 - 229	86	30	35	31	36	17	20	8	9
181 - 208	86	7	8	23	27	34	40	22	26
≤ 180	86	5	6	10	12	20	23	51	60

Note. Percentages rounded to whole integer; total percentages may exceed 100.

The resulting Sankey diagram shown in Figure 4.6 can appear overwhelming at first glance. However, the diagram holds a great deal of useful information provided it is examined with patience. One note of particular importance is that what can appear to be negative downward flow is not a loss of points, only a relatively smaller AKG when compared to Exam 1 scores.

Figure 4.6

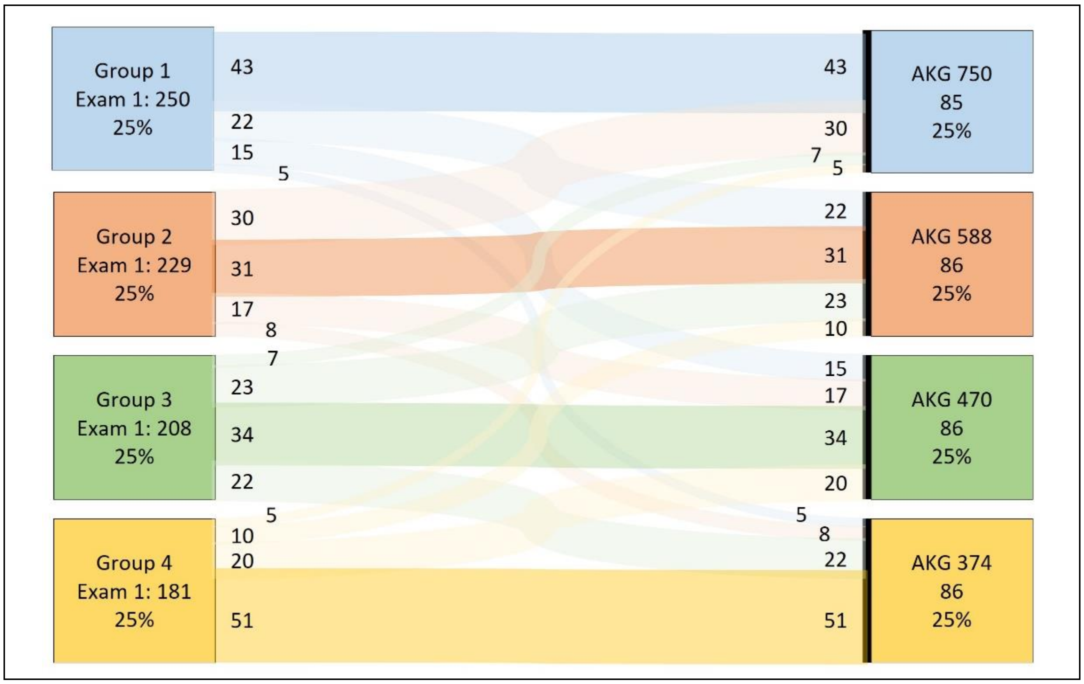
Sankey Diagram of Student Performance on Exam 1 to Absolute Knowledge Gain



The following trends are of particular importance and so have been explained in greater detail. Figure 4.7 demonstrates the general stability of student ranking from Exam 1 to the end of the semester. Of the four groups, 59%, 50%, 40% and 36% of students in Groups 4, 1, 3 and 2, respectively, did not change from one quartile to the next.

Figure 4.7

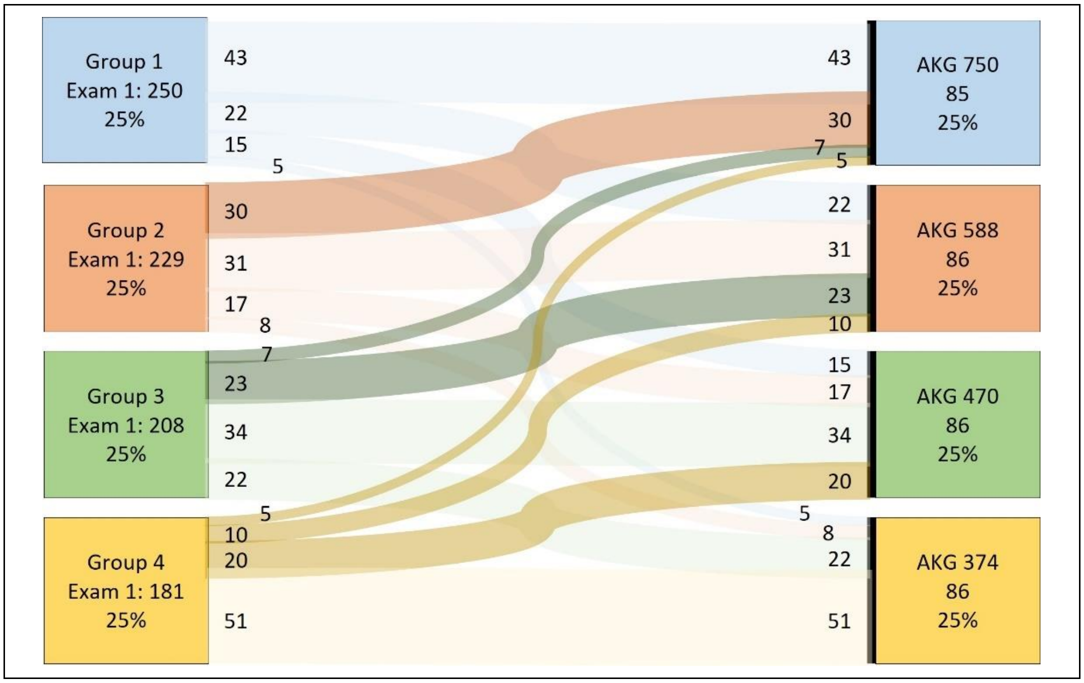
Ranking Continuity of Students from Exam Quartile to AKG Quartile



The upward trends in student quartile ranking have been highlighted in Figure 4.8. Of the three groups where upward mobility could occur, 35%, 35% and 30% were from Groups 2, 4 and 3, respectively. However, only 8% and 6% of students from the Groups 3 and 4 had large enough gains to finish in AKG quartile 1. On the other hand, 27% and nearly 9% of students from Groups 3 and 4, respectively had gains which allowed them to finish in the AKG quartile 2. About 23% of students form Group 4 had gains which allowed them to end in AKG quartile 3. Whether these gains had significant effects on the final course is unknown since grades and scores had been decoupled prior to analysis. Students who benefitted the most were the 30 students or approximately 35% of Group 2 who ended in the highest ranking quartile.

Figure 4.8

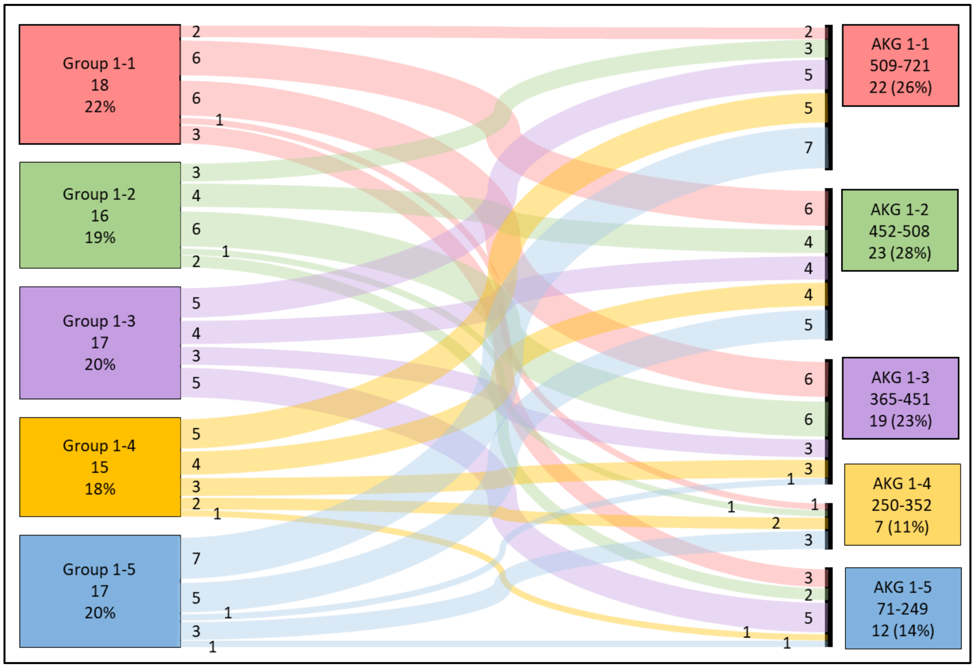
Ranking Change of Students Upwards from Exam Quartile to AKG Quartile



A more detailed examination was made of a single group of students, those from Group 4, who had scored less than 180 on matched Exam 1, Figure 4.9. Even within this group there is a significant degree of variability in terms of AKG accrued over the course of the semester. It is worth noting that the dynamic process demonstrated is consistent with the idea that learning is a complex series of cognitive events that occur over extended periods of time.

Figure 4.9

Ranking Change Within Group1 and Lowest AKG Quartile



Conclusion

Empirical data generated by six GC cohorts, N = 409 was analyzed for evidence that a cumulative advantage process might be responsible for differential student performance. Analyses were performed as suggested by DiPrete and Eirich (2006): (1) distributional analysis, (2) modeling and (3) signs of growing inequality or variance. The results of those analyses, as well as learning curve modeling suggested the presence of differences in initial preparation, which grew over time, consistent with a process of cumulative advantage.

Chapter 5: Discussion

Introduction

Prior to a discussion of the study results, an acknowledgement of the chemistry students who generated the data is offered. Each of the data sets analyzed represented a student, who remained nameless and faceless. Care must be taken to avoid blaming a student for their own misfortunes, particularly when students are anonymous (e.g., Bensimon, 2005). We, as educators, must always remember that each student has a unique story, a once imagined future, which did not involve failing GC.

Chemistry education research appears to be in the midst of welcome reform, a sign that all faculty are not satisfied by the status quo. Researchers have called for new attitudes towards students (e.g., Vyas & Reid, 2023), the use of a cognitive learning theory (e.g., Cooper & Stowe, 2018), and curriculum reform (e.g., Tashiro & Talanquer, 2021). After years of near silence on the subject, issues related to diversity, equity, and inclusion have begun to be openly discussed in American chemistry education literature (Ryu et al., 2021; Stitzel & Raje, 2022).

Summary of Findings

Cumulative advantage is the mechanism used to explain underlying achievement gaps (Baumert et al., 2012) and achievement gaps have been described for incoming GC students (Harris et al. 2020). It was presumed that cumulative advantage could be problematic for some portion of the students who enroll in GC. A number of analyses were performed on empirical GC student performance data for signs a cumulative advantage process could be present.

The distribution of Exam 1 scores, $N = 409$ depicted as a histogram was skewed left with a long tail. The number of data points in the tail was consistent with a fat-tailed distribution (e.g. Hayes, 2007). Despite the number of observations, the Central Limit Theorem limit was never realized. The data never reached normality, suggesting the influence of a CA-like process was responsible for the distribution.

Mathematical modeling of the trendline associated with the skewed histogram differences between the upper and lower 50% of the cumulative frequency distribution. Mathematically, the trendlines were qualitatively different, with one defined by an exponential function, the other a power-law. Qualitatively, this suggested that students with high exam scores could achieve exponential growth when compared to lower scoring students.

This hypothesis was supported numerically by considering the theoretical test trajectory of the student who had scored 35.83 on the first exam. Supposing that the student doubled their score for each of the remaining three exams; their total score would only be 286.64, as opposed to 1000 points for the highest scoring student. Even though the lowest scoring student increased their scores multiplicatively, their gap with the higher scoring student, which began at a factor of 6.98 could only be reduced by a factor of 3.49. An even more sobering fact was that the highest scoring student did score 1000 for the four exams. The student with 35.83 did not complete the course.

The path-dependent form of CA was used to learning as the acquisition of new knowledge as a function of a previous time point (DiPrete & Eirich, 2006). The correlation between matched Exam 1 and AKG scores, N = 343, was calculated as the Spearman correlation coefficient, ρ= 0.57002, p = 0.05. A Spearman correlation coefficient is a measure of the monotonic, or nonlinear relationship between two variables.

Since nonlinear functions are more difficult to visualize than linear functions an idealized model was constructed. Nearest neighbor matching was used to construct ideal monotone pairs, with which the model was constructed. The result was a nearly perfect correlation between matched Exam 1 and AKG scores, N = 343: ρ = 0.99995, p = .05. The mathematical fit of the model was exponential, $y = 59.419e^{0.01x}$ and $R^2 = .9961$. The fit was consistent with the common forms of a learning curve (e.g., Evans et al., 2018).

The shape of the learning curve was best described as an increasing returns model, normally associated with economic or productivity growth (Arthur, 1996). An increasing returns model suggested that greater returns to learning would be realized for students with more prior knowledge and experience. For example, prior knowledge is thought to increase learning rates (e.g., Newell, 2001) due to the increased efficiency of multiple cognitive processes (Chi & Glaser, 1988).

The presence of a cumulative advantage process was expected to result in increasing variance associated with performance over time. A Sankey flow diagram was used to demonstrate changes in student ranking from the first exam to the AKG. The Sankey diagram demonstrated three factors associated with CA: (1) the dynamic nature of CA, (2) increasing variance over time, and (3) student rank mobility.

Several trends could be discerned from the Sankey diagram. For example, Groups 2 and 3, due to their position in the diagram, were able to manifest the complete range of variance, in both directions. As a result, both Groups 2 and 3 demonstrated a fan-spread like shape based on rank changes from the Group ranking to final AKG ranking, $N = 343$.

Particular attention was paid to upward student mobility, since chemistry students are implicitly expected to catch up to their better prepared peers in a matter of a few weeks. Such expectations are usually communicated with fill in the blank type statements such "If the student would just...[do something] they could succeed." The blank is usually described in terms of affective or cognitive deficiencies (e.g. Bensimon, 2005; Ferrare & Miller, 2019).

Two instances of significant upward mobility were observed. In the first case, 30 students ranked in Exam 1 Group 2 earned enough points on the last three exams that they were able to end the course in the top AKG category. The move upward demonstrated increased performance on one or more of the last three exams when compared to their initial exam performance. Naturally, the upward mobility of some students required that a commensurate number of students drop in rank.

In all, a total of 43 students were deflected downward from their initial position in Group 1. However, the majority of these students merely dropped to the second AKG rank of AKGs of less than 588. Since grades were decoupled from the data before the analysis, it was not possible to speculate about how this may have affected student course grades.

Larger learning gains were rare, with only 12 students ranked in Groups 3 or 4 earning unusually high gains. Of those twelve students, seven were ranked in Group 3 and five in Group 4. Since Groups 3 and 4 earned less than 208 and 181, respectively, on the first exam, their later performance was impressive. It was possible that the boost in performance could have saved those students from an ignominious GC performance.

Two students in particular performed spectacularly, likely ending the course with at least a B or C. One student from Group 4 earned 160.15 on Exam 1 and an AKG of 720.91, out of the possible 750.00 for a combined score 881.06. The other student from Group 3 earned 184.03 on Exam 1 and an AKG of 735 out of the possible 750 for a total score of 919.99.

While it was presumed the two students worked diligently in the course, large performance improvements over the course of a semester seemed unlikely. As noted by Ericsson and Kintsch (1994) “Researchers have not uncovered some simple strategies that would allow nonexperts to rapidly acquire expert performance, except in a few isolated cases, such as the sexing of chickens” (p. 737). In the intervening years since that statement, no known method of overcoming the biological constraints of brain development has been reported.

A number of other possible scenarios could be considered, each with varying degrees of probability. For example, the Exam 1 scores could have been recorded in error, although that theory could neither be confirmed nor denied. The students may have had an eidetic, or photographic memory. This would allow them to use rote methods of learning, which can be highly effective in some cases. Lastly, the students may have taken AP chemistry in high school, and underestimated their ability to take Exam 1 with little to no review. While the latter explanation was considered the most likely, more research would be needed to appease critics of this conclusion.

Research Limitations

Any research study performed in chemistry, regardless of whether it is performed in the laboratory or classroom, requires replication. A second study using different data would be advised if possible. This does not imply or suggest some deficiency in the original research, only the degree of experimental rigor.

One obvious limitation to the research study is the lack of generalizability, or the ability to make conclusions regarding students other than those described. To demonstrate the validity and even the robustness of the analysis method, replication would be required. Replication by at least three separate investigators at three different educational institutions would be advised.

Another limitation is specific to archival data, which usually represents the “answer” for which the correct question must be framed. Consequently, there can be a sense of working backwards to discover the cause or causes. While it was hypothesized that inadequate high school preparation contributed to poor GC performance, no causal relationship could be established under the conditions of the study.

Implications

It was not possible to identify whether students pursuing a particular major disproportionally represented high or low performing students. However, it was strongly suspected that the highest performing students were likely engineering or physics majors, both of which usually have extensive mathematics backgrounds (Maltese & Tai, 2011; Warne et al., 2018).

In that case, an analysis of institutional records could help clarify that point, particularly high school transcript data and achievement test scores. Of particular interest would be determining the number of years of science and mathematics a student had taken in high school, and whether that affected GC performance. In any event, that step represents the logical progression towards identifying GC students and their high school experiences.

As described earlier, the site of this research was an institution serving a number of surrounding counties, with the bulk of students matriculating from those locations. A comparison of student performance based on high school attended could provide important information. The object would not be to lay blame on a particular high school for inadequate student preparation. Rather, the identification of inadequately sourced school districts would be the objective.

Suggestions

Surprisingly, in many chemistry education research studies, faculty expectations appeared to be above and beyond what is expected even for newly minted chemists. In the limited experience of the author, most newly graduated chemists have had little to no practical experience. Consequently, It was difficult to make sense of faculty expectations, particularly of students taking the course as a prerequisite.

Some chemistry education research studies managed to convey a palpable sense of dislike or even contempt towards the students described. As a consequence, the following remarks and recommendations may have a slight degree of subjectivity. However, any subjectivity was considered offset by the importance of questioning traditional practices related to the education of chemistry students.

The data collected in this study suggested that a significant number of students enrolled in GC were not as prepared for the course as their classmates. Since all students planning to enroll in GC are advised to take a placement test, the mismatch was puzzling. The institution where the study was conducted offers GC in a variety of formats to meet the needs of their students. However, a placement exam is not mandatory, and students are free to enroll in the course of their choice.

The dilemma of mandatory placement exams or enrollment is an issue many institutions grapple with (e.g., Donovan & Wheland, 2009; Mills et al., 2009). For example, literature accounts describe instances where students refused to enroll in a course more closely aligned with their experience (Mills et al. 2009). Conversely, there were faculty who described remediation as remediation as ineffective (Bentley & Gellene, 2005). However, very few expressed remorse for funding GC departments with revenue generated by students who were likely to fail the course.

On the other hand, college students, particularly younger ones, may not be well informed or able to make the best choices for themselves. Many students are anxious to finish postsecondary studies as quickly as possible. They may see a remedial or slower paced course as a mark of shame they are unwilling to endure.

General chemistry serves a number of majors, including the natural sciences, physical sciences and engineering. Based on the widely varying backgrounds of, for example, engineering and biology students, it is questionable whether these students belong in the same class. They are unlikely to have the same mathematics background making some majors more vulnerable than others. Arguably, neither the highest nor lowest performing students in GC are best served by taking the same course.

Conclusion

The presence of a CA process was inferred from empirical GC student data using four major methods: distributional analysis, correlational analysis, modeling, and graphical analysis. A histogram of Exam 1 scores, *N* = 409, indicated there were large differences in preparation and performance. The left-skew of the histogram and the presence of a skewed and fat trailing tail suggested students entered the course with different degrees of preparation (e.g., Perc, 2014).

Students who scored less than 181 points out of the possible 250 on the first exam had almost no chance of catching up. In that sense, Exam 1 scores appeared to lock students into a path they were unable to leave. Many students experienced an ever-widening gap between their performance and their better prepared classmates. Learning is the consequence of permanent changes to long-term memory. How students are expected to accomplish such a task in a single semester is unclear.

The modeling of AKG as a function of first exam score suggested the relationship could be conceptualized as one of increasing returns. The more experienced the student, as indicated by first exam score, the likelier they were to continue earning commensurate scores. However, according to the theory of deliberate practice (Ericsson et al., 1993) increases in performance are monotonic. Consequently, students are not expected to learn at the same rate or in the same way.

While many chemistry faculty assume students are able to surmount large preparation gaps, there was no evidence to support that conclusion. Only 30 students demonstrated what could be considered enough knowledge gain to make a speculative difference in final grades. In general, however, students tended to begin and end in the rank they began in. This was especially true for students who began the course less prepared. In those cases, students who began behind, stayed behind.

References

Albanese, M., Brooks, D. W., Day, V. W., Koehler, R. A., Lewis, J. D., Marianelli, R. S., Rack, E. P., & Tomlinson-Keasey, C. (1976). Piagetian criteria as predictors of success in first year courses. *Journal of Chemical Education*, *53*(9), 571. https://doi.org/10.1021/ed053p571

Allison, P.D., Long, J.S., Krauze, T. K. (1982). Cumulative advantage and inequality in science. *American Sociological Review, 47*, 615–25.

An, B. P. (2022). Formal and social-psychological elements of secondary schooling on college major selection. *Research in Higher Education Journal*, *42*, 1–18.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369-406. https://doi.org/10.1037/0033-295X.89.4.369

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review*, *94*(2), 192–210. https://doi.org/10.1037/0033-295x.94.2.192

Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. Erlbaum.

Anderson, J. R., Betts, S., Bothell, D., Hope, R., & Lebiere, C. (2019). Learning rapid and precise skills. *Psychological Review*, *126*(5), 727–760. https://doi.org/10.1037/rev0000152

Andrews, M. H., & Andrews, L. (1979). First-year chemistry grades. *Journal of Chemistry Education*, *56*(4), 231–232.

Baluyut, J. Y. (2015). *Unpacking Students' Atomistic Understanding of Stoichiometry*. (Publication No. 10009038). [Doctoral Dissertation, Iowa State University]. ProQuest Dissertations and Theses Global.

Barr, D. A., Matsui, J., Wanat, S. F., & Gonzalez, M. E. (2010). Chemistry courses as the turning point for premedical students. *Advances in Health Sciences Education*, *15*(1), 45–54. https://doi.org/10.1007/s10459-009-9165-3

Bauer, C. F. (2005). Beyond "student attitudes": Chemistry self-concept inventory for assessment of the affective component of student learning. *Journal of Chemistry Education*, *82* (12), 1864−1869.

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary Schools? *Child Development*, *83*(4), 1347–1367. https://doi.org/10.1111/j.1467-8624.2012.01779.x

Bäckman, O., & Edling, C. (1999). Mathematical matters: On the absence of mathematical models in quantitative sociology. *Acta Sociologica*, *42*, 69–78.

Bensimon, E. M. (2005). Closing the achievement gap in higher education: An organizational learning perspective. *New Directions for Higher Education*, *2005*(131), 99–111. https://doi.org/10.1002/he.190

Bird, L. (2010). Logical reasoning ability and student performance in general chemistry. *Journal of Chemical Education*, *87*(5), 541–546. https://doi.org/10.1021/ed8001754

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Bloom, B. S. (1968). Learning for mastery. *UCLA Center for the Study of Evaluation of Instructional Programs*, *1*(2), 1–12.

Bloom, B. S. (1984). The 2 sigma problem. The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 3-16.

Boliver, V., & Schindler, S. (2021). Introduction to "Mechanisms of educational stratification." *Research in Social Stratification and Mobility*, *76*, 100641. https://doi.org/10.1016/j.rssm.2021.100641

Bothner, M. S., Haynes, R., Lee, W., & Smith, E. B. (2010). When do Matthew effects occur? *The Journal of Mathematical Sociology*, *34*(2), 80–114. https://doi.org/10.1080/00222500903310960

Bowen, R. S., & Cooper, M. M. (2022). Grading on a curve as a systemic issue of equity in chemistry education. *Journal of Chemical Education*, *99*(1), 185–194. https://doi.org/10.1021/acs.jchemed.1c00369

Bowen, W. G., Kurzweil, M. A., Tobin, E. G. (2005). *Equity and Excellence in American Higher Education.* University of Virginia Press.

Brod, G., Werkle-Bergner, M., & Shing, Y. L. (2013). The influence of prior knowledge on memory: a developmental cognitive neuroscience perspective. *Frontiers in Behavioral Neuroscience*, *7*, 139. https://doi.org/10.3389/fnbeh.2013.00139

Bryson, M. C. (1974). Heavy-tailed distributions: Properties and tests. *Technometrics*, *16*(1), 61–68. https://doi.org/10.1080/00401706.1974.10489150

Buis, J. H. (2016). *Increasing the Understanding of Chemical Concepts: The Effectiveness of Multiple Exposures.* (Publication No. 10129299). [Doctoral Dissertation, University of Kansas]. ProQuest Dissertations and Theses Global.

Bunce, D. M., & Hutchinson, K. D. (1993). The use of the GALT (Group Assessment of Logical Thinking) as a predictor of academic success in college chemistry. *Journal of Chemical Education*, *70*(3), 183. https://doi.org/10.1021/ed070p183

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Carpiano, R. M., & Daley, D. M. (2006). A guide and glossary on postpositivist theory building for population health. *Journal of Epidemiology and Community Health*, *60*(7), 564. https://doi.org/10.1136/jech.2004.031534

Carroll, J. B. (1989). The Carroll model. *Educational Researcher*, *18*(1), 26–31. https://doi.org/10.3102/0013189x018001026

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing. *American Psychologist*, *60*(2), 149–160. https://doi.org/10.1037/0003-066x.60.2.149

Cheng, S. (2014). A life course trajectory framework for understanding the intracohort pattern of wage inequality. *American Journal of Sociology*, *120*(3), 633–700. https://doi.org/10.1086/679103

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–151.

Chi, M. T. H., Glaser, R., & Farr, M.J. (Eds.). (1988). *The Nature of Expertise*. Psychology Press. https://doi.org/10.4324/9781315799681

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828–833. https://doi.org/10.1126/science.229.4716.828

Coley N. R. (1973), Prediction of success in general chemistry in a community college, *Journal of Chemistry Education, 50*(9), 613.

Conlin, K. (2023, March 14). Ten hardest college majors in 2023. *College Transitions.* *https://www.collegetransitions.com/blog/hardest-college-majors/*

Cook, T. D. & Campbell, D. T. (1979) *Quasi-experimentation*. Rand-McNally.

Cooper, C. I., & Pearson, P. T. (2012). A genetically optimized predictive system for success in general chemistry using a diagnostic algebra test. *Journal of Science Education and Technology*, *21*(1), 197–205. https://doi.org/10.1007/s10956-011-9318-z

Cooper, M. M., & Stowe, R. L. (2018). Chemistry education research—From personal empiricism to evidence, theory, and informed practice. *Chemical Reviews*, *118*(12), 6053–6087. https://doi.org/10.1021/acs.chemrev.8b00020

Cooper, M. M. & Klymkowsky, M. (2020, November 18). Curbing the malpractice of curved grades and high-stakes exams. ASBMB Today. https://www.asbmb.org/asbmb-today/opinions/111820/curbing-the-malpractice-of-curved-grades-and-high

Cornog, J., & Stoddard, G. D. (1925). Predicting performance in chemistry. *Journal of Chemical Education*, *2*(8), 701. https://doi.org/10.1021/ed002p701

Cracolice, M. S., & Busby, B. D. (2015). Preparation for college general chemistry: More than Just a matter of content knowledge acquisition. *Journal of Chemical Education*, *92*(11), 1790–1797. https://doi.org/10.1021/acs.jchemed.5b00146

Davenport, J. L., Yaron, D., Koedinger, K. R., & Klahr, D. (2008). Development of conceptual understanding and problem solving expertise in chemistry. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *The 30th Annual Conference of the Cognitive Science Society* (pp. 1699–1704). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.485.4188&rep=rep1&type=pdf

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, *32*(1), 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Donovan, W. J., & Wheland, E. R. (2009). Comparisons of success and retention in a general chemistry course before and after the adoption of a mathematics prerequisite. *School Science and Mathematics*, *109*(7), 371–382. https://doi.org/10.1111/j.1949-8594.2009.tb17868.x

Dornbusch, S. M., Glasgow, K. L., and Lin, I.-C. (1996). The social structure of schooling. *Psychology*, *47*(1), 401–429. https://doi.org/10.1146/annurev.psych.47.1.401

Duignan, B. (2023, August 15). Occam's Razor. Britannica.com https://www.britannica.com/topic/Occams-razor

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037//0033-295x.100.3.363

Ericsson, K. A., & Charness, N. (1994). Expert performance. *American Psychologist*, *49*(8), 725–747. https://doi.org/10.1037/0003-066x.49.8.725

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295x.102.2.211

Ericsson, K. A. (2017). Expertise and individual differences: the search for the structure and acquisition of experts' superior performance. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1–2), e1382. https://doi.org/10.1002/wcs.1382

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*(4), 592–605. https://doi.org/10.1037/rev0000105

Fendler, L., & Muzaffar, I. (2008). The history of the bell curve: sorting and the idea of normal. *Educational Theory*, *58*(1), 63–82. https://doi.org/10.1111/j.1741-5446.2007.0276.x

Ferrare, J. J., & Miller, J. M. (2019). Making sense of persistence in scientific purgatory: a multi-institutional analysis of instructors in introductory science, technology, engineering, and mathematics (STEM) courses. *The Journal of Higher Education*, *91*(1), 1–26. https://doi.org/10.1080/00221546.2019.1602392

Finney, R. M. (2008). *Student categorization of general chemistry problems*. (Publication No. 3348788). [Doctoral Dissertation, University of Northern Colorado]. ProQuest Dissertations and Theses Global.

Fredrick, W. C., & Walberg, H. J. (1980). Learning as a function of time. *The Journal of Educational Research*, *73*(4), 183–194. https://doi.org/10.1080/00220671.1980.10885233

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The Nature of Expertise* (pp. xv-xxviii). Erlbaum.

Grodsky, E., Warren, J. R., & Felts, E. (2008). Testing and social stratification in American education. *Annual Review of Sociology*, *34*(1), 385–404. https://doi.org/10.1146/annurev.soc.34.040507.134711

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research*. Sage Publications, p. 105–117.

Harris, R. B., Mack, M. R., Bryant, J., Theobald, E. J., & Freeman, S. (2020). Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a “hyperpersistent zone.” *Science Advances*, *6*(24), 108. https://doi.org/10.1126/sciadv.aaz5687

Hasl, A., Voelkle, M., Kretschmann, J., Richter, D., & Brunner, M. (2022). A dynamic structural equation approach to modeling wage dynamics and cumulative advantage across the lifespan. *Multivariate Behavioral Research*, 1–22. https://doi.org/10.1080/00273171.2022.2029339

Hawks, M. (2015, May 14). Educational testing. https://eugenicsarchive.ca/discover/tree/5554c24335ae9d9e7f0000af

Hedström, P. & Swedberg, R. (2010). *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press.

Herron, J. D., & Greenbowe, T. J. (1986). What can we do about Sue: A case study of competence. *Journal of Chemical Education*, *63*(6), 528–531.

Hovey, N. W., & Krohn, A. (1958). Predicting failures in general chemistry. *Journal of Chemical Education*, *35*(10), 507. https://doi.org/10.1021/ed035p507

Irvin, M., Byun, S., Smiley, W. S., & Hutchins, B. C. (2017). Relation of opportunity to learn advanced math to the educational attainment of rural youth. *American Journal of Education*, *123*(3), 475–510. https://doi.org/10.1086/691231

Johnstone A.H., (1982), Macro and microchemistry, *School Science Review*, 64, 377-379.

Johnstone, A. H., & Elbanna, H. (1986). Capacities, demands and processes: A predictive model for science education. *Education in Chemistry, 123*(3), 80–84.

Johnstone A. H., (1991), Why is science difficult to learn? Things are seldom what they seem, *Journal of Computer Assisted Learning*, 7, 75-83. https://doi.org/10.1111/j.1365-2729.1991.tb00230.x

Johnstone, A. H. (1993). The development of chemistry teaching. *Journal of Chemical Education*, *70*(9), 701–705.

Johnstone, A. H. (1997). Chemistry teaching - Science or alchemy? 1996 Brasted Lecture. *Journal of Chemical Education*, *74*(3), 262. https://doi.org/10.1021/ed074p262

Johnstone, A. H. (2006). Chemical education research in Glasgow in perspective. *Chemistry Education Research and Practice*, *7*(2), 49–63. https://doi.org/10.1039/b5rp90021b

Johnstone, A. H. (2010). You can't get there from here. *Journal of Chemical Education, 87*(1), 22–29. https://doi.org/10.1021/ed800026d

Jordan, M. & Achenbach, J. (1993, February 11). By any other name, still the SAT. *Washington Post*. https://www.washingtonpost.com/archive/politics/1993/02/11/by-any-other-name-still-the-sat/6d82b289-e5c0-4b6f-9651-d287c6513be0/

Jong, T. de, & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist*, *31*(2), 105–113.

Kerckhoff, A. C., & Glennie, E. (1999). The Matthew effect in American education. *Research in Sociology of Education and Socialization*, *12*, 35–36.

Kennedy, A. B. W., & Sankey, H. R. (1898). The thermal efficiency of steam engines. Report of the Committee Appointed to the Council Upon the Subject of the Definition of a Standard or Standards of Thermal Efficiency for Steam Engines. Available from Zenodo.org/record/2036163. https://doi.org/10.1680/imotp.1898.19100

Koch, A. K., & Drake, B. M. (2018). Digging into the Disciplines: The Impact of Gateway Courses in Accounting, Calculus, and Chemistry on Student Success. *John N. Gardner Institute for Excellence in Undergraduate Education*.

Kotnik, L. J. (1974), What is being done to help the underprepared student? *Journal of Chemical Education, 51*(3), 165–167.

Kozma, R. B., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, *34*(9), 949–968.

Kileen, P. R. (1999). Modeling modeling. *Journal of the Experimental Analysis of Behavior*, *71*, 275–280.

Kuhfeld, M., & Soland, J. (2021). The Learning Curve: Revisiting the Assumption of Linear Growth during the School Year. *Journal of Research on Educational Effectiveness*, *14*(1), 143–171. https://doi.org/10.1080/19345747.2020.1839990

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*(4450), 1335–1342.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*(1), 11–24. https://doi.org/10.1002/tea.3660150103

Lawson, D., & Marion, G. (1995). An introduction to mathematical modelling. *Choice Reviews Online*, 1–32. https://doi.org/10.5860/choice.32-5134

Leech (2002). A call for greater use of nonparametric statistics. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Chattanooga, TN, November 6-8.

Legg, J. C., Greenbowe, T. J., & Legg, M. J. (2001). Analysis of success in general chemistry based on diagnostic testing using logistic regression. *Journal of Chemical Education*, *78*(8), 1117-1121. https://doi.org/10.1021/ed078p1117

Letukas, L. (2015). Nine facts about the SAT that might surprise you. College Board Research. https://eric.ed.gov/?id=ED562751

Lewis, S. E., & Lewis, J. E. (2007). Predicting at-risk students in general chemistry: comparing formal thought to a general achievement measure. *Chemistry Education Research and Practice, 8*(1), 32–51. https://doi.org/10.1126/science.208.4450.1335

Lindquist, E. F. (1970). The Iowa Testing Programs - A retrospective view. *Education*, *90*(4), 13–19.

Long, M., Conger, D., & Iatarola, P. (2012). Effects of high school course taking on secondary and postsecondary success. *American Educational Research Journal, 49* (2): 285–322.

Mackenzie, N. & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues In Educational Research*, 16(2), 193-205. http://www.iier.org.au/iier16/mackenzie.html

Mack, M. R., Hensen, C., and Barbera, J. (2019). Metrics and methods used to compare student performance data in chemistry education research articles. *Journal of Chemical Education*, *96*(3), 401–413. https://doi.org/10.1021/acs.jchemed.8b00713

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. *Science Education*, *95*(5), 877–907. https://doi.org/10.1002/sce.20441

Martin, F. D. (1942). A diagnostic and remedial study. *Journal of Chemical Education*, *19*, 274–277.

Mast, J. de, Steiner, S. H., Nuijten, W. P. M., & Kapitan, D. (2023). Analytical Problem Solving Based on Causal, Correlational and Deductive Models. *The American Statistician*, *77*(1), 51–61. https://doi.org/10.1080/00031305.2021.2023633

Mata, L. E. (2020). *The Effect of POGIL on Chemistry EOC scores and ACT Science Scores.* (Publication No. 28088231). [Doctoral Dissertation, Grand Canyon University]. ProQuest Dissertations and Theses Global.

Mazur, J. E. (2006). Mathematical models and the experimental analysis of behavior. *Journal of the Experimental Analysis of Behavior*, *85*(2), 275–291. https://doi.org/10.1901/jeab.2006.65-05

Mazur, J. E., & Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin*, *85*(6), 1256–1274. https://doi.org/10.1037/0033-2909.85.6.1256

McFate, C., & Olmsted, J. (1999). Assessing student preparation through placement tests. *Journal of Chemical Education*, *76*(4), 562. https://doi.org/10.1021/ed076p562

Merriam Webster Dictionary (2023). "Aptitude." Merriam-Webster.com. Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/aptitude

Merton, R. K. (1942/1973). *The Sociology of Science: Theoretical and Empirical Investigations.* University of Chicago Press.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. https://www.jstor.org/stable/1723414

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis, 79*(4), 606–623. http://www.jstor.org/stable/234750

Mills, P., Sweeney, W., & Bonner, S. M. (2009). Using the first exam for student placement in beginning chemistry courses. *Journal of Chemical Education*, *86*(6), 738–743.

Morgan, T. L., Zakem, D., & Cooper, W. L. (2018). From high school access to postsecondary success: an exploratory study of the impact of high-rigor coursework. *Education Sciences*, *8*(4), 191. https://doi.org/10.3390/educsci8040191

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal Chemical Education, 79*(6). 739-744. https://doi.org/10.1021/ed079p739

Nakhleh, M. B., & Mitchell, R. C. (1993). Concept Learning versus Problem Solving. *Journal of Chemical Education*, *70*(3), 190–192.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* The National Academies Press.

National Research Council. (2013). *A Framework for K-12 Science Education: Appendix F.* The National Academies Press. https://www.nextgenscience.org/sites/default/files/Appendix%20F%20%20Science%20and%20Engineering%20Practices%20in%20the%20NGSS%20-%20FINAL%20060513.pdf

National Research Council. (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering.* The National Academies Press. https://doi.org/10.17226/13362.

National Science Board (2018). *National Science Indicators 2018*. National Science Foundation.

NIST/SEMATECH e-Handbook of Statistical Methods (2023). http://www.itl.nist.gov/div898/handbook/

Newell, K. M., Mayer-Kress, G., & Liu, Y.-T. (2001). Time scales in motor learning and development. *Psychological Review*, *108*(1), 57–82.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Novak, E. (2022). Mathematical modeling for theory-oriented research in educational technology. *Educational Technology Research and Development*, *70*(1), 149–167. https://doi.org/10.1007/s11423-021-10069-6

Nurrenbern, S., & Pickering, M. C. (1987). Concept learning versus problem solving is there a difference? *Journal of Chemical Education*, *64*(6), 508–510.

Orfield, G. (2009). *Reviving the Goal of an Integrated Society: A 21st Century Challenge, Los Angeles, CA.* (pp. 1–33). The Civil Rights Project/Proyecto Derechos Civiles at UCLA. https://civilrightsproject.ucla.edu/research/k-12-education/integration-and-diversity/reviving-the-goal-of-an-integrated-society-a-21st-century-challenge/orfield-reviving-the-goal-mlk-2009.pdf

Ozsogomonyan, A., & Loftus, D. (1979). Predictors of general chemistry grades. *Journal of Chemical Education*, *56*(3), 173. https://doi.org/10.1021/ed056p173

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface*, *11*(98), 20140378. https://doi.org/10.1098/rsif.2014.0378

Peltokorpi, J., & Jaber, M. Y. (2022). An interference-adjusted power learning curve for tasks with cognitive and motor elements. *Applied Mathematical Modelling*, *101*, 157–170. https://doi.org/10.1016/j.apm.2021.08.016

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development*, *51*(1), 40–47. https://doi.org/10.1159/000112531

Pickering, M. (1990). Further Studies on Concept Learning. *Journal of Chemical Education*, *67*(3), 254–255.

Powell, C. B., Simpson, J., Williamson, V. M., Dubrovskiy, A., Walker, D. R., Jang, B., Shelton, G. R., & Mason, D. (2020). Impact of arithmetic automaticity on students' success in second-semester general chemistry. *Chemistry Education Research and Practice*, *21*(4), 1028–1041. https://doi.org/10.1039/d0rp00006j

Price, P. C., Rajiv, J., Chiang, I-C. A., Leighton, D. C., Cutler, C. (2023). *Research Methods in Psychology*, Third Edition. Pressbooks. http://www.saylor.org/site/textbooks/

Ralph, V. R., & Lewis, S. E. (2018). Chemistry topics posing incommensurate difficulty to students with low math aptitude scores. *Chemistry Education Research and Practice*, *19*(3), 867–884. https://doi.org/10.1039/c8rp00115d

Ralph, V. R., & Lewis, S. E. (2019). An explanative basis for the differential performance of students with low math aptitude in general chemistry. *Chemistry Education Research and Practice*, *20*(3), 570–593. https://doi.org/10.1039/c9rp00068b

Ralph, V. R., Scharlott, L. J., Schafer, A. G. L., Deshaye, M. Y., Becker, N. M., & Stowe, R. L. (2022). Advancing equity in STEM: The impact assessment design has on who succeeds in undergraduate introductory chemistry. *JACS Au*, *2*(8), 1869–1880. https://doi.org/10.1021/jacsau.2c00221

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. G. J. Duncan & R. J. Murnane, Eds. p. 91–115. Russell Sage Foundation.

Rigney, D. (2010). The Matthew Effect: How Advantage Begets Further Advantage. Columbia University Press.

Rosen, S. (1956). A century of high-school science. *The Science Teacher*, *23*(7), 321–324. https://www.jstor.org/stable/24143179

Ryu, M., Bano, R., & Wu, Q. (2021). Where does CER stand on diversity, equity, and inclusion? Insights from a literature review. *Journal of Chemical Education*, *98*(12), 3621–3632. https://doi.org/10.1021/acs.jchemed.1c00613

Sadler, P. M., & Tai, R. H. (2007). Advanced placement exam scores as a predictor of performance in introductory college biology, chemistry and physics courses. science educator, 16(2), 1–19.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2014). The role of advanced high school coursework in increasing STEM career interest. *Science Educator*, *23*(1), 1–13.

Sadler, P., & Sonnert, G. (2018). The path to college calculus: The impact of high school mathematics coursework. *Journal for Research in Mathematics Education*, *49*(3), 292–329. https://doi.org/10.5951/jresematheduc.49.3.0292

Salehi, S., Burkholder, E., Lepage, G. P., Pollock, S., and Wieman, C. (2019). Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics. *Physical Review Physics Education Research*, *15*(2), 020114. https://doi.org/10.1103/physrevphyseducres.15.020114

Saltelli, A. (2019). A short comment on statistical versus mathematical modelling. *Nature Communications*, *10*(1), 3870. https://doi.org/10.1038/s41467-019-11865-8

Saltelli, A., and Puy, A. (2023). What can mathematical modelling contribute to a sociology of quantification? *Humanities and Social Sciences Communications*, *10*(1), 213. https://doi.org/10.1057/s41599-023-01704-z

Santos, D. L., Gallo, H., Barbera, J., & Mooring, S. R. (2021). Student perspectives on chemistry intelligence and their implications for measuring chemistry-specific mindset. *Chemistry Education Research and Practice*. 4, 1-18. https://doi.org/10.1039/d1rp00092f

Sawrey, B. A. (1990). Concept learning versus problem solving: Revisited. *Journal of Chemical Education*, *67*(3), 253–254.Simon

Scerri, E. R. (2003). Philosophical confusion in chemical education research. *Journal of Chemical Education*, *80*(5), 468–474.

Shadish, W., Cook, T., Campbell, T. (2002). *Quasi-Experimental Design for Generalized Causal Inference*, Houghton Mifflin.

Shah, L., Basner, E. B., Ferraro, K., Sajan, A., Fatima, A., & Rushton, G. T. (2020). Diversifying undergraduate chemistry course pathways to improve outcomes for at-risk students. *Journal of Chemical Education*, *97*(7), 1822–1831. https://doi.org/10.1021/acs.jchemed.9b00972

Shah, L., Fatima, A., Syed, A., & Glasser, E. (2022). Investigating the impact of assessment practices on the performance of students perceived to be at risk of failure in second-semester general chemistry. *Journal of Chemical Education*, *99*(1), 14–24. https://doi.org/10.1021/acs.jchemed.0c01463

Shelton, G. R., Mamiya, B., Weber, R., Walker, D. R., Powell, C. B., Jang, B., Dubrovskiy, A. V., Villalta-Cerdas, A., & Mason, D. (2021). Early warning signals from automaticity diagnostic instruments for first- and second-semester general chemistry. *Journal of Chemical Education*, *98*(10), 3061–3072. https://doi.org/10.1021/acs.jchemed.1c00714

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill Book Company.

Simmons, A. B., & Heckler, A. F. (2020). Grades, grade component weighting, and demographic disparities in introductory physics. *Physical Review Physics Education Research*, *16*(2), 020125. https://doi.org/10.1103/physrevphyseducres.16.020125

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*(3), 425–440.

Simon, H. A., & Chase, W G. (1973). Skill in chess. *American Scientist, 61*, 394-403.

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., and Schneider, M. (2021). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, 1–24. https://doi.org/10.1080/00461520.2021.1939700

Spencer, H. E. (1996). Mathematical SAT test scores and college chemistry grades. *Journal of Chemical Education*, *73*(12), 1150-1153. https://doi.org/10.1021/ed073p1150

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: the case for emphasizing chemistry on chemistry assessments. *Journal of Chemical Education*, *98*(8), 2490–2495. https://doi.org/10.1021/acs.jchemed.1c00532

Stuart E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science 25*(1):1-21. https://projecteuclid.org/journals/statistical-science/volume-25/issue-1/Matching-Methods-for-Causal-Inference--A-Review-and-a/10.1214/09-STS313.full

Streitberger, E. (1977). What should we be teaching them in high school chemistry? *The Science Teacher*, *44*(8), 35–37. https://www.jstor.org/stable/24128595

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

Tai, R. H., Sadler, P. M., & Loehr, J. F. (2005). Factors influencing success in introductory college chemistry. *Journal of Research in Science Teaching*, *42*(9), 987–1012. https://doi.org/10.1002/tea.20082

Tai, R. H., Ward, R. B., & Sadler, P. M. (2006). High school chemistry content background of introductory college chemistry students and its association with college chemistry grades. *Journal of Chemical Education*, *83*(11), 1703–1711.

Tai, R. H., & Sadler, P. M. (2009). Same science for all? Interactive association of structure in learning activities and academic attainment background on college science performance in the USA. *International Journal of Science Education*, *31*(5), 675–696. https://doi.org/10.1080/09500690701750600

Talanquer, V. (2006). Commonsense Chemistry. *Journal of Chemical Education*, *83*(5), 811–816. https://doi.org/10.1021/ed083p811

Talanquer, V. (2015). Threshold Concepts in Chemistry: The Critical Role of Implicit Schemas. *Journal of Chemical Education*, *92*(1), 3–9. https://doi.org/10.1021/ed500679k

Tashiro, J., & Talanquer, V. (2021). Exploring inequities in a traditional and a reformed general chemistry course. *Journal of Chemical Education*, *98*(12), 3680–3692. https://doi.org/10.1021/acs.jchemed.1c00821

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, *26*(2), 265–283.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.

Vyas, V. S., & Reid, S. A. (2023). What moves the needle on DFW rates and student success in general chemistry? A Quarter-Century Perspective. *Journal of Chemical Education*, *100*(4), 1547–1556. https://doi.org/10.1021/acs.jchemed.2c01121

Wagner, E. P., Sasser, H., & DiBasse, W. J. (2002). Predicting students at risk in general chemistry using pre-semester assessments and demographic information. *Journal of Chemical Education, 79*(6), 749-755. https://doi.org/https://doi.org/10.1021/ed079p749

Walker, T., & Rosales, J. (2021). The racist beginnings of standardized testing. *NEA Today*. https://www.nea.org/advocating-for-change/new-from-nea/racist-beginnings-standardized-testing

Wang, Y., Rocabado, G. A., Lewis, J. E., & Lewis, S. E. (2021). Prompts to promote success: evaluating utility value and growth mindset interventions on general chemistry students' attitude and academic performance. *Journal of Chemical Education*, *98*(5), 1476–1488. https://doi.org/10.1021/acs.jchemed.0c01497

Warne, R. T., Sonnert, G., & Sadler, P. M. (2018). The relationship between advanced placement mathematics courses and students' STEM career interest. *Educational Researcher*, *48*(2), 101–111. https://doi.org/10.3102/0013189x19825811

Wells, R. S., Manly, C. A., Kommers, S., & Kimball, E. (2019). Narrowed gaps and persistent challenges: examining rural-nonrural disparities in postsecondary outcomes over time. *American Journal of Education*, *126*(1), 1–31. https://doi.org/10.1086/705498

Weston, T. J., Seymour, E., Koch, A. K., & Drake, B. M. (2019). Weed-out classes and their consequences. In Seymour, E., Hunter, A. B., (Eds.) *Talking about Leaving Revisited*; Seymour, E., Hunter, A. B., Eds. (197−244). Springer: Basel, Switzerland.

White, P., & Mitchelmore, M. C. (1996). Conceptual knowledge in introductory calculus. *Journal for Research in Mathematics Education*, *27*, 79–95.

Wildemuth B. M. (1993). Post-positivist research: Two examples of methodological pluralism. *The Library Quarterly: Information, Community, Policy, 63*(4), 450–468. http://www.jstor.org/stable/4308866

Wright, T. P. (1936/ 2012). Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences (Institute of the Aeronautical Sciences)*, *3*(4), 122–128. https://doi.org/10.2514/8.155

Xie, Y., Fang, M., & Shauman, K. (2015). STEM Education. *Annual Review of Sociology*, *41*(1), 1–27. https://doi.org/10.1146/annurev-soc-071312-145659

Yuan, K.-H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis. *Sociological Methods & Research*, *34*(2), 240–258. https://doi.org/10.1177/0049124105280200

Ye, L., Shuniak, C., Oueini, R., Robert, J., & Lewis, S. (2016). Can they succeed? Exploring at-risk students' study habits in college general chemistry. *Chemistry Education Research and Practice, 17*(4), 878–892. https://doi.org/10.1039/c6rp00101g

Zimmerman, J. (2022, October 10). 'Rigorous' and 'Weed-Out' Are Not Synonymous. Inside Higher Ed. https://www.insidehighered.com/views/2022/10/11/weed-out-courses-need-be-rethought-opinion

Appendix 1: IRB #2065542

IRB Determination Notice Project #2065542 Review #329621

UMSL eCompliance <ecompliance-do-not-reply@umsl.ecompliance.umsystem.edu>
via missouri.edu

Thu 11/11/2021 4:09 PM

To: Barvian, Nicole (UMSL-Student) <ncbh3r@mail.umsl.edu>;Miller, Keith <keith.w.miller@umsl.edu>

UMSL eCompliance

IRB Determination Notice Project #2065542 Review #329621

Project #2065542
Project Title: The Role of Prior Chemistry Knowledge on Postsecondary Introductory Chemistry Student Performance
Principal Investigator: Nicole Barvian (UMSL-Student)
Primary Contact: Nicole Barvian (UMSL-Student)

Dear Investigator,

The UMSL Institutional Review Board reviewed your application and supportive documents. It has been determined that this project does not constitute human subjects research according to the Department of Health and Human Services regulatory definitions. As such, there are no further IRB requirements.

If you have questions, please feel free to contact the UMSL IRB office at 314-516-5972 or email at irb@umsl.edu.

Sincerely,

UMSL Institutional Review Board

Appendix 2: Preliminary Statistics

Statistics for All Initial Exam Data

Table 1

Summary Statistics All Initial Exam Data N= 409

Group	Variable	*N*	Mean	Std Dev	Skew	Kurtosis	95% CI	
							Lower	Upper
F 2016	Exam 1	85	197.60	41.94	-1.57	3.25	188.56	206.65
	Exam 2	75	152.63	48.18	-0.31	-0.42	141.54	163.72
	Exam 3	74	176.23	50.38	-0.66	-0.25	164.56	187.90
	Exam 4	69	143.75	57.07	-0.11	-1.05	130.04	157.45
S 2017	Exam 1	52	193.55	39.59	-1.10	0.44	182.52	204.57
	Exam 2	48	139.94	54.98	-0.06	-0.79	123.98	155.91
	Exam 3	48	163.48	61.65	-0.69	-0.38	145.58	181.38
	Exam 4	43	133.40	57.90	0.10	-0.96	115.58	151.22
F 2017	Exam 1	90	191.73	45.62	-1.31	1.18	182.17	201.28
	Exam 2	82	154.82	51.24	-0.35	-0.08	143.56	166.08
	Exam 3	77	176.86	54.30	-0.79	0.03	164.54	189.19
	Exam 4	71	147.50	54.82	0.01	-0.99	134.52	160.47
S 2018	Exam 1	68	193.23	41.73	-0.80	0.05	183.12	203.33
	Exam 2	63	137.74	56.31	-0.08	-0.96	123.56	151.93
	Exam 3	66	169.27	52.74	-0.66	-0.17	156.30	182.23
	Exam 4	63	138.68	61.15	0.01	-1.02	123.28	154.08
F 2018	Exam 1	57	199.86	41.19	-1.62	3.55	188.93	210.79
	Exam 2	54	151.63	47.93	-0.85	0.80	138.55	164.71
	Exam 3	55	170.61	54.65	-0.62	-0.32	155.84	185.39
	Exam 4	54	146.57	57.04	-0.33	-0.71	131.00	162.14
S 2019	Exam 1	57	180.75	48.37	-0.55	-0.08	167.92	193.59
	Exam 2	52	148.16	59.51	-0.20	-0.99	131.59	164.72
	Exam 3	52	172.87	56.92	-0.90	0.06	157.02	188.72
	Exam 4	48	167.01	59.95	-0.66	-0.36	149.61	184.43

Table 2

Goodness-of-Fit Tests for Normal Distribution, N = 409

Group	*N*	D-Statistic	p Value
1	85	0.10655	0.018
2	52	0.15020	< 0.010
3	90	0.15939	< 0.010
4	68	0.10242	0.077
5	57	0.16096	< 0.010
6	57	0.09851	> 0.150

Statistics for All Matched Exams 1-4, *N* = 343

Table 3

Summary Statistics All Matched 1-4 Exams N= 343

Group	Variable	*N*	Mean	Std Dev	Skew	Kurtosis	95% CL Lower	95% CL Upper
	Exam 1	69	205.87	31.63	-1.37	4.28	198.27	213.47
	Exam 2	69	154.42	47.65	-0.34	-0.26	142.98	165.87
F 2016	Exam 3	69	178.78	47.34	-0.49	-0.66	167.40	190.16
	Exam 4	69	143.75	57.07	-0.11	-1.05	130.04	157.45
	AKG	69	479.95	139.01	-0.25	-0.84	443.56	510.34
	Exam 1	43	201.39	33.86	-1.24	1.04	190.97	211.81
	Exam 2	43	145.63	53.12	-0.11	-0.65	129.28	161.97
S 2017	Exam 3	43	174.11	54.72	-0.90	0.46	157.27	190.94
	Exam 4	43	133.40	57.90	0.10	-0.96	115.58	151.22
	AKG	43	453.13	143.76	- 1.10	- 0.36	408.89	497.37
	Exam 1	71	203.28	33.36	-1.40	2.75	195.38	211.18
	Exam 2	71	160.59	50.23	-0.45	0.20	148.70	172.48
F 2017	Exam 3	71	178.58	54.49	-0.83	0.13	165.69	191.48
	Exam 4	71	147.50	54.82	0.01	-0.99	134.52	160.47
	AKG	71	486.67	144.34	- 0.40	- 0.11	452.51	520.84
	Exam 1	61	197.80	39.24	0.95	0.76	187.75	207.85
	Exam 2	61	139.79	55.12	-0.04	-1.07	125.67	153.90
S 2018	Exam 3	61	172.80	49.72	-0.69	0.23	160.07	185.54
	Exam 4	61	138.32	61.55	0.02	-1.02	122.56	154.08
	AKG	61	446.81	150.53	0.01	- 0.89	408.26	485.36
	Exam 1	53	204.45	34.75	-1.11	0.77	194.87	214.02
	Exam 2	53	151.79	48.37	-0.85	0.75	138.46	165.12
F 2018	Exam 3	53	173.20	53.02	-0.63	-0.28	158.59	187.81
	Exam 4	53	146.15	57.51	-0.30	-0.75	130.30	162.00
	AKG	53	470.60	139.30	- 0.54	- 0.05	432.21	509.00
	Exam 1	46	190.44	42.45	-0.37	-0.78	177.83	203.04
	Exam 2	46	155.91	55.26	-0.17	-1.00	139.50	172.32
S 2019	Exam 3	46	178.69	52.97	-0.93	0.20	162.96	194.42
	Exam 4	46	170.16	56.67	-0.52	-0.77	153.33	186.98
	AKG	46	504.76	144.31	- 0.42	- 0.71	461.90	547.61

Table 4

Goodness-of-Fit Tests for Normal Distribution, N = 343

Group	*N*	D-Statistic	Pr>D
F2016	69	0.08245	> 0.150
		0.08664	> 0.150
S2017	43	0.16791	< 0.010
		0.07895	> 0.150
F2017	71	0.10191	0.068
		0.04623	> 0.150
S2018	61	0.10696	0.082
		0.06855	> 0.150
F2018	53	0.16150	< 0.010
		0.10272	> 0.150
S2019	46	0.09564	> 0.150
		0.11915	0.098

Table 5

Kruskal-Wallis One Way Analysis of Variance First Exam Scores By Group

Group	*N*	Sum of Scores	Expected H_0	Std Dev H_0	Mean Score
F2016	85	18380.5	17425	970	216.24
S2017	52	10561.5	10660	796	203.11
F2017	90	18327.0	18450	990	203.63
S2018	68	13891.0	13940	890	204.28
F2018	57	12830.0	11685	828	225.09
S2019	57	9855.0	11685	828	172.89

Table 6

Kruskal-Wallis One Way Analysis of Variance First Exam Scores By Group

Group	N	Label	Sum of Scores	Expected H_0	Std Dev H_0	Mean Score
F2016	69	Exam 1	12523.0	11868	736.15	181.49
		AKG	12052.0	11868	736.19	174.67
S2017	43	Exam 1	7387.0	7396	608.08	171.79
		AKG	6710.0	7396	744.05	156.04
F2017	71	Exam 1	12474.5	12212	744.01	175.70
		AKG	12868.5	12212	702.13	181.25
S2018	61	Exam 1	10148.0	10492	702.19	166.36
		AKG	9370.5	10492	702.23	153.61
F2018	53	Exam 1	9667.5	9116	663.75	182.41
		AKG	9086.5	9116	663.78	171.44
S2019	46	Exam 1	6796.0	7912	625.78	147.74
		AKG	8908.5	7912	625.81	193.66