

DEVELOPMENT AND EVALUATION OF A PROFESSIONAL TRAINING
MODULE FOR CHEMISTRY GRADUATE TEACHING ASSISTANTS

by

GAURI RAMASUBRAMANIAN

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2022

THESIS APPROVAL PAGE

Student: Gauri Ramasubramanian

Title: Development and Evaluation of a Professional Training Module for Chemistry Graduate Teaching Assistants.

This thesis has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Chemistry and Biochemistry by:

Mike Haley	Chairperson
Thomas J Greenbowe	Advisor
Deborah Exton	Core Member
Juliet Baxter	Institutional Representative
Jared Danielson	External Member

and

Krista Chronister	Vice Provost for Graduate Studies
-------------------	-----------------------------------

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2022.

© 2022 Gauri Ramasubramanian
This work is licensed under a Creative Commons
Attribution-Noncommercial-NoDerivs (United States) License



DISSERTATION ABSTRACT

Gauri Ramasubramanian

Doctor of Philosophy

Department of Chemistry and Biochemistry

June 2022

Title: Development and Evaluation of a Professional Training Module for Chemistry Graduate Teaching Assistants.

Graduate teaching assistants (GTAs) are an integral part of the instructional personnel in undergraduate general chemistry courses. Professional development programs for GTAs are well-established across STEM disciplines ranging from two-day to week-long sessions, covering a wide variety of topics, ranging from pedagogy and safety standards to duties, expectations, and resources as a university employee. Chemistry Education research (CER) is an evidence-based approach to chemistry teaching and learning. Assessment of weekly laboratory reports is one of the primary responsibilities of Chemistry GTAs. Professional development (PD) of GTAs specifically for assessment of student work is a largely unexplored area in CER.

My doctoral research in CER aims to address this gap and explores professional development modules designed to help GTAs become reliable and consistent graders. This dissertation describes the design and implementation of specialized grading activities that enhance GTAs' understanding of grading criteria, chemistry misconceptions, and common technical errors in undergraduate students' writing. Program evaluation is a rare feature in such initiatives. This work is the first-known report of using growth models as a novel statistical approach for evaluation of GTAs or GTA training programs. I examine factors

influencing chemistry GTAs' professional development as reliable graders. This longitudinal research was informed by extensive individual and group interactions with GTAs, needs assessments, technological, pedagogical, and online learning management system (LMS) tools to support GTAs in developing reliable and consistent grading skills.

ACKNOWLEDGMENTS

I wish to express sincere appreciation to Professors Greenbowe and Exton for their guidance and encouragement throughout my doctoral program. The design, and rigorous data collection implementation of protocols in my chemistry education research study was possible largely due to the generous allocation of time, effort, resources on their part along with a modicum of autonomy as a chemistry educator. Words alone will not suffice to express my gratitude to my PhD committee, Professors Haley (Chair), Baxter, and Danielson (ISU) members who have supported my pursuits patiently and unequivocally during transitions between Oregon and Iowa and also through parenthood. The staff members and laboratory preparators at University of Oregon have my deepest thanks for their willingness to navigate and support my research to the best of their abilities.

I owe a deep debt of gratitude to Dr. Burke, Dr. Fernando, Dr. Burnett, and Dr. Kingston from the Department of Chemistry at Iowa State University for their valuable guidance during the preliminary stages of my doctoral work and continued support through my dissertation. My deepest thanks to all the graduate and undergraduate teaching assistants at Iowa State University and University of Oregon for their participation in my research and for their thoughtful inputs as my peers. Every interaction with teaching assistants has molded my skills and perspectives in so many ways, and only increased my awe at the multi-faceted potential of the TA diaspora around me. There are always individuals whose contributions remain unnamed in every glorious journey seeking new knowledge and overcoming hurdles. A Stephen, L Werbel and S Zuber have been my closest confidantes in this extremely challenging journey, and for their friendship, kindness and advice, I am forever indebted.

Lastly, I would never have made it this far without the immeasurable support and love from my husband, Pranav, inspiration from my children Abhay and Sharanya, my parents, my in-laws and extended family for always believing in my abilities and being my rays of hope in dark times. Everything I was, am and ever will be, is simply because of your love and faith in my dreams.

Dedicated to the three women in my life I am forever indebted to for their
infinite wisdom

“Faith can move mountains”

-RR

“Quitting is never an option”

-JJ

“One day at a time”

-KB

TABLE OF CONTENTS

Chapter	Page
CHAPTER I: GENERAL INTRODUCTION	16
1.1 Overview of Teaching Assistants in US Universities	16
1.2 Graduate Teaching Assistants Training Programs	19
1.3 Theoretical Framework	21
1.4 Research Motivation	22
1.4.1 Preliminary Study Conducted at Iowa State University, Ames, IA	22
1.4.2 Preliminary Study Conducted at University of Oregon, Eugene, OR	25
1.5 Problem Statement	26
1.6 Summary	28
CHAPTER II: INCORPORATING A BACK-READING PROFESSIONAL DEVELOPMENT MODULE FOR CHEMISTRY GRADUATE TEACHING ASSISTANTS	29
2.1 Chapter Abstract	29
2.2 Introduction	30
2.2.1 Review of the Literature on Teaching Assistant Training Programs	31
2.2.2 Context of present research problem	35
2.2.3 Research setting	36
2.2.4 Research motivation	36
2.3 Theoretical basis of the study	41
2.3.1 Assumptions of the present study	44
2.4 Formal Methodology	46
2.4.1 Research Questions	46
2.4.2 The College Board AP Chemistry Exam Scoring Process	46
2.4.3 Summary of the General Chemistry Laboratory Course at the University of Oregon ..	49
2.4.4 Adapting the Back-Reading Process for Training GTAs In Grading Laboratory Report.	51
2.5 Results	59
2.5.1 Grading Sample Laboratory Reports at Staff Meetings	59
2.5.2 Back-Reading Discussions at Staff Meetings	59
2.5.3 Individual Back-Reading for Grading Laboratory Reports	63
2.5.4 Post-Training Protocol Results	65
2.6 Analysis & Discussion	67

2.6.1 Grading Accuracy	67
2.6.2 Grading Consistency	68
2.6.3 Course Grade and Overall Learning Outcomes	69
2.6.4 GTA Feedback Comments	71
2.7 Conclusion.....	72
2.7.1 Implications for GTA Training:.....	72
2.7.2 Limitations and Challenges in Back-reading Protocol.....	73
CHAPTER III: QUALITATIVE ANALYSIS OF GRADED STUDENT LABORATORY	
REPORTS.....	77
3.1 Chapter Abstract	77
3.2 Introduction	77
3.2.1 Qualitative Research Context	79
3.2.2 Qualitative Analysis Subjects.....	81
3.2.3 Qualitative Data Sources	82
3.2.4 The Food Dyes Laboratory Experiment.....	84
3.2.5 Examples of Student Work for Inductive Analysis.....	84
3.3 GTA Molly’s High Quality Student Report.....	88
3.3.1 Introduction Section	88
3.3.2 Claims and Evidence Section	91
3.3.3 Reflection Section	93
3.3.4 Summary	95
3.4 GTA Klaus’ High Quality Student Report.....	97
3.4.1 Introduction section.....	97
3.4.2 Claims and Evidence	99
3.4.3 Reflection section	101
3.4.4 Summary	103
3.5 GTA Molly’s Low Quality Student Report	104
3.5.1 Introduction section.....	104
3.5.2 Claims and Evidence section	106
3.5.3 Reflection section	108
3.5.4 Summary	109
3.6 GTA Klaus’ Low Quality Student Report	110
3.6.1 Introduction section.....	110

3.6.2 Claims and Evidence section	113
3.6.3 Reflection section	115
3.6.4 Summary	117
3.7 Conceptual Analysis Rubric (CAR) And Statistical Data	118
3.7.1 Objectivity in Using the CAR Rubric	118
3.7.2 CAR Data Analysis and Graphs.....	120
3.8 Discussion.....	123
3.9 Conclusion.....	128
CHAPTER IV: GROWTH MODELS AS A UNIQUE APPROACH FOR EVALUATION OF GRADING TRAINING	130
4.1 Chapter Abstract	130
4.2 Introduction	130
4.2.1 Literature Review: Evaluation of Teaching Assistant Training Programs.....	130
4.2.2 Training Programs Focused on International TAs (ITAs)	134
4.2.3 Training Programs for TAs In Inquiry Teaching.....	135
4.2.4 Training Programs Focused on Student Outcomes	135
4.3 Research Context.....	137
4.3.1 The Issue with Existing Program Evaluation Techniques Measures, And Approaches	137
4.3.2 Justification of Interest in Growth Models for Assessing TA Training Programs.....	139
4.3.3 Literature Review on The Use of Growth Models in Academic or Intervention Studies	141
4.4 Growth Model Theory	143
4.4.1 Explanation of A Growth Model Using Individual GTA Grading Data	143
4.4.2 Examining Multiple Individuals in A Group for Overall Growth Trends.....	145
4.4.3 Missing Data	149
4.4.4 Other Useful Results from A Growth Model Output	150
4.4.5 Summary	152
4.4.6 Research Questions	153
4.5 Methods, Data and Results	153
4.5.1 Data Collection	153
4.5.2 Data Preparation.....	155
4.5.3 Example of GM Data for A Single GTA's Grading Pattern	156
4.5.4 Example of GM Data Over Two Years of Back-Reading for Groups of GTAs.....	161

4.5.5 Example of Two Years GM Data of Back-Reading for All GTAs.....	166
4.6 Discussion.....	171
CHAPTER V: CONCLUSION AND FUTURE WORK.....	173
5.1 Conclusions (Summary of Results from Previous Chapters)	175
5.1.1 Design and Development of Back-Reading for Training GTAs In Grading.....	175
5.1.2 Qualitative Analysis of Graded Student Laboratory Reports	177
5.1.3 Growth Models as A Unique Approach for Evaluation of Grading Training	178
5.2 Future Work Ideas	180
5.2.1 Furthering the TA Training Protocol: Expectancy Value Theory.....	180
5.2.2 A Unique Rewards program: GTA Digital Badges	182
5.2.3 Vision for The Growth Model Framework to Analyze GTA Training Programs	185
5.2.4 The learning curve for grading in online courses and virtual classrooms.....	187
5.2.5 Summary of Research Results	189
APPENDIX A: COMPONENTS OF LAB REPORT.....	190
APPENDIX B: GENERIC RADING RUBRIC	191
APPENDIX C: SAMPLE LABORATORY EXPERIMENT.....	192
APPENDIX D: STUDENT REPORT GUIDELINES	200
REFERENCES	202

LIST OF FIGURES

Figure	Page
Figure 1: Primary sources of financial support for U.S. doctorate recipients 2020	16
Figure 2: Undergraduate enrollment in U.S. 4-year postsecondary institutions 1970-2029	16
Figure 3: Percentage of international students in US from 2005-2021	17
Figure 4: Percentage of international students in the US by field of study	18
Figure 5: Scatter plot of TA scores for laboratory reports (ISU)	24
Figure 6: A scatterplot showing varying scores for samples graded during Week 1,2 and 3 at UO.....	27
Figure 7: The same sample chemistry laboratory report was given varied scores during GTA orientation program prior to the start of classes (Year 1 of this study)	37
Figure 8: The same sample chemistry laboratory report was given varied scores during GTA orientation program prior to the start of classes (Year 2 of this study)	37
Figure 9: Faculty responses to survey questions on GTAs' grading and teaching practices.....	38
Figure 10: Frequency graph of coded GTA responses to survey in Fall 2015.....	39
Figure 11: Andragogy as a learning systems model with a feedback loop	43
Figure 12: Proposed model of andragogy for back-reading approach in training GTAs	45
Figure 13: A timeline depicting the AP scoring rubric development and validation leading to implementation at the AP scoring and back-reading.....	47
Figure 14: Sample laboratory report on the density experiment.....	54
Figure 15: An individual back-reading meeting in progress.....	57
Figure 16: Scatter plot and box-whisker plots for scoring of density sample report by GTAs before and after back-reading	61
Figure 17:GTA MC's back-reading data during weeks 1-4.....	64
Figure 18:GTA TB's back-reading results during training weeks 1-4.....	65
Figure 19:GTA MC's and GTA TB's data for the Fall term of Year 2 back-reading study.	67
Figure 20:Comparison of overall course grades and laboratory practical exam performance for GTAs MC, TB with general chemistry laboratory course total student population	70
Figure 21:GTA Molly; Introduction section from high-quality report.....	89
Figure 22: GTA Molly; Claims-evidence section from high- quality report	91
Figure 23: GTA Molly; Reflection section from high-quality report	93
Figure 24: GTA Klaus; Introduction section from high quality report	98
Figure 25: GTA Klaus; Claims and evidence section from high-quality report.....	99
Figure 26 : GTA Klaus; reflection section from high-quality report	102
Figure 27: GTA Molly; Introduction section from low-quality report.....	105
Figure 28: GTA Molly; Claims and evidence section from low-quality report.....	107
Figure 29: GTA Molly; Reflection section from low-quality report.....	108
Figure 30: GTA Klaus; Introduction section from low-quality report	111

Figure 31: GTA Klaus; Claims and Evidence section form low-quality report.....	113
Figure 32: Graph output of calibration curve blue dye (GTA Klaus, low-quality report)	114
Figure 33: Graph output of calibration curve for blue dye (GTA Klaus, low-quality report)	114
Figure 34:GTA Klaus, Reflection section from low-quality report	116
Figure 35: Conceptual analysis rubric used for qualitative examination of thirty selected reports from GTA Molly and Klaus.....	119
Figure 36: Plot of normalized CAR scores versus BR scores for GTAs Molly(blue) and Klaus (orange).....	121
Figure 37: Box-whisker plots of CAR scores, BR scores and final exam scores for Molly's students (n =15)	122
Figure 38: Box-whisker plots of CAR scores, BR scores and final exam scores for Klaus' students (n =15)	123
Figure 39: Histogram showing different modes of data collection used for training program evaluations reported in literature.	138
Figure 40: Measurements recorded for an individual at various points in time	143
Figure 41: Graph of dependent variable vs. time, the trendline represents the growth trajectory for this individual	144
Figure 42: Statistical representation and terms in an unconditional, level- 1growth model	145
Figure 43: Statistical representation and terms in an unconditional, level-2 growth model	148
Figure 44: Model output for GTA Molly's longitudinal data, unconditional means model	157
Figure 45: Level-1 model output for GTA Molly's back-reading data.....	158
Figure 46: Level-1 model graph for GTA Molly's back-reading data.....	159
Figure 47: Growth model output for Four individuals using TXCOMPLI (participation) as predictor variable.....	163
Figure 48: Graph output for level-2 growth model using data from four GTAs, TXCOMPLI and Individual GTA code predictor variables	164
Figure 49: Graph output for level-2 growth model using data from four GTAs, TXCOMPLI predictor variable.....	165
Figure 50: Level-1 model output for 108 GTAs	167
Figure 51: Graph output for level-1 model for 108 GTAs	168
Figure 52: Level-2 model output for 108 GTAs using participation and gender as predictor variables.....	169
Figure 53: Graph output for level-2 model for 108 GTAs using participation and gender as predictor variables	171
Figure 54: Adapted data from Statista for doctoral degrees awarded between 1950-2020	173
Figure 55: Scatter plot showing the rise of publications on "Digital Badges" between 2004-2022.....	182
Figure 56: Visualization of an individual GTA as part of a nested population	186

LIST OF TABLES

Table	Page
Table 1: Summary statistics for of high, medium, and low-quality laboratory reports graded during weekly staff meetings	24
Table 2: Comparative Summary of Pedagogy and Andragogy	42
Table 3: Summary Statistics for sample grading and back-reading of the Density Exploration laboratory report “Density Exploration” was graded by GTAs and the process	55
Table 4: Point breakdown for scoring of the laboratory report seen in Figure 14	60
Table 5: Summary of discussion between instructor and TAs after grading training sample report.	62
Table 6: GTA comments for open-response question on the usefulness of back-reading	71
Table 7: List of laboratory experiments performed during winter term.....	83
Table 8: Reflection section prompts for student laboratory reports	87
Table 9: Summary of scores awarded for the introduction section in GTA Molly's high-quality report.	90
Table 10: Recorded values for different food dyes	92
Table 11: Summary of scores awarded by GTA and researcher for reflection section	95
Table 12: Total scores for sections in Molly's high-quality report	96
Table 13: GTA Klaus; Scores awarded to introduction section in high-quality report	99
Table 14: GTA Klaus; scores for Reflection section in high-quality report.....	103
Table 15: Total scores for GTA Klaus' high quality report	104
Table 16: Total scores for low-quality report; GTA Molly	110
Table 17: Scores for introduction section from GTA Klaus' low-quality report	112
Table 18: Total scores for GTA Klaus' low-quality report	117
Table 19: Time-coding for longitudinal data in two-year back-reading study	154
Table 20: Predictor Coding by Participation and Gender	155
Table 21: Data for GTA Molly; coded for weeks in the back-reading study	156
Table 22: Estimated parameters for GTA Molly, level-1 model	160
Table 23: Estimated parameters for FOUR GTAs (Molly, Mickey, Mimi, and Molly), level-2 model with participation as predictor.....	164
Table 24: Estimated parameters for all GTAs (n =108), level-1 model.....	166
Table 25: Estimated parameters for 108 GTAS level-2 model with participation and gender as predictors.....	170
Table 26: Proposed ideas for GTA Digital Badges in grading	184

LIST OF TEXTBOXES

Textbox	Page
Textbox 1: Excerpt from student 1 regarding GTA Klaus' grading of laboratory reports.....	125
Textbox 2: Excerpt from student 2 regarding GTA Klaus' grading of laboratory report.....	126
Textbox 3: Excerpt from student 3 about GTA Klaus' grading of laboratory reports	126
Textbox 4: Excerpt from student 4 regarding GTA Klaus' grading of laboratory reports	127

CHAPTER I: GENERAL INTRODUCTION

1.1 Overview of Teaching Assistants in US Universities

Teaching assistantships are the third largest source (highlighted in orange in Figure 1 below) of financial support for pursuing doctoral degrees at Universities in the United States ¹.

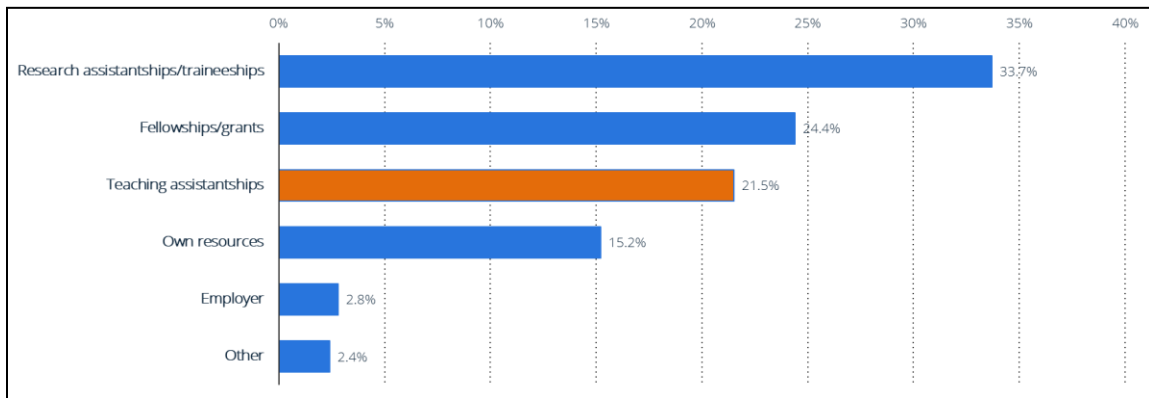


Figure 1: Primary sources of financial support for U.S. doctorate recipients 2020

Even though projected enrollment in US 4-year post-secondary institutions² has followed a strongly upward trend (Figure 2) up until 2019. However, graduate teaching assistants (GTAs) will continue to play an integral and ever-increasing role as undergraduate instruction personnel.

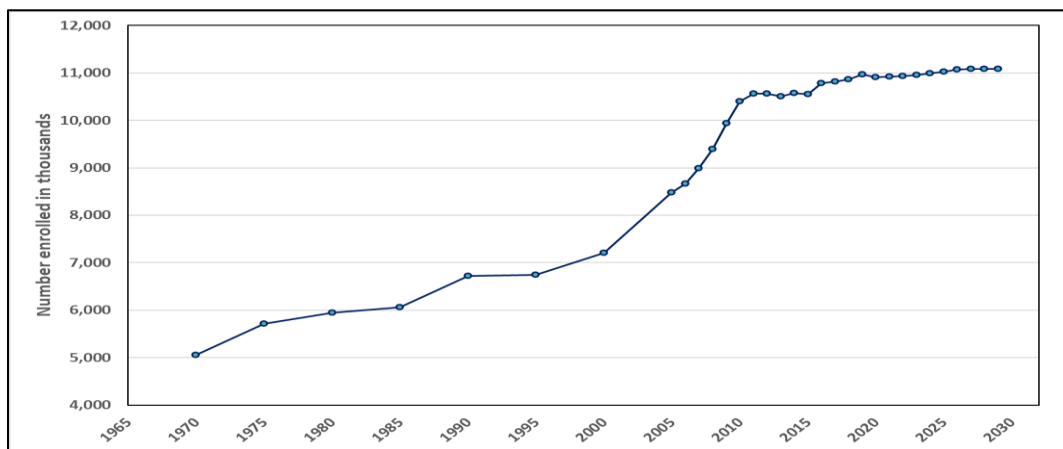


Figure 2: Undergraduate enrollment in U.S. 4-year postsecondary institutions 1970-2029

In the words of Gardner and Jones³, undergraduate teaching at research universities often rests solidly on the backs of graduate teaching assistants (GTAs) who teach large proportions of the introductory curriculum and have been described in a variety of ways from being the “bridge between faculty and students”⁴ to the “first line of defense for instruction”⁵.

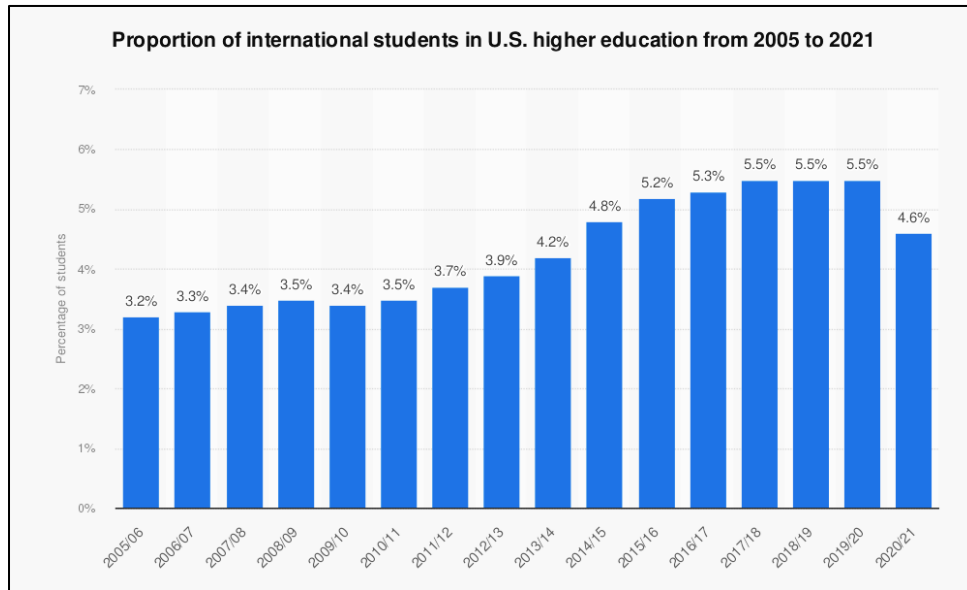


Figure 3: Percentage of international students in US from 2005-2021

Encouragement of program expansions, increased enrollment, and demand for higher education in the US has also ensured that the number of incoming international students (both undergraduate and graduate) has consistently been on the rise in the last decade (Figure 3). Due to the global COVID-19 pandemic between 2019-21, exceptions to this upward trend are noted. The growing rise of this demography is specially notable in STEM disciplines and management programs, which largely employ teaching assistants for large lecture and laboratory courses (Figure 4). Therefore, both the student population and the

teacher population indicate a rapid influx, and current predictions indicate that international student population at US universities will continue on an upward trend.

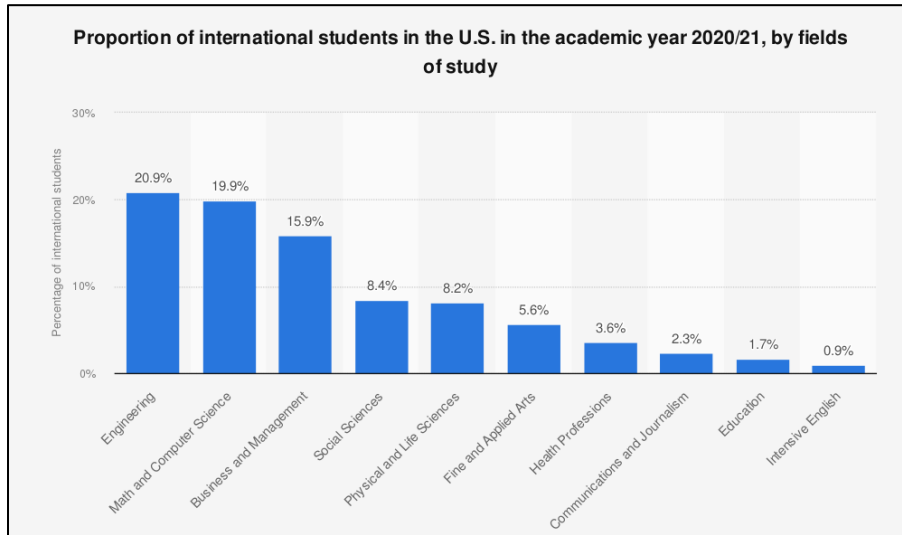


Figure 4: Percentage of international students in the US by field of study

Most universities have orientation week programs to provide a wealth of information about resources on campus as well as opportunities to experience the culture and lifestyle that international undergraduate students will experience during the education in the US. For international *graduate* students there is almost always some form of financial support, either as scholarships, research or teaching positions provided at the university to help support their education and living expenses in the US. With the graph projections seen in Fig 1-4, it makes sense to expect that with higher enrollments and incoming students' employment opportunities for graduate students as a research or teaching assistants would also have to increase proportionally. Because new graduate students have not completed graduate coursework and have yet to prove themselves as capable of productivity in a research laboratory, PhD students are often hired as teaching assistants during their first or second year of the program. As graduate students demonstrate progress towards their

doctoral degree (successfully completing courses, passing qualifying examinations and joining a research group) and becoming doctoral candidates (usually third year of PhD program or later) research professors who garner funding through research grant proposals, allocate research funding to enable their students to dedicate most of their time to their research.⁶ During a graduate student's first two years, there is competition between teaching and research responsibilities. Often during the second year, a graduate student being paid to serve as a teaching assistant for 20 hours per week, simultaneously works as an unpaid scientist in a research group. Graduate students must be successful in the research group and as well as in their graduate courses in order to continue their studies and remain in a doctoral program. However, they only need to be acceptable to continue to serve as a teaching assistant. This unwritten hierarchy of research-over-teaching is a critical point to bear in mind when examining the present literature in this field. Examining some of the challenges that GTAs face from this perspective, also highlight gaps in graduate students' current professional development which are recently being addressed in reported literature. This brings us to the nature and status of teaching assistant training programs in US universities.

1.2 Graduate Teaching Assistants Training Programs

Graduate students pursuing higher education are a valuable talent pool for their respective departments. Most GTA training programs involve discussions of teaching styles, technology and classroom culture, course-specific materials, and general responsibilities. There is always so much to cover during the first week before classes commence, that training programs invariably 'skim through' some of the key tasks and challenges GTAs face as stand-alone instructors. That is, the amount of preparation that is required for

teaching assistants is so intense that covering all of it in a week-long orientation is nearly impossible. Therefore, inclusion of as much information as possible in the GTAs reference materials, resources etc. is the next best option. There are several textbooks, and TA manuals with anywhere between 20-50 pages worth of information covering just the “first day of class” for GTAs.⁷⁻¹⁰ Many studies explore GTA training for orientation¹¹⁻¹³, lecture courses, stand-alone recitations, laboratory courses implementing traditional chemistry¹⁴⁻¹⁹, guided inquiry and open inquiry^{17, 20-23}, and teaching in the organic chemistry laboratory²⁴⁻²⁷. Studies exploring GTAs self-concepts, beliefs and other affective characteristics as teachers-in-training are also abundant in the literature.^{3, 28-35}. Studies for training programs focused on international GTAs are often limited to ensuring good communication skills and ability to overcome language and cultural barriers as stand-alone recitation or laboratory instructors.³⁶⁻⁵⁰

One of the ‘touch-and-go’ topics in many GTA training programs is assessment or grading of student responses. Teaching and grading are both equally important, since the former represents skilled delivery of new knowledge to students while the latter is an evaluation of how this knowledge was conveyed and if the students understood it well enough to be called ‘proficient’. However, the focus on the overwhelming task of simply facing students to teach, overcoming anxiety and communicating clearly, for both domestic and international GTAs largely eclipses the value of GTAs’ grading practices and assessment skills in their professional development or training. Although there are multiple reported articles evaluating the successful implementation training programs for GTAs^{40, 50-53}, most of them are geared towards preparatory skills such as the first day of classes, managing student behavior, communication and interacting with students, microteaching experiences

to alleviate anxiety and first-time teaching jitters. A very negligible portion, if at all, examines grading as a key piece of teaching assistant training. These are addressed in more detail as topic -specific literature reviews in each chapter that follows.

It cannot be claimed that grading practices of GTAs are not of any interest. There are a notable number of studies on what assessments should look like⁵⁴, types of rubrics and observation protocols, criteria of assessment, and what graders must know prior to grading and issues with grading. Several papers about faculty or teaching assistant beliefs about grading and grading practices^{22, 48, 55} serve to inform the research community on the nitty-gritties of assessment of students as novice learners. However, from a general perspective, training teachers and GTAs to grade student work accurately and consistently remains a largely unexplored area in chemical education. Perhaps it is because being able to assess students' chemistry content knowledge and laboratory skills, in and of itself does nothing by way of rewards or being a milestone for the GTA. Accurate and reliable grading is not one of the categories for which GTAs are evaluated. Grading, therefore, should be viewed as a heuristic process, from which GTAs learn and modify their next step, repeating this loop until the goal of reliable assessment is attained.

1.3 Theoretical Framework

Teacher training programs are often informed by theoretical perspectives such as constructivism, growth mindset, and sociocultural learning theory. We initially considered a developmental model for TAs as proposed by Nyquist,⁵⁶ Based on this framework, considering GTAs as adults learning complex skills and entering new organizations as beginning teachers. This developmental model suggests four stages of growth from novice-

colleague in training- junior colleague- and future faculty. This growth is accompanied by changes in approaching each aspect of their teaching career with growing competence based on the role models GTAs are provided with along with supporting tasks to reinforce the skills. However, we found significantly better alignment with Knowles model of andragogy⁵⁷⁻⁵⁹ considering adult learners' prior knowledge and ability to process new ideas via a feedback loop that can be summarized as input-action-evaluation.

An andragogy approach provides us a baseline to consider GTAs as self-regulating learners having reliable subject matter knowledge (chemistry) but lacking the ability to use this as a tool for assessment and evaluation. One component of constructing a professional development module and also the driving hypothesis of our study is to help GTAs see their own subject matter knowledge as a tool and use it skillfully to perform grading tasks as well as they would teaching chemistry concepts, techniques and phenomena to their students.

1.4 Research Motivation

1.4.1 Preliminary Study Conducted at Iowa State University, Ames, IA

We began our exploratory research on GTAs grading practices in the general chemistry courses at Iowa State University (ISU). Our initial premise was to add to existing literature about GTA training with an emphasis on GTA grading practices, since we identified it as a largely unexplored topic thus far. Forty-eight (48) TAs were assigned to teach over seventy (70) laboratory sections in the undergraduate general chemistry laboratory course. Each GTA attended a mandatory, hour-long staff meeting on Friday every week. The upcoming week's experiment, logistics, teaching strategies were discussed with course instructors and laboratory staff members at these meetings. Laboratory sections comprised

of 20 students on average and GTAs were assigned one or two laboratory sections depending on the discretion of the hiring coordinator and course instructor(s).

As part of our study, a grading element was introduced as a part of the weekly staff meeting at ISU. GTAs were provided with paper copies of three laboratory reports of varying (high, medium and low) quality, after redaction of any identifying information along with a grading rubric.

The criteria for pre-selection of these samples of varying quality were the *assigned scores* for the reports from the term they were drawn. A report with a score between 80-90% (32-36 points out of 40) was drawn as a “high quality” report. A report with a score of 70-79% (28-31 points out of 40) was drawn as a “medium quality” report. A report with a score of 60% or below (24 points or below out of 40 points) was drawn as “low quality” report. GTAs graded samples using the grading rubric and returned it with a numerical score and/or feedback comments. This exercise was repeated for at least five weeks in each semester. Our preliminary results are shown in Figure 5 and Table 1 below.

We hypothesized that GTAs would accurately use the rubric and award scores matching the actual quality of the student report. Also, that we would be able to observe a consistent pattern over the duration of the grading exercises at staff meetings i.e., low quality reports would consistently receive lower scores and copious feedback comments to the student. Since GTAs were provided three varying quality reports, we also expected to see a consistent differentiation of quality from the scores award by GTAs. The results in figure 5 and Table 1, showing large spreads of scores, high standard deviation, and little distinction between quality of reports tell of an exactly opposite picture. This preliminary data became our primary motivation to address GTAs’ grading practices.

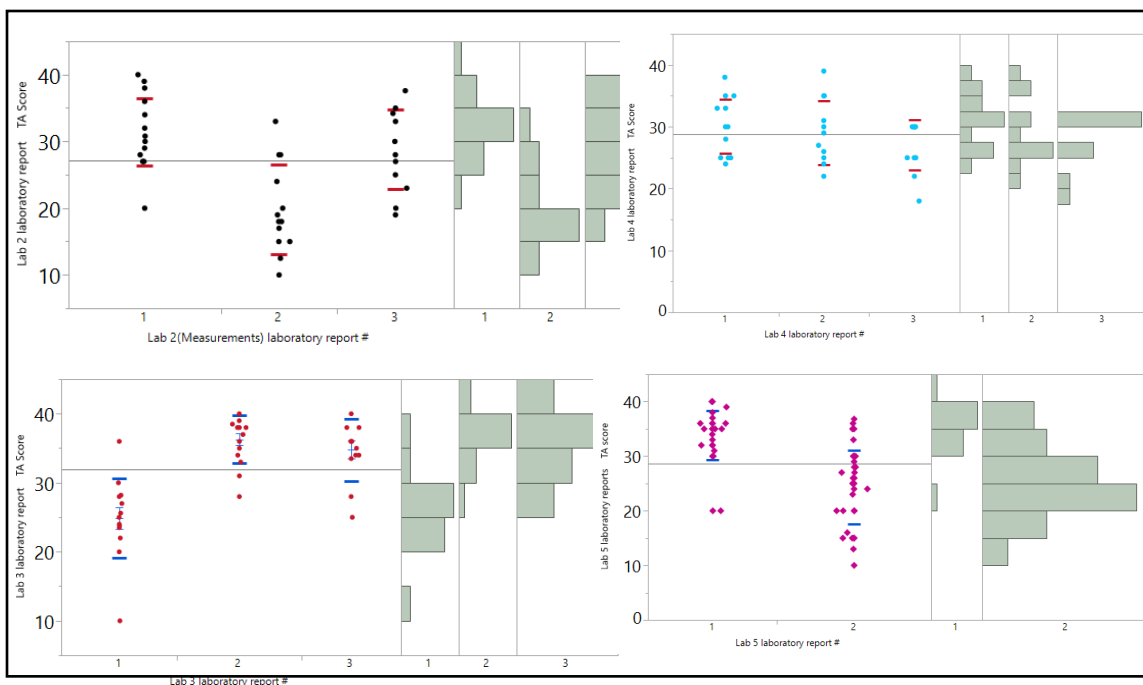


Figure 5: Scatter plot of TA scores for laboratory reports (ISU)

1.Experiment Title	Quality	Sample #	(n)	Mean	Standard Deviation	% Standard deviation
2.Measurements	High	2.1	17	31.46	5.03	12.575
	Medium	2.2	13	19.81	6.69	16.725
	Low	2.3	13	28.23	5.98	14.95
3.Empirical Formula of unknown (CuO)	High	3.1	14	24.86	5.77	14.425
	Medium	3.2	15	36.30	3.50	8.75
	Low	3.3	12	34.79	4.50	11.25
4.Preparation and properties of a hydrated salt/ double-salt	High	4.1	19	29.05	4.30	10.75
	Medium	4.2	15	28.26	4.92	12.3
	Low	4.3	13	26.84	3.97	9.925
5.Acid-Base Titrations	Medium	5.1	31	33.84	4.53	11.325
	Medium	5.2	37	24.29	6.79	16.975

Table 1: Summary statistics for of high, medium, and low-quality laboratory reports graded during weekly staff meetings

1.4.2 Preliminary Study Conducted at University of Oregon, Eugene, OR

We continued our exploration of GTA grading practices at University of Oregon (UO) from 2015 onward, shortly after our preliminary findings at ISU. Graduate Teaching Fellows, referred to as GTFs (GTAs in this dissertation for simplicity) for the undergraduate general chemistry laboratory course, at UO were invited to participate in a study exploring their grading practices. A total of 24 GTFs were employed during this preliminary exploration of GTAs grading practices, serving as instructors for 44 sections of 18 students each of the course. All GTAs attended a weekly, two-hour staff meeting on Mondays where the upcoming week's experiment goals, technical logistics, teaching strategies and specific material (such as the pre-lab slides used for instruction) were discussed.

In the last half hour of this meeting, GTAs were provided with two or more sample laboratory reports of varying quality with redacted identifying information about the student authors along with a grading rubric for the experiment that was just discussed. The GTAs were requested to either grade the entire report (or depending on time availability, only a section of the report) individually. The course instructor(s) also simultaneously graded the sample and provided prompts for generating discussions among GTAs.

After a show-of-hands to assess score agreements between all graders, GTA questions about grading or grading rubric criteria were answered. Instructors also provided input on the appropriate use and expected interpretation of criteria specified in the rubric when grading student responses. The graded sample laboratory reports were collected and recorded as preliminary data from UO (Figure 6) with similar hypotheses proposed at ISU: GTAs would grade would accurately use the rubric and award scores matching the actual

quality of the student report; a consistent pattern in grading would be observable i.e., low quality reports would consistently receive lower scores and copious feedback comments to the student and if grading of varying quality reports was performed, a consistent differentiation of quality would be evident from the scores awarded by GTAs. This exercise was repeated each week.

Our findings indicate the GTAs at UO were not able to use the designated rubric to grade the same report accurately and reliably. The same lab report received a range of scores from a grade of “A” to a grade of “C”. This is an issue with far-reaching implications, because a student’s lab report score depends on the GTF who is grading the report not on chemistry content or the ability to communicate chemistry, and further reinforced our motivation to continue researching GTAs grading practices and develop a training or professional development module to address grading discrepancies in assessment of chemistry laboratory reports.

1.5 Problem Statement

Dedicating a portion of the time in staff meetings to grading sample student work and discussing the scoring creates a low-stakes, cohesive environment for GTAs to engage with fair and accurate assessment practices as learners. The weekly staff meetings were a unique opportunity to expose GTAs’ to assessment goals and best practices in addition to the teaching logistics by simple provision of sample student reports for grading exercises. Based on data presented in Figure 5 and 6, and our discussions with GTAs at staff meetings, we inferred that without adequate training in grading, GTAs cannot be expected to grade student laboratory reports accurately and consistently. Requiring accurate and consistent assessment from all TAs is as critical as providing instruction in the laboratory. This is

achievable by focusing on training GTAs in grading, providing guidance for consistency in grading and explicitly evaluating whether or not GTAs can grade reliably.

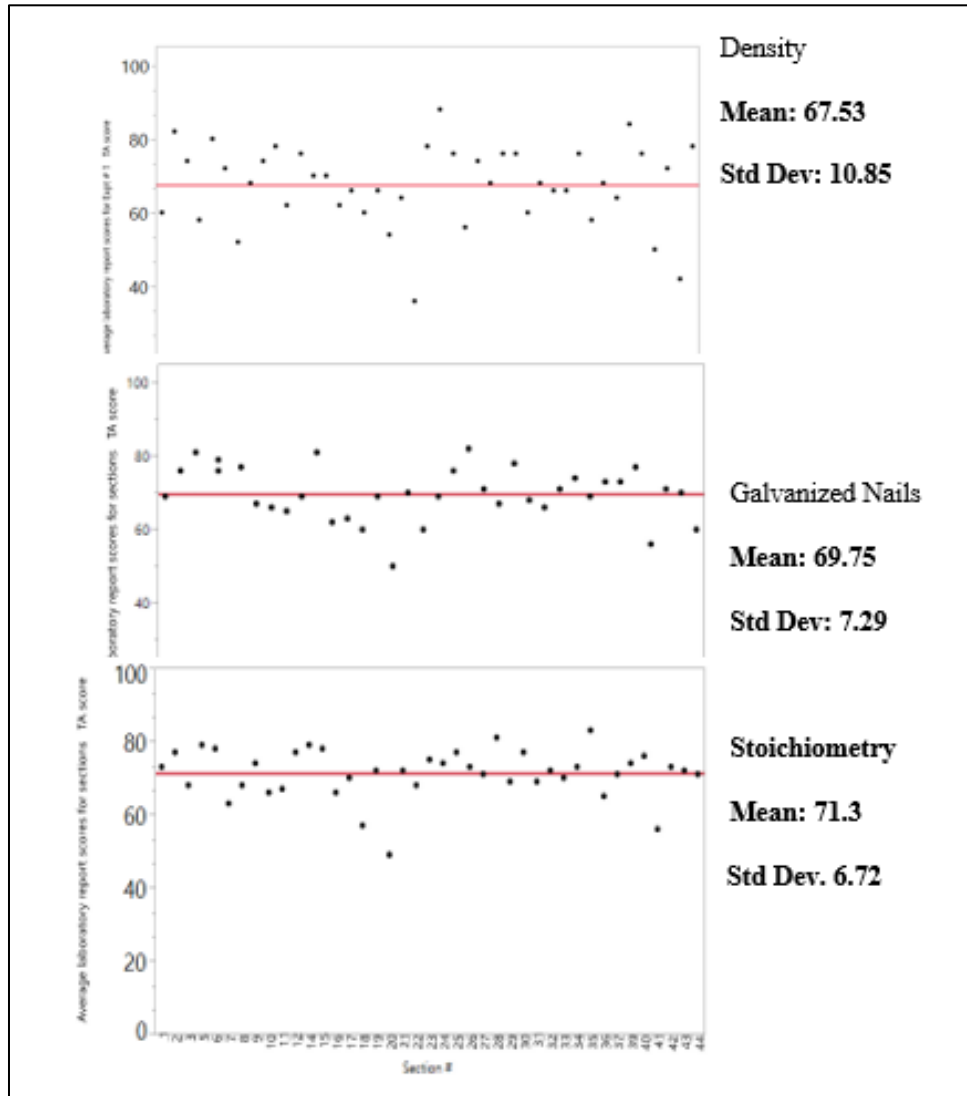


Figure 6: A scatterplot showing varying scores for samples graded during Week 1,2 and 3 at UO

With this objective, we framed our research question as:

1. How do we design a grading-specific training program for TAs to ensure laboratory reports and laboratory practical exams are graded accurately and reliably?
2. What factors influence GTAs' assimilation and continued utilization of training in grading?

3. How can we ensure the grading process serves intended course learning outcomes?

reports.

1.6 Summary

We have provided a general overview of the nature of existing literature and our motivation to pursue a professional development protocol for GTAs in Chemistry. This protocol emphasizes reliable grading practices in assessment of general chemistry laboratory reports. The design and development of such a specific protocol to train GTAs in grading chemistry laboratory reports is discussed in Chapter two. Chapter three explores a qualitative examination of two high-quality and two low quality reports from individuals who represent the GTAs in at least two specific parameters, participation and gender. This chapter highlights the impacts and shortcomings of our training protocol. Chapter four reports a first known implementation of a growth model approach for evaluating a GTA training program. This is a significant departure from conventional training program evaluation methods such as surveys and semi-structured interviews with participants. The concluding chapter provides information and ideas for further extension of this work and potential challenges that will need to be addressed in the future. As Ernest Boyer⁵⁶ rightly puts it, “to be careless about assessment, is to be careless about human life. [...] ...We are talking about the future of human beings and [TAs] need to understand that their assessment of students should be fair, ethical, and consistent.”

CHAPTER II: INCORPORATING A BACK-READING PROFESSIONAL DEVELOPMENT MODULE FOR CHEMISTRY GRADUATE TEACHING ASSISTANTS

2.1 Chapter Abstract

Graduate teaching assistants (GTAs) are an integral part of the instructional team for the general chemistry lecture and laboratory sequence at major colleges and universities. Before an academic term begins, chemistry GTAs at most universities receive professional development from departmental faculty. The duration of such orientation programs is usually 2- 4 days and involves GTAs' familiarization with regulations, duties, and relevant course materials. While this report focuses on the characteristics of how faculty in one Department of Chemistry conducts weekly meetings with GTAs, weekly meetings conducted by faculty in other Departments of Chemistry have similar components. The GTAs in this study were working at a Pacific Northwest University. The GTAs attended weekly staff meetings with course instructors to prepare for the upcoming experiment. We report our findings using a modified format of weekly staff meetings with chemistry GTAs. The focus of these modifications was GTAs' grading of student chemistry laboratory reports. We determined there were significant errors in GTAs' ability to grade a laboratory report accurately and consistently in our preliminary exploration. Our modified approach to weekly staff meetings is adapted from existing protocols implemented in College Board Advance Placement (AP) Readings. Results of our study provide valuable information about GTAs' requirements for training in grading laboratory reports and the positive impact of such on-going professional development on GTAs' grading accuracy and consistency.

2.2 Introduction

General chemistry laboratory courses are often referred to as a “gateway” course for undergraduate majors such as pre-med or similar discipline-specific tracks. Chemistry course content and laboratory techniques play a significant role in students’ developing skills to succeed in future academic and career paths. General chemistry laboratory courses are also a ‘bridge’ for students transitioning from high school settings to a full-fledged college setting. Since high school chemistry, depending on resources, curriculum, and teaching approach, may not necessarily cover laboratory experiments to tie in with chemistry concepts taught in the classroom. There is sufficient evidence in literature that performing laboratory experiments⁶⁰ or actively participating in demonstrations of chemistry phenomena helps students visualize and better assimilate relationships between macro, micro and symbolic representation levels⁶¹. This in turn, enhances students’ conceptual understanding and academic success. The curriculum for general chemistry courses often begins with basic ideas such as symbols, units, conversions and builds up gradually to relationships between matter, forms of matter, energy, and other chemistry phenomena. General chemistry courses are also designed to help students examine real-life scenarios and understand the chemistry in everyday events around them. This endeavor certainly merits the use of laboratory experiments to ensure hands-on learning or, learning by doing.

Graduate and undergraduate teaching assistants (GTAs) play an integral part by serving as assistant instructors in chemistry courses. GTAs perform two important components of instruction: they lead stand-alone recitation sections and /or laboratory sections, and they grade assignments to provide feedback to students. The format of general chemistry

laboratory courses has also undergone significant evolution from being traditional cookie-cutter laboratory experiments to more elaborate discovery or inquiry type experiments that are the highlight of current times.

2.2.1 Review of the Literature on Teaching Assistant Training Programs

The ever-increasing responsibilities and challenges faced by TAs have been well-documented in many recent studies³. Undergraduate teaching at research universities involves TAs who have been described in a variety of ways from being the ‘first line of defense for instruction’⁵ to the ‘bridge between faculty and students’⁴.

In a typical large enrollment university, approximately 30% of TAs are international graduate students and often referred to as international TAs (ITAs). The primary focus of several reported studies have been ITAs’ ability to (a) communicate with native speakers of the English language^{37, 41, 62, 63}, and (b) conduct instructional activities effectively to foster undergraduate student learning^{47, 64, 65}, along with exploration of issues faced by ITAs such as managing classroom diversity^{47, 62}, teacher self-identity^{33, 35, 45, 62}, and cultural challenges faced by ITAs^{13, 30, 40, 44, 56}.

Another popular research focus with respect to training TAs is the development of teaching and presentation skills in a classroom or laboratory setting. These range from training TAs to (a) become stand-alone instructors of a lecture classroom⁹, (b) prepare and present discipline-specific content for students, (c) design assignments and assessment activities^{8, 10, 56} and (d) incorporate more recent methods of teaching such as active learning using flipped classrooms or just-in time-teaching⁶⁶ and process-oriented guided-inquiry learning (POGIL)⁶⁷.

Several studies also report on the design and implementation of a general TA orientation program^{68, 69}, orientations for inquiry-type teaching^{69, 70}, a graduate course for TAs^{4, 22, 71}, and theoretical models for developing teaching assistants' instructional skills and pedagogical knowledge⁷²⁻⁷⁴. However, reviewing the available literature indicates that TAs are mostly assigned laboratory teaching and related responsibilities but provided with little or no minimal training depending on the institution/program^{5, 29, 70, 72, 75}. Wheeler and colleagues also point out the scarcity of research studies on the training for inquiry-based teaching in the chemistry laboratory²². These studies often focus on highlighting issues faced by GTAs during training or strategies for resolving these while designing such a program for GTAs.

Some studies have examined the GTAs' own understanding of the discipline-specific content or inquiry teaching^{32, 74, 76}, self-perceptions towards their role as teachers^{32, 77}, and benefits of gaining teaching experiences⁷⁸ while GTAs are in graduate school. Such studies tend to reiterate approaches like school-level teacher-training, shifting the focus toward GTAs' identity and characteristics as a teacher-in-training and away from training that is necessary skill development such as pedagogical content knowledge mastery, teaching and academic assessments for college level students.

Despite the existence of many TA training programs and several research studies informing the community about the issues and possible strategies to resolve challenges, there is a persistent concern that training provided to TAs continues to be insufficient in addressing the goal of effective and holistic professional development of TAs⁷⁹. One silver lining that has successfully addressed some of these concerns is the Preparing Future Faculty (PFF) program instituted at various north American universities. Such programs continue

exploring effective contemporary strategies to develop future citizens of the academic community^{8, 56}. However, there is a gap in the on-going research on TAs' abilities as trained and reliable assessors of students' writing. The primary objective of this chapter is an attempt at highlighting this issue and proposing a likely solution to address it.

One of the critical skills undergraduate students are expected to master through four years of college is communicating scientific knowledge in different written formats (reports, memos, posters, blogs etc.) to a wide variety of readers. Writing assignments are designed to provide students an opportunity to think through ideas and compile them in a logical, presentable fashion. With the recent paradigm of "writing to learn"⁸⁰ gaining momentum, assignments exploring different written formats are now integral to the process of learning for students. Naturally, students also need to be provided with feedback and instruction regarding such written communications^{81, 82}.

There is little doubt that the majority of students enrolled in a college level general chemistry course need help in improving their writing skills. Wackerly⁸³ reported that students' ability to write is a complex thing to understand and it requires a gradual approach to improve students' writing skills. Submitting and receiving feedback on laboratory reports is one method wherein students can gradually improve their writing skills⁸⁴. Therefore, one of the major responsibilities that TAs in a general chemistry laboratory course are expected to fulfill is *grading laboratory reports*. This aspect is addressed by some training programs, albeit briefly. A study on generating desirable attributes of TAs by Cho *et.al.* has its basis on a needs assessment response from faculty⁸⁵. The TAs and students in this study indicated 'grading student work in a fair and consistent way' to be of importance. Teaching assistants and faculty rated this attribute on a 5-point Likert scale as

4.50 and 3.83 (consistently higher) under ‘instructional practices’ compared to other categories rated 3 (high) such as ‘preparedness’, ‘engagement with students’ and ‘classroom management’. Nyquist and Wulff⁹ address the important aspects pertaining to grading very briefly in their book chapter on preparing graduate teaching assistants (GTAs) for specific instructional roles: (i) reinforce the link between goals and grading criteria, (ii) inform GTAs about all relevant grading procedures and policies, and (iii) provide opportunities to practice consistency in grading. These points are stated briefly and mostly targeted toward the *faculty* who develop and coordinate training programs. Sadly, there is no illustration of how to conduct activities that address these goals. Schoem and colleagues⁵⁶ discuss the GTA training program at the University of Michigan designed as a one-credit course over a period of six weeks. Of these, week six addresses testing and grading. Wheeler and colleagues²² present the design and findings for their inquiry-based general chemistry course curriculum as interview and survey data. Although grading sample science laboratory reports, plans and summaries are integral and discussed in the methodology as well as results of this study, there is no explicit mention of using these findings on grading to examine the actual, real-time practices of the GTAs. A study on teaching-focused graduate student professional development by Withers *et.al.* discusses a course for chemistry graduate students. The course materials are designed extensively for graduate students to learn and practice skills as ‘future faculty’ wherein grading is addressed rather briefly. Thus, there is limited availability of literature that examines training GTAs in grading techniques, provides illustrations or examples of how such training is conducted, what challenges are involved, or explore GTAs’ actual practices as graders. These make the examination of how GTAs grade written student assignments, the

nature of feedback provided by GTA's, and overall evaluation of best practices for training TAs in grading to be worthwhile research questions.

2.2.2 Context of present research problem

Most universities in the USA have a TA orientation program to equip GTAs with the basic tools and skills for leading a class or laboratory as instructors^{18, 56, 65, 86}. Several studies summarize the importance of including training programs for teaching assistants⁸⁷. These programs are often focused on orienting teaching assistants toward the expected content knowledge and pedagogy^{19, 22}, providing tactics for logistical and class management scenarios^{13, 44, 52, 88, 89}, teaching and presentation skills⁹⁰ and even tracking the development of specific abilities of graduate students in their role as teachers^{65, 91}. However, there is little, or no information on the examination of how GTAs grade laboratory reports. A novice GTA often starts teaching a chemistry laboratory course with no grading experience. An inadequate preparation in the correct use of a grading scheme or rubric to score student laboratory reports often leads to grading that is not consistent with what is intended by the instructor.

The purpose of this study is to present evidence of grading discrepancies that arise from lack of training specific to assessment of students' written work. This is followed by our findings when implementing a grading-specific training protocol to address these issues. Lastly, we provide a discussion of our results and some recommendations for TA training programs.

2.2.3 Research setting

General chemistry laboratory courses at the University of Oregon (UO) regularly employ graduate teaching assistants and sometimes even undergraduate teaching assistants. Under the supervision of a faculty member (course instructor) each GTA serves as a facilitator or teacher for one or more sections of 18-20 students. GTAs' role involves interacting, monitoring, and guiding students through a series of chemistry experiments for a duration of 3 hours. During the orientation program at UO, GTAs' attention was directed to the course learning outcomes which include students developing an understanding of how to conduct experiments by formulating questions, following instructions, and recording data; the ability to think analytically from analyzing data; and the ability to write effective, properly formatted scientific reports. With the specific modifications made to weekly staff meetings, GTAs were provided guidance to examine student laboratory reports and in the use of the corresponding grading rubric.

2.2.4 Research motivation

GTAs at UO were provided with a *sample* laboratory report and a grading scheme (rubric) during the general chemistry course orientation sessions in the fall of year 1 and fall of year 2. Each GTA was requested to grade the same sample lab report independently, and then anonymously respond to a clicker question on the score they had provided. Results for the scores provided by GTAs in year 1 and year 2 are displayed in Figure 7 and 8. The diverse range of scores and repeated discrepancy patterns in both years for *the same laboratory report*, with *different* sets of GTAs involved in the course is remarkable and a cause for concern. Below are the range of scores assigned by GTAs to the same student's laboratory

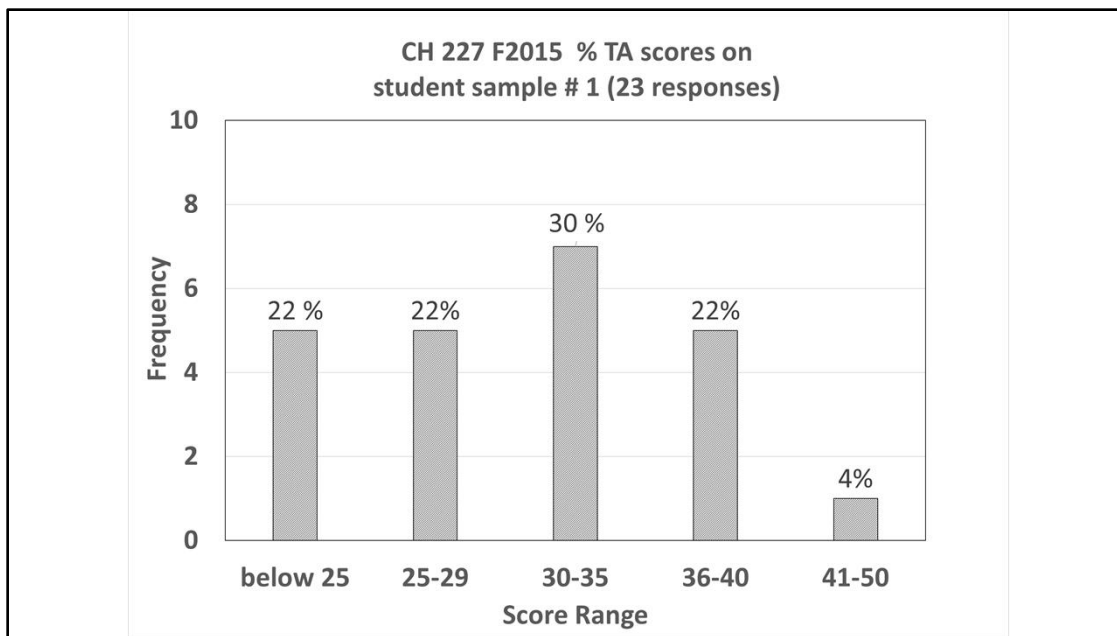


Figure 7: The same sample chemistry laboratory report was given varied scores during GTA orientation program prior to the start of classes (Year 1 of this study)

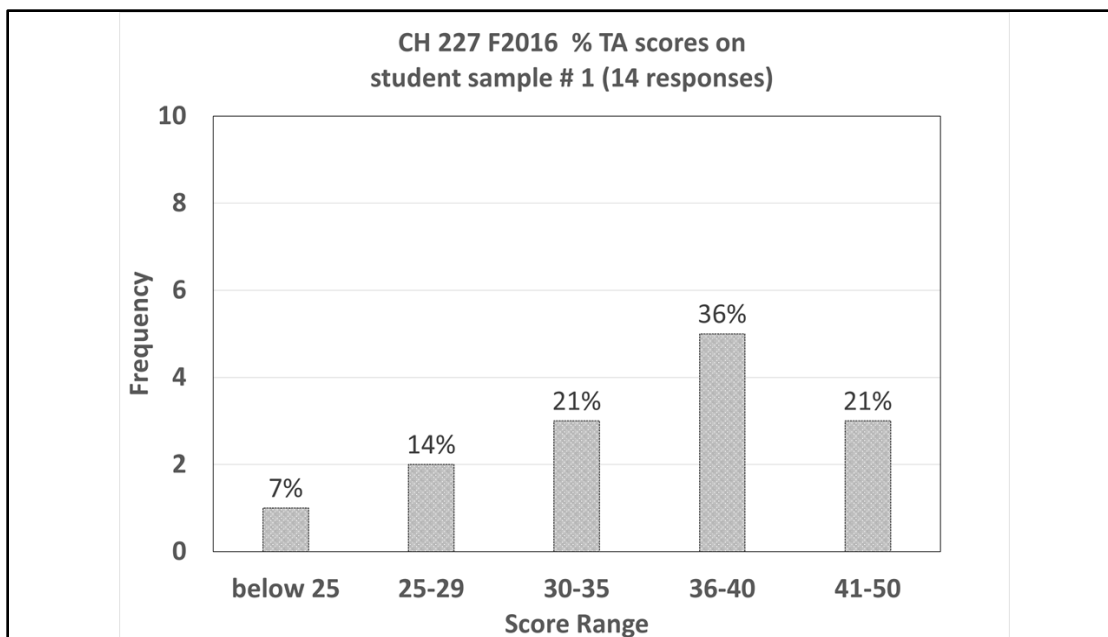


Figure 8: The same sample chemistry laboratory report was given varied scores during GTA orientation program prior to the start of classes (Year 2 of this study)

report during a GTA Training Program year 1 (Figure 7) and year 2 (Figure 8). Anecdotal evidence from faculty assigned to oversee teaching assistants in their general chemistry program indicated that GTAs' teaching and grading practices are not considered to be reliable or accurate (personal communication, 2014, 2015). Results of a survey of chemistry faculty at various institutions and information regarding the faculty's view of the reliability, consistency, and accuracy of their GTAs' grading skills are provided in Figure 9.

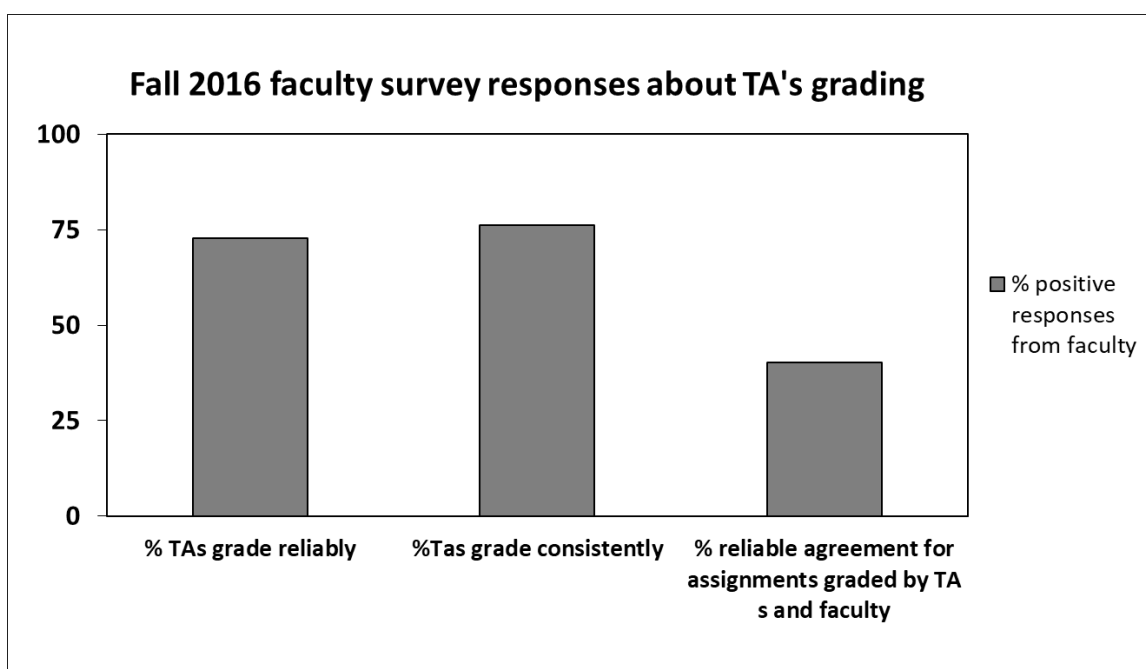


Figure 9: Faculty responses to survey questions on GTAs' grading and teaching practices

Although multiple respondents agreed that GTAs at their respective institutions graded reliably and consistently, the responses for the GTAs' laboratory report scores being in good agreement with faculty laboratory reports scores for the same chemistry laboratory report paint a contrasting picture. Due to the lack of published studies in addressing this

specific problem and from our preliminary findings, the purpose of our study was to address one major concern.

GTAs cannot be considered 'ready' to assess student work without receiving grading-specific training. We addressed this concern by developing and implementing a professional training module designed for chemistry GTAs to grade general chemistry laboratory reports accurately and reliably.

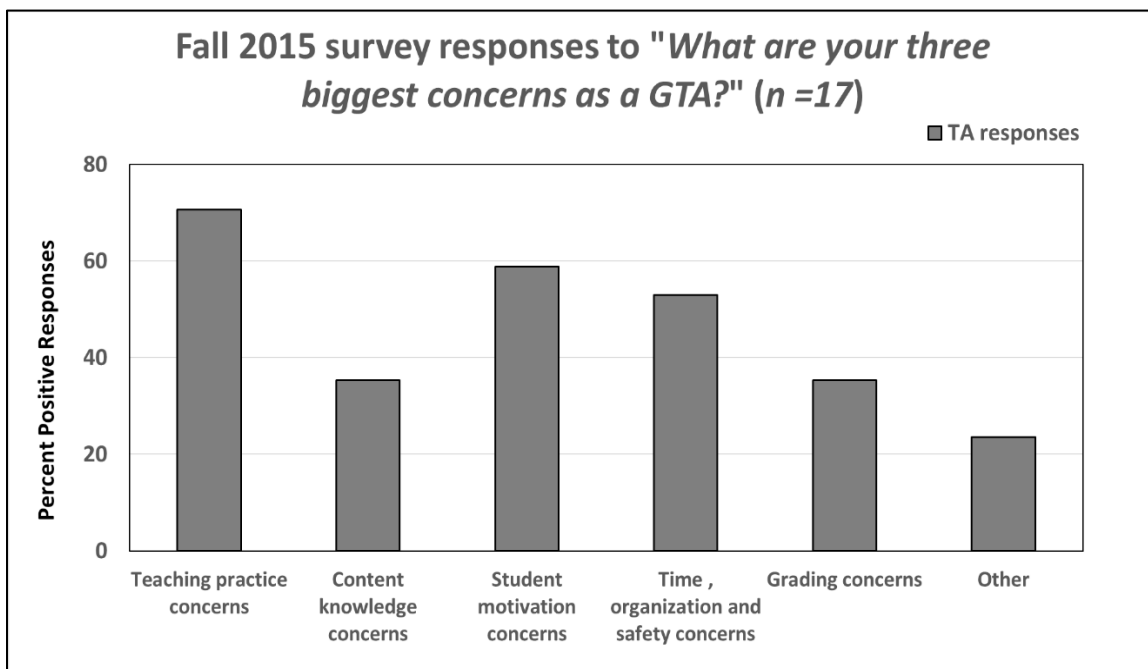


Figure 10: Frequency graph of coded GTA responses to survey in Fall 2015

GTAs for the general chemistry laboratory course at UO were mostly first-year graduate students from the department of chemistry & biochemistry. Depending on availability, two or three undergraduate chemistry majors were occasionally hired as UTAs. Most GTAs had little or no experience teaching a chemistry laboratory session or grading general chemistry laboratory reports. Novice GTAs struggle with accurately teaching using

guided-inquiry and communicating relevant chemistry concepts and laboratory techniques while performing their role as laboratory teaching assistants.

A survey of GTAs at UO was administered in the initial stages of this study and asked, “*What are your three biggest concerns as a general chemistry teaching assistant?*” Free responses from the GTAs were coded and grouped into themes, as shown in Figure 10.

Seventeen (17) UO GTA respondents completed the survey and reported a majority of teaching-related concerns such as “teaching a class by themselves” (70%) or “teaching wrong concepts by mistake” (35%) as well as “controlling the class environment” and “being liked and/or respected” (58%) by their students. Less than 40% of the GTA survey respondents reported “grading” or related tasks as a concern in comparison to other teaching-specific duties. This pattern may be attributed to (a) novice GTAs being unaware of the actual requirements as a grader until they grade student work and/ or (b) the anxiety associated with teaching/being a stand-alone leader for a laboratory group which eclipses any other GTA-role related concern(s).

Thus, even when GTAs are provided a grading rubric, detailed instructions and practice grading sample reports, the same laboratory report graded by different GTAs was awarded a “D,” “C,” “B,” or “A” letter grade depending on who is assessing the laboratory report. This disparity, therefore, was the basic source of concern. GTA’s beliefs and attitudes towards teaching and grading (as seen from anecdotal and survey responses above) were symptoms of a deeper need for training programs to be focused on encouraging best practices in grading such as (a) GTAs understanding the educational approach in teaching laboratory courses and (b) the importance of reliable assessment practices. Apart from having a direct impact on students through effective feedback and reliable assessment of

their chemistry understanding, significant grading discrepancies reflect poorly on the laboratory course instructors and have long-term implications for student attrition and discipline-specific academic success statistics.

2.3 Theoretical basis of the study

The term pedagogy can be traced to Greek origin and translated as “to guide children” (*paid* = children and *agogos* = leading). In the other words, pedagogy is the art and science of *teaching children* ⁹². However, with progressive pursuit of “how people learn” and “why people learn”, the children part of this definition was gradually lost and today, pedagogy is, by and large, the art and science of teaching ⁹³. In pedagogy, the assumptions that guide the theories are based on children as (1) novices or dependent learners ⁵⁸, (2) learners who encounter the knowledge presented to them for the very first time ⁹⁴, and (3) learners with minimal background knowledge or prior experiences that could affect the learning process. However, these same assumptions may not hold when teaching adults or adult learners.

The apropos term for the “art and science of teaching adults” known today is “andragogy.” This term was developed in the early 1970s by Malcolm Knowles ⁵⁸, who is often referred to as the “founder” of andragogy. However, it is also known that the origins of andragogy as an idea distinct from pedagogy are also credited to Alexander Kapp in the context of European adult education. ⁹⁵ According to both Kapp and Knowles, the education of adults requires focusing on what knowledge and skills are already present within the learner, not just the knowledge the learner is aspiring to gain. Andragogy is based on six principles: Learners (1) Need to know (2) Self-concept (3) prior experience (4) Readiness to learn (5) Orientation to learn and (6) Motivation to learn^{93, 94}. Surprisingly, an examination of literature shows us that andragogy as a framework or guiding theoretical perspective is

quite popular in adult education for a wide variety of occupations ranging from training athletes⁹⁶, swimmers⁹⁷, healthcare workers or nursing⁹⁸, teachers^{57, 99} and even training law enforcement ¹⁰⁰ and firefighting personnel ¹⁰¹.

The existence and persistence of andragogy-driven studies is evidence of a growing awareness of the differences between the education of children and that of adults. The simple acknowledgement that adult learners possess prior knowledge and experiences that guide motivation coupled current research on new learning strategies and self-regulation have now made a world of difference. These are now evident in recent theoretical frameworks such as transformative learning and experiential learning which are guided by the fundamental principles of andragogy.¹⁰².

There are four basic differences in the approaches to education of adults (andragogy) versus that of children (pedagogy). These are summarily presented in Table 2 below.

Table 2: Comparative Summary of Pedagogy and Andragogy

	Pedagogy	Andragogy
Self-concept	Dependent learner	Autonomous learner
Prior Experiences	Little or no contribution to learning process	Rich resource for learning process
Orientation to learning	Teacher-directed learning; teacher decides curriculum and methods	Learner-directed learning; based on current needs of learners, uses variety of methods
Timeframe for application knowledge gained	Learning as a preparatory step for the future	Learning as a mean to identifying and solving problems in the present

Teacher training is often explored using a pedagogical lens/framework such as constructivism or phenomenology. Several GTA training programs outline the basis of their design and implantation using various pedagogical frameworks such as the expectancy-value theory of motivation ²³, self-regulated learning ³¹, sociocultural theory and constructivism.¹⁰³.

Malcolm Knowles' ⁵⁸ framework describes a 7-step process in an andragogical approach, a learning systems model with a feedback loop (Figure 11). The feedback learning loop is a three-step cycle allowing for needs assessments and input of information followed by targeted activities to address these goals. The loop finally concludes with an end-of-activity assessment which evaluates the activity and/or the resolution of the issue(s).

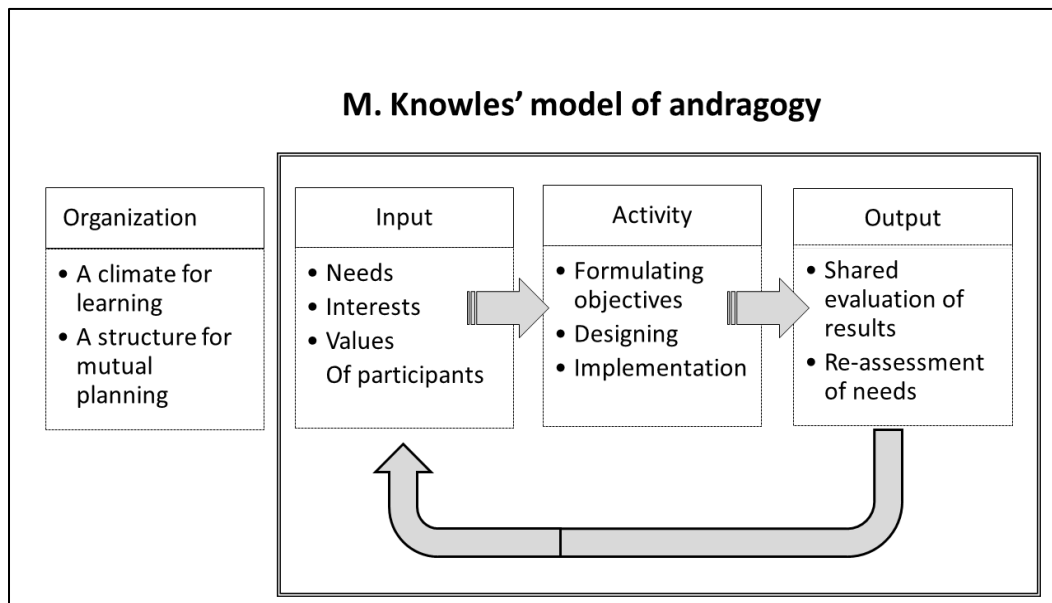


Figure 11: Andragogy as a learning systems model with a feedback loop

This model considers the prior knowledge and experiences of adult learners in the “input” stage, and this is the key difference between this learning model and any other pedagogical model focused on learning.

The present study involves the training of GTAs to reliably assess and provide feedback on students' written assignments. Examining the principles of andragogy helped us identify several characteristics of GTAs which match Knowles' assumptions about adult learners. Graduate students pursuing a master's or doctoral degree in Chemistry have been through the general chemistry course as undergraduate students themselves. This would support the assumption that GTAs chemistry content knowledge is rich. Additionally, their own lived experiences as students would imply firsthand knowledge of being assessed and receiving feedback on their written assignments. These lived experiences have tangible impacts on an individual's abilities as a GTA. For example, an individual who was evaluated too rigorously as an undergraduate student in a chemistry course, might tend to be lenient in their role as a GTA to over-compensate for their experiences, to avoid conflicts or simply being disliked. Alternatively, a student who experienced reliable and productive feedback from their instructors might be inclined to include similar approaches as a GTA because they find value in such skills. Therefore, training in grading for GTAs is very much a case of exploring adult learners and their abilities and led us to use andragogy as our research framework for this study.

2.3.1 Assumptions of the present study

GTAs are adult learners with

- (i) A developing sense of autonomy over their learning of any required skill set(s).
- (ii) An ability to identify or recognize areas where they face a problem as GTAs.
- (iii) An ability to orient themselves to a facilitated process of overcoming/solving such identified problems.

- (iv) Readiness to learn the skills that will solve the problems/ implement improvement(s) using various techniques such as collaborative tasks, discussions, time-bound or target-specific tasks.
- (v) Prior experiences as students themselves (receiving graded written assignments with feedback); these experiences translate to an information-rich resource for collaborative learning and bring rich perspectives towards grading written assignments.
- (vi) Demonstrated willingness and commitment to solve such problems by gaining required skills through interpersonal activities such as training and assimilating continuous feedback from the facilitators or peers.

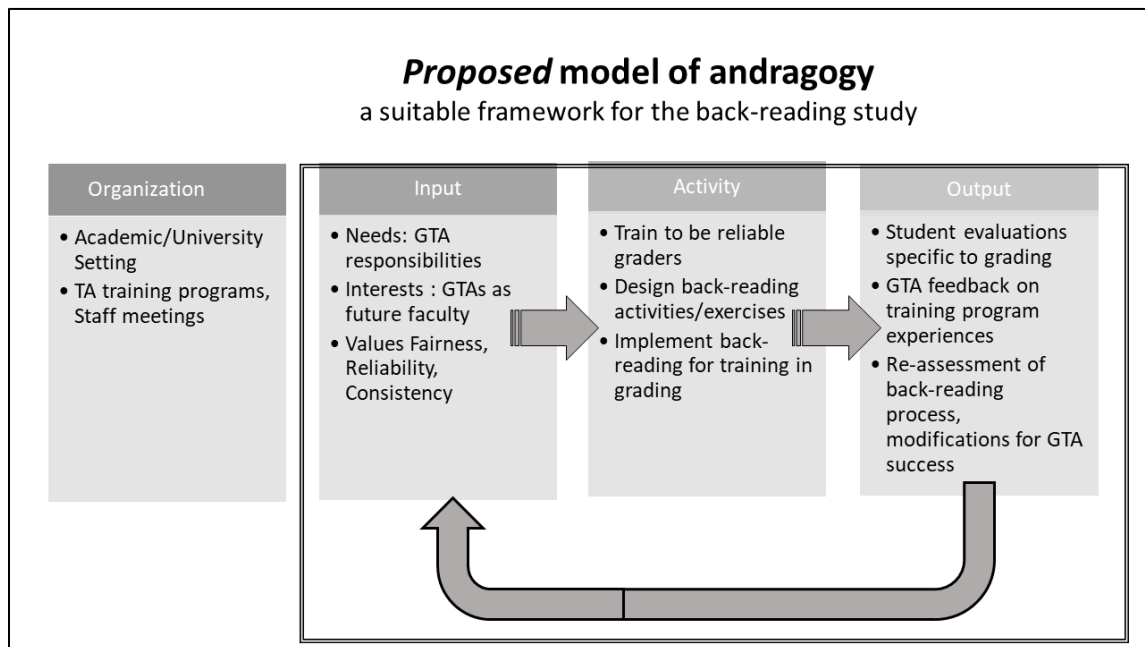


Figure 12: Proposed model of andragogy for back-reading approach in training GTAs

We reconstructed the model in figure 11 using our assumptions and goals of our study. As shown in figure 12, the feedback loop for GTAs agrees with Knowles' ideas and justifies an andragogical guiding framework.

2.4 Formal Methodology

2.4.1 Research Questions

- What components are necessary in a GTA training module to ensure that teaching assistants gain competency with grading laboratory reports accurately and reliably?
- To what extent can such a grading-specific professional development (PD) module positively influence GTAs' grading practices?

2.4.2 The College Board AP Chemistry Exam Scoring Process

In the context of andragogical framework, we implemented an adaptation of a “back-reading” (BR) approach from the College Board’s AP Chemistry Scoring Process ®. This was carried out with an intention of designing a continuous professional development module for GTAs to replace the conventional start-of-term training. “Orientation week” or “start-of-term” training may last anywhere between 3-7 days at the beginning of an academic term. The “back-reading” (BR) method is adapted specifically for GTAs grading student laboratory reports in a general chemistry course and holds promising results for the professional development of novice GTAs as accurate and reliable laboratory report graders.

In 2014, the College Board revised the format of the AP chemistry curriculum and AP chemistry examination to better align with the NRC 2002 report and recommendations. The “legacy” exams administered since 1956 underwent a significant review and were redesigned by the College Board to better align the widely used assessment with best practices in chemistry education. An overview of the redesign procedure and outcomes ¹⁰⁴ was used to develop Figure 13 above, which shows the hierarchical arrangement of AP

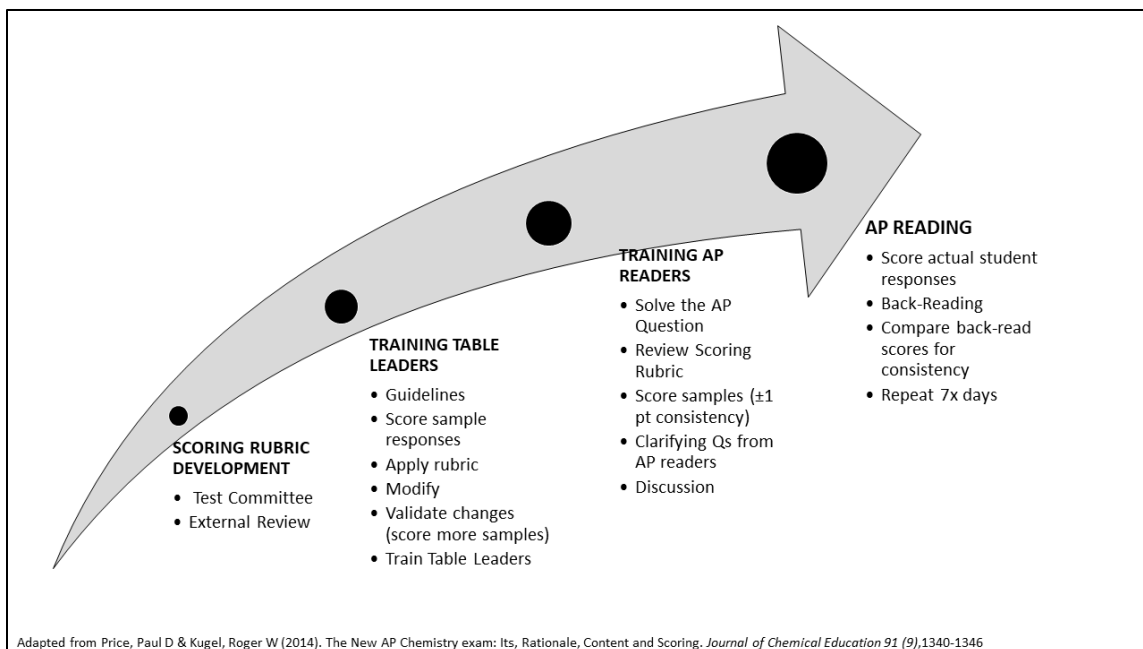


Figure 13: A timeline depicting the AP scoring rubric development and validation leading to implementation at the AP scoring and back-reading

readers during an AP Chemistry examination scoring session. The AP Chemistry examination has a free-response section consisting of three long answer questions and four short answer questions. Over 180,000 exam booklets are scored by AP Chem readers in a period of seven days. Sample student responses are assessed using a scoring rubric developed by the AP Chem Test Development Committee. Prior to the arrival of the readers, the Chief AP Chem Reader and AP Chem Question Leaders review and score sample student responses. Based on actual student work the scoring rubric is revised and details are added with specific examples prior to the start of the Reading. Before official scoring of student work begins, all AP Chemistry readers assigned to a specific question are provided with the rubric and receive training on how to score sample student papers. One component of this training is “back-reading,” (see excerpt below) a process developed for the College Board AP® examination scoring process by Educational Testing Services

(ETS) experts. The actual grading of students' responses begins when all AP Chem readers, who are assigned to a question, have demonstrated acceptable consistency and accuracy in applying the scoring rubric

“Readers begin reading fresh papers, and as they do so, their papers are back-read by the TLs. When the TL finds a discrepancy of more than ± 1 between their score and the reader’s score, they talk to the reader and discuss the discrepancy to make sure the reader is interpreting the rubric correctly. Finally, after a full day of reading, reader statistics are monitored daily to make sure their average score is consistent with the question’s average. If a reader’s score deviates significantly from the question average, they are back-read more frequently to correct, if necessary, their application of the rubric. By the end of this process, it will not matter who, in a group of 36 or so readers at a given table, gets a given student’s paper, it will end up with the same score within ± 1 point. The process developed by ETS statisticians and implemented at the reading ensures that the students’ scores on each of the free response questions will be consistent with the question’s rubric, regardless of which of the 300+ readers actually did the scoring.”

From personal communication with authors ¹⁰⁴.

Reading a large quantity of free responses, about five hundred per day at the AP reading, presents several challenges. One of the major concerns is scoring individual student responses consistently for quality and content by applying the rubric. AP Readers must be consistent in their grading over the course of seven days. It is common for some individual AP Readers to “drift” from the rubric and apply their own scoring standards. Because the work of all AP Readers is “back-read” by their Table Leaders each day, a Table Leader will detect when an AP Reader is not using the rubric appropriately. The

actual grading of students' responses begins when all AP Chemistry Readers, who are assigned to a question, have demonstrated acceptable consistency and accuracy in applying the rubric.

2.4.3 Summary of the General Chemistry Laboratory Course at the University of Oregon

The General Chemistry Laboratory Course at UO offered over three-quarter terms (or three 10-week terms) is concurrent with the affiliated General Chemistry lecture course. Each laboratory course lasts for 10 weeks and covers experiments that are often in tandem with the topics covered during the lecture course. A complete listing of the experiments offered is available in the course laboratory manual ^{105, 106}. Average enrollment in this general chemistry laboratory course is about 800 students. GTAs are employed as stand-alone instructors for the laboratory sections. Students attend a weekly 50-minute lab-lecture where the course instructor (a senior faculty member) provides information about the upcoming experiment for the week as well as an explanation of relevant theoretical concepts prior to performing the experiment in the laboratory. Students are expected to prepare a pre-laboratory experiment write-up in their laboratory notebook before attending a three-hour laboratory session supervised by a GTA. Each week's experiment includes 1) a pre-laboratory write-up, 2) in-lab work such as data collection, pooling/sharing class data, and 3) a post-laboratory report. A modified Science Writing Heuristic (SWH) format ^{15, 19, 82, 107, 108} for the post-laboratory report is implemented for this course, consisting of four major sections: (1) introduction, (2) relevant data tables, calculations, graphs, (3) discussion of results, and (4) conclusion section. Experiment-specific report writing

guidelines are also provided to students using Canvas, the university-wide on-line learning platform at UO. The post-laboratory report is due one week after the experiment was performed. It is graded by the GTA using a rubric provided by the instructor.

The Department of Chemistry and Biochemistry at UO provides a start-of-the term GTA training program during fall term to ensure GTAs are prepared for their roles as laboratory instructors. This is usually a four-day program which covers several aspects of teaching, GTA expectations and policies at this university and then specific course topics. The itinerary of the UO GTA training is like existing programs reported in the literature at numerous institutions. According to one of the instructors who leads this session every year,

“For the departmental training, we spend about 30 minutes talking about grading issues and strategies and have experienced GTAs give suggestions. We have done a variety of things over the years. GTAs have always been given instructions about what to look for in lab reports, how to use the grading keys and expected grade distributions. We have occasionally had all TAs grade and discuss sample reports at the beginning of the term. A lot of this has depended on how much time is available during training.” (Personal communication, 2014)

For the general chemistry laboratory course, GTAs also attend a weekly 90-minute staff meeting. During this staff meeting, the pedagogical goals for an upcoming experiment are discussed, any new equipment/techniques to be used are demonstrated, and time is also allocated for GTAs to perform a part of the experiment, to practice laboratory techniques, and to ask clarifying questions. From the time the present study was initiated at UO, the last 20-25 minutes of the weekly staff meeting were allocated for training in grading

samples of laboratory reports for the upcoming experiment and clarifying any questions or doubts regarding the grading scheme or grading rubric. The start-of-fall-term GTA orientation and weekly staff meetings offer excellent opportunities for course instructors to provide GTAs with professional development in experiment specific pedagogy as well as development of reliable grading skills. Our preliminary findings on GTAs grading of laboratory reports were presented earlier in the research motivation section of this paper.

2.4.4 Adapting the Back-Reading Process for Training GTAs In Grading

Laboratory Report

“Developing the ability to write an effective properly formatted scientific laboratory report” is listed as one of the course objectives¹ for the general chemistry course at UO where the present study was conducted. Laboratory reports are thus, a formative assessment of the student’s ability to communicate scientific information in writing. This mode of assessment requires graders to provide verbal feedback and/or written comments on graded student work. Modifications to the AP chemistry back-reading process were made in alignment with this objective as the framework.

The UO GTAs were provided with professional training for scoring laboratory reports using a grading rubric accurately and consistently. Additionally, they were also provided with guidance on (a) common student errors or misconceptions, (b) how to provide feedback to students and (c) time and effort management while grading. This training module was specifically designed keeping in mind that first year graduate students who are appointed as GTAs are themselves enrolled in graduate courses in addition to having

¹Source: UO General chemistry course syllabus

research commitments. At this stage, it would be worthwhile here to make note of the *differences* between the AP® scoring process and the laboratory report grading process presented in this study. The AP examination serves as a means of summative assessment, allowing students to demonstrate their proficiency in various chemistry concepts, and critical thinking skills. Therefore, the scoring of the students' responses over a period of six days is focused on the accuracy or degree of validity of the responses while applying the scoring rubric. The AP readers never write comments on the exam booklets or provide any form of feedback to the students. On the other hand, GTAs in a general chemistry laboratory course participate in formative assessment. They work with students on a weekly basis over a period of ten weeks, grading laboratory reports using a rubric and providing feedback to each student. This process of receiving written comments or feedback on graded laboratory reports provides students with valuable information on what they could do to improve their laboratory report writing skills.

Training In Grading Using Sample Laboratory Reports at Staff Meetings

Setting the expectation of accurate interpretation and consistent use of a grading scheme² or rubric often involves a clear demonstration of what this would look like in practice. As described earlier, about 20-25 minutes of the weekly staff meeting time are allocated to grading sample laboratory reports for the upcoming week's experiment. At UO, "Density Exploration" is usually the experiment of choice to commence with the general chemistry laboratory course. This experiment involves some basic chemistry topics (measurement, unit conversions); use of equipment such as an analytical balance, and glassware such as

²A grading scheme may be interchangeably referred to as a rubric in some sections of this paper.

volumetric pipets, beakers, and cylinders. Students were also provided instructions regarding the laboratory layout and safety protocols at this first laboratory session.

Many students enrolled in a college chemistry laboratory course for science or engineering majors have never used an analytical balance or a volumetric pipet prior to their first laboratory experiment in a general chemistry course. Fluency with fundamental mathematical and chemical concepts (such as calculating density, using significant figures, interpreting results based on chemical composition of samples analyzed) are major learning outcomes of the laboratory work for the “density exploration” experiment. For further details, readers are referred to the complete experiment described in the course laboratory manual¹⁰⁶. Students are provided with a simple outline of the expected format (*see Appendix A*) of the post-laboratory report for this experiment, which integrates the Science Writing Heuristic approach^{15, 82, 108} into student scientific writing development.

At the weekly staff meeting (before the “Density” experiment was performed by students), GTAs were provided with resources for leading the density laboratory experiment using a modified SWH guided-inquiry pedagogy^{19, 82} such as notes on managing laboratory logistics as well as the laboratory report outline and a grading scheme. A copy of these resources is available upon request.

Towards the last 20 minutes of the staff meeting, each GTA was provided with a *sample* laboratory report written by student “A” (student report content reproduced in the Figure 14). Each GTA was requested to grade this report using the grading scheme provided (*see Appendix B*) and then asked to respond to a clicker question about the score awarded to this report. After viewing the frequency distribution of clicker responses, the instructors asked the GTAs to discuss the implications of having a wide range in scores awarded for

the same report. This was followed by an explanation of the instructor's scoring, including interpretation and use of the grading scheme for the sample laboratory report. This brief

The results of this experiment led me to conclude that the type of sweeteners used in the two different cokes does alter the density of the drinks. The regular coke consistently showed the density being greater than 1g/mL whereas the diet coke was consistently less than 1g/mL. At the end of the experiment, we even looked to find that the diet coke floated in water and the regular coke sunk to the bottom. The differences in densities were likely a result of the types of sweeteners used because diet coke uses artificial sweeteners and regular coke uses regular sugars.

The possible sources of error I had in these experiments mostly came down to measuring volume. The pipetting was not always 100% precise and measuring the volume was difficult to see at times. I feel like it was minimal but could have easily been a source of error. Also, doing the mathematic calculations could have held possible sources of error because the math could have been done incorrectly thus yielding false results. These sources of error could have made my results different because the measurements would not have been as accurate, and the standard deviation would be greater.

Prior to the experiment I believed that diet coke and regular coke had the same densities but my ideas around that have since changed. New questions I have would be about different kinds of drinks and foods and how ingredients can alter the densities. I am curious to see what the density of a completely sugar free drink would be as opposed to its sugary counterpart. Also, I wonder how other ingredients would affect the densities as well. Finding the volumes, masses, and densities was relevant to what I've been studying because we have been learning about these concepts and how they connect in class. I knew the formula to finding the density due to the classwork. I knew that density could be found by the formula mass/volume and was able to use this in the class to determine the densities of the sodas and also the copper.

The experiment I conducted can relate to my chemistry 221 work because some of the same concepts were studied in both. I learned about densities in chemistry 221 and was able to apply these skills throughout the lab. I can connect the things I learned to my real life because I plan to go into a science career and I am sure I will need to know about densities in marine biology. Hopefully I will be able to apply it to find information about the oceans and other parts of my career.

One other experiment that I saw that was able to confirm my findings about the density of diet coke versus coke is the floating in water experiment. Water has a density of 1 g/cm³ so anything that is higher than that will sink in the water and anything lower will float. This proved that the density of the coke was higher than 1 g/cm³. Additionally, I have read in my textbook about densities and the information I found in my experiment was confirmed by what I read.

Figure 14: Sample laboratory report on the density experiment

discussion was followed by a request to GTAs to re-grade or revise their scoring and share a consensus score aloud or via clicker response. The post-discussion score was in overall agreement for most GTAs as seen in Table 3 below (see post-discussion SD values). A second sample laboratory report was used for repeating this exercise. Most of the GTAs graded the second sample report within ± 2 pts of the instructor's score.

Table 3: Summary Statistics for sample grading and back-reading of the Density Exploration laboratory report "Density Exploration" was graded by GTAs and the process

Scoring round	Mean	SD	(n)
Sample 1- before back-reading	35.4	5.6	19
Sample 1 – after back-reading discussion	30.9	1.5	19
Sample 2- before back-reading	25.4	8.2	19
Sample 2– after back-reading discussion	27.9	1.0	19

Back-Reading Personnel

At UO, a graduate student with at least one or more years of experience teaching the course was appointed as the Head GTA. The responsibilities of the Head GTA would be akin to a coordinator—assisting with the smooth flow of information and ensuring availability of resources to GTAs. The Head GTA was also considered a reliable grader based on their experience. That is to say, he/she would have a verifiable history of implementing each week's rubric accurately. Their consistency in grading would be evident from repeated, consistent scores recorded after time-gaps between grading the same report twice (or a greater number of times, if needed) for verification of consistency. After grading sample laboratory reports as part of training in grading at the staff meeting, an option to participate in a one-on-one *back-reading* session with the course Head GTA and/or instructors was

announced and highly recommended for GTAs as they began grading actual student reports.

This option to participate in a back-reading session was offered to GTAs with a two-fold intention: (1) the GTA could gain insight by grading actual student work on a one-on-one basis along with a peer (the Head TA) or the course instructor; (2) staff could monitor GTAs' understanding and implementation of the training in grading laboratory reports. A description of the individual back-reading can be found in the next section.

Individual Back-Reading for Grading Laboratory Reports

During the period of this study, back-reading of actual student reports was never mandatory for GTAs but highly recommended by the senior course instructors. Our back-reading study commenced from year 1 (fall, winter, spring terms) and through year 2 (fall, winter, spring terms). During the fall term of year 1, 14 out of 18 GTAs (77%) attended the back-reading session for grading laboratory reports on “Density” (week 1) and 15 out of 18 GTAs (83%) attended the back-reading sessions for grading “Galvanized Nails” laboratory reports (week 2). About 50% of the GTAs attended back-reading during “Metal Carbonates” grading for week 3, and attendance was sparse (below 50%) from week 4 onward. This was attributed to the general assumption that most GTAs felt they were sufficiently prepared to grade laboratory reports. Additionally, GTAs being fully occupied with graduate courses or research rotations was also a possible factor in the decreased attendance for individual back-reading meetings.

At the start of an individual back-reading session, each GTA along with the Head GTA or the course instructor independently graded 3 to 6 randomly selected copies of student laboratory reports from each individual GTAs laboratory sections. Once graded, the

scoring from the GTA and experts were compared, and any discrepancies were discussed. That is to say, the reports were “back-read” by the two graders. A consensus score was then awarded to the student laboratory reports. Figure 15 is a photograph of an individual back-reading session in progress and included for the reader’s visualization of the back-reading process.

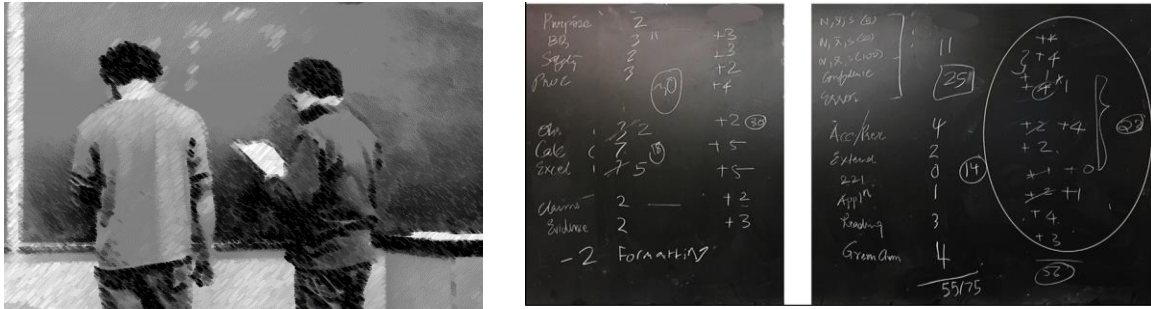


Figure 15: An individual back-reading meeting in progress

Back-reading on multiple student laboratory reports was repeated to check the GTA’s accurate implementation of the scoring rubric and implementation consistency. If the difference between the scores was within ± 3 points of the expert for three or more student laboratory reports at the back-reading session, the GTA was encouraged to proceed with grading independently. Further one-on-one grading was not pursued unless the GTA requested more back-reading or persistent grading discrepancies were recorded.

Post-Training Protocol: Checks on Grading Reliability and Consistency

As a follow-up to the one-on-one back-reading session, scanned copies of 2-3 student laboratory reports graded by each GTA were obtained for post-back-reading checks and monitoring uptake of training in grading. For post-back-reading checks, an expert or researcher graded unmarked copy of student reports and compared their scoring with that of the GTA. If discrepancies more than ± 3 points were recorded, the GTA was informed

of these differences and the discrepancies in scoring were addressed before returning the laboratory report to the student. The GTA was also advised to attend back-reading sessions for the upcoming assignment(s) to be graded. Re-grading laboratory reports was always an option to address discrepancies, but not recommended if the GTA had already returned the graded assignments to students during that week as it would be an added burden and an unnecessary expenditure of valuable time and effort.

During weeks 2 through 8 of the general chemistry courses, GTAs graded 2-3 sample laboratory reports during staff meetings. As described previously, GTAs responded to clicker polls and participated in discussions when discrepancies in scoring occurred. Discussions included how the sample student response and grading criteria was interpreted by the GTAs and instructors. The sample grading sessions provided us with a baseline to examine which GTA(s) might potentially have the need for one-on-one meetings because of the consistent discrepancies observed when grading sample reports. However, individual back-reading sessions were offered to all GTAs throughout the duration of the course.

Back-reading sessions for grading student laboratory reports was repeated every week. The majority of the GTAs attended only 2 or more during the first 3 consecutive weeks until their scoring results were within ± 3 points of the expert's score. Two GTAs repeatedly demonstrated outlier patterns in their grading at the back-reading sessions; one opted not to participate in the back-reading sessions after week 1 and the other GTA did not implement the grading scheme as intended. Nevertheless, three or more graded student reports were collected from every GTA for post-back-reading checks, thus maintaining consistent expectations for all GTAs. Results for individual back-reading sessions with two GTAs are discussed in the following section.

2.5 Results

2.5.1 Grading Sample Laboratory Reports at Staff Meetings

The “results and discussion” section of the sample laboratory report for the experiment on Density (reproduced in Figure 14) was used as a sample for training GTAs in grading. Paper copies of this section along with a grading rubric were provided (*see Appendix A*), and GTAs were requested to score the sample report as described in earlier sections of this paper. Paper copies of graded density sample laboratory report were collected and analyzed for scores provided by GTAs. Table 4 shows a snapshot of how five different GTAs graded this sample laboratory report. Instructor(s) scores are also provided for comparison. This grading task was followed by a discussion between GTAs and instructors, and a second round of scoring (back-reading). However, revised scores were collected only as show-of-hands or clicker responses. The second round of scores on the same samples are included as one summarized result in the table 4 ($mean = 19.8, SD = 1.8, n=5$). Based on these data, a stark difference between the grading by the expert and the GTAs was observed before back-reading and noticeable improvement in score agreement between all graders after back-reading (Figure 16).

2.5.2 Back-Reading Discussions at Staff Meetings

During back-reading discussions at staff meetings, GTAs were asked to share their observations and describe the nature of feedback that would best help the student (the original author of the sample laboratory report) improve their scientific writing skills. Most GTAs noticed the finer points of how the expert graded by the time they evaluated the second or third back-reading sample report. Table 5 summarizes the points of the discussion between the instructor and GTAs with reference to the density sample report.

Table 4: Point breakdown for scoring of the laboratory report seen in Figure 14

Rubric Criteria	Maximum Points possible	TA1	TA2	TA3	TA4	TA5	Instructor
general details	1	1	0	0	0	0	0
BQs	4	3	4	4	4	4	3
Claims	5	3	4	2	5	5	3
Evidence	5	2	3	4	2	5	2
Sources of error	5	0	3	4	4	4	2
Extending the experiment	5	3	2	5	5	5	3
Connecting lab to lecture	5	2	4	4	5	5	3
Real-life applications	5	3	3	4	5	5	2
Related reading/literature	5	2	3	2	5	1	3
TOTAL (Before BR)	40	19	26	29	35	34	21
<i>Revised scores (After BR)</i>	<i>40</i>	<i>20</i>	<i>17</i>	<i>20</i>	<i>20</i>	<i>22</i>	<i>21</i>

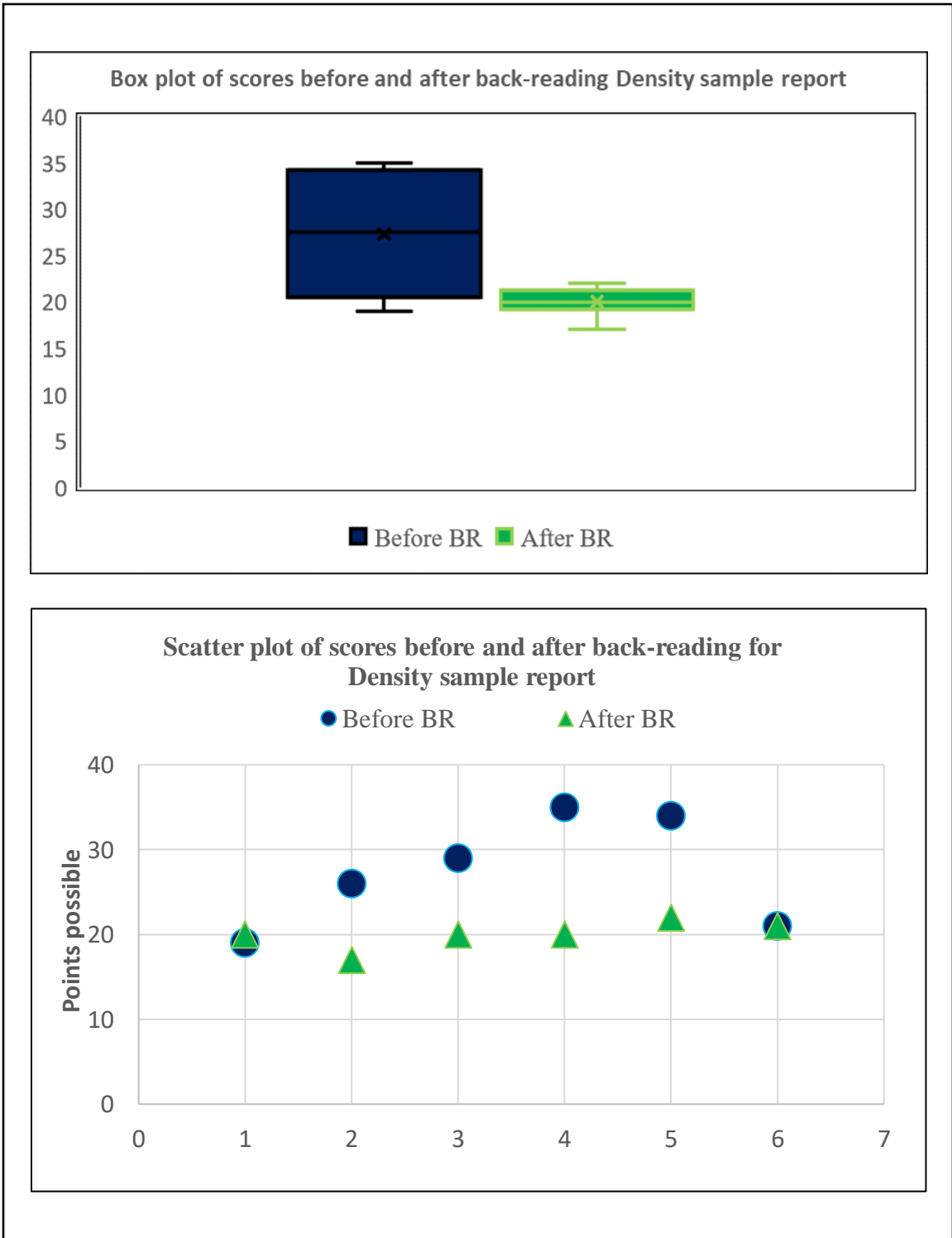


Figure 16: Scatter plot and box-whisker plots for scoring of density sample report by GTAs before and after back-reading

Table 5: Summary of discussion between instructor and TAs after grading training sample report.

IMPLICATIONS OF DISCREPANCIES IN GRADING
<p><i>“The same laboratory report graded using the same rubric must have significant agreement between multiple raters.”</i></p> <p>If the AP scoring expectations were to be applied to the UO GTAs, as specified by Price and Kugel (2014), only 6 out of 23 TAs (26%) were within ± 1 point of the expert’s score in year 1. However, 2 of the 18 TAs were within ± 2 points of the expert, which is slightly encouraging. There must be a clear agreement among raters, and this requires ‘practice grading.’</p>
<p><i>“The spread of scores indicates that each rater or GTA has a different interpretation or understanding of the grading scheme criteria specified.”</i></p> <p>It is not possible to provide a “perfect” rubric for each assignment, therefore the interpretation by various raters should be clear and fall within an acceptable range OR mutually agreed set of interpretations. The need for discussing the rubric and allowing for revisions or clarifications is essential during training in grading.</p>
<p><i>“Without training, GTAs examined the sample report either too critically or very leniently compared to the expert.”</i> The score provided by an expert (in this case, the course instructor) for the sample assignment was (62%) while GTAs scoring ran the gamut from being assessed as both a low-quality report (44%) and a high-quality report (88%) by GTAs. This highlighted the need for an explanation of how the expert went through grading.</p>
<p>“Grading discrepancies can have profound consequences.”</p> <p>Grading a sample report is a no-stakes scenario, where the score provided by the GTA does not impact an actual student’s grade. The results and discussion following sample laboratory report grading serve as a reminder that inaccurate and/or inconsistent use of the grading scheme can have serious consequences for GTAs as well as students. For example, GTAs may have to re-grade all the reports again or spend many hours trying to validate and address student complaints about fairness in grading.</p>

Table 5 (continued): Summary of discussion between instructor and TAs after grading training sample report.

“Turnaround time and quality of feedback provided to students is critical.”

The expectation of written feedback has often resulted in TAs spending copious amounts of time grading student work by writing comments or annotations for formatting, symbols, formulae etc. This information is generic enough to be provided as a class-wide email summary for everyone’s benefit. It is important to understand specific feedback that each student should be given prior to their submission of the next assignment versus general feedback. Accurate, consistent, and fair grading accompanied by constructive feedback is more likely to result in improvements of students’ laboratory reports as well as evidence of their understanding of chemistry content. The sample laboratory report graded during training sessions OR weekly staff meetings were geared towards achieving consistency and accurate interpretation of the grading criteria with minimal focus on feedback/comments provided by TAs in these sessions. There is a need for including knowledge of feedback types and methods of conveying generic versus specific feedback while minimizing time and effort investment on GTAs’ part.

2.5.3 Individual Back-Reading for Grading Laboratory Reports

Graduate students M Clifford (MC) and T Baldwin (TB) (*names changed*) were enrolled as first-year graduate students at UO and described their prior teaching experience as “almost none” during initial orientation. Both GTAs were also employed in the subsequent (Year 2) year of this study as general chemistry laboratory course GTAs. Both participated in all weekly staff meetings and attended back-reading sessions for the first four weeks to grade laboratory reports in a back-reading setting. Based on their grading approach and score discrepancies with the expert at the individual back-reading sessions, MC and TB were advised to attend further back-reading sessions during weeks 2-7. Both GTAs complied with this recommendation for weeks 2, 3 and 4 but did not attend any sessions

after week 5. For our results section, we present data from back-reading sessions during weeks 1-3. Additionally, post-check graded lab reports were collected during weeks 4-7 for analysis on uptake of training in grading.

During individual back-reading sessions, GTA Clifford and the expert (Head TA) graded a total of four laboratory reports independently, discussed scoring, and made revisions until satisfactory agreement was achieved. The difference in scores each week were significant as seen in Figure 17. During week 2 the difference in scores reflect GTA's tendency to grade more strictly, i.e., award much lower scores than expert. However, with progressive discussion and back-reading revisions, Clifford and the expert were able to come to an agreement on majority of the scores.

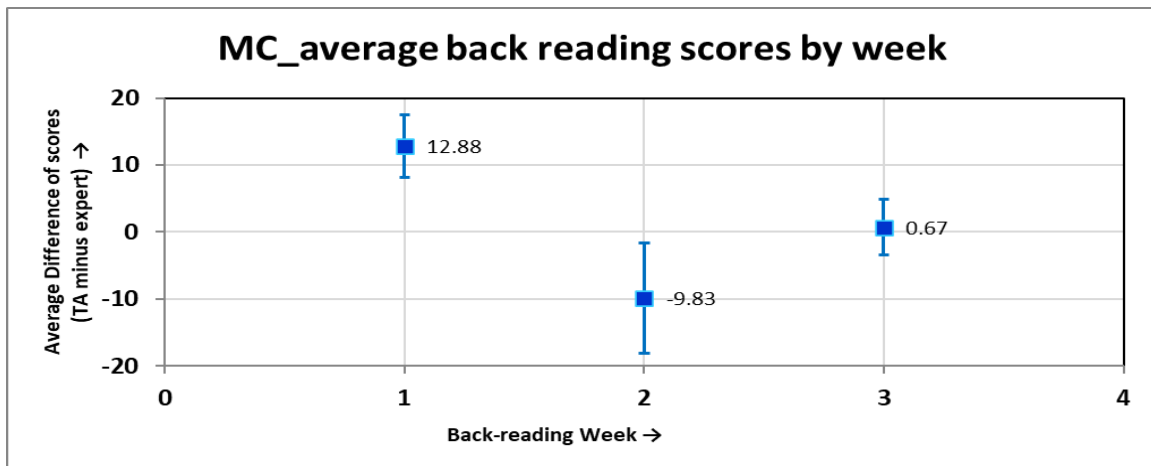


Figure 17:GTA MC's back-reading data during weeks 1-4

By week 4, GTA Clifford was more confident of the reliability of their grading and decided to stop attending back-reading meetings but agreed to provide reports for post checks.

With GTA Baldwin, the grading discrepancies were on the other end of the spectrum. While this GTA was off to a promising start in Week 1, subsequent individual back-reading meeting reflected a lenient grading approach and significant discrepancies wherein GTA

Baldwin consistently awarded higher scores relative to the expert. Discussion and back-reading revisions eventually resulted in overall agreement, but our observations also suggested that GTA Baldwin would need further support through back-reading.

Although GTA Baldwin did not attend any back-reading sessions after week 4, he agreed to provide graded reports for post-checks like all their peers. Figure 18 is a box-whisker plot showing GTA Baldwin's trend for individual back-reading during weeks 1-3.

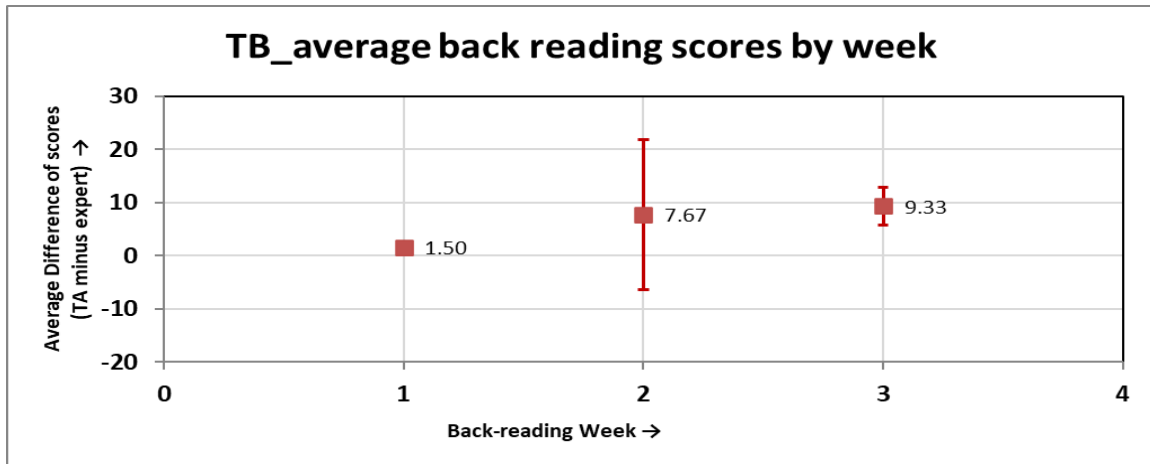


Figure 18:GTA TB's back-reading results during training weeks 1-4

2.5.4 Post-Training Protocol Results

Examination of back-reading data for weeks 4-8 was performed using the post-checks samples provided by GTAs. The head GTA obtained unmarked photocopies of at least three student reports as post-checks and graded them independently. Corresponding graded versions from GTAs was also photocopied and studied for comparison of scoring agreement.

The difference between GTA and expert was computed for each week, and a plot of average difference of scores was generated. Figure 19 shows the average difference and standard

deviation of back-read scores trends for GTAs Clifford and Baldwin for the entire duration of the fall academic term in Year 2 of our back-reading study. The *ideal* result would be absolute agreement for every back-read sample, i.e., within ± 1 point or even zero difference between raters. Absolute agreement of scores between the GTAs and expert(s) was impossible to achieve. However, the decrease in standard deviation or spread of scores is indicative of growing agreement with time and can be attributed in part, to GTA training in grading. For GTA Clifford, the average difference of scores was within the ± 1 point. We considered this as a promising trend as well as evidence GTA Clifford's positive uptake of the training in grading. A similar analysis for GTA Baldwin's individual back-read scores and post-checks shows us that the standard deviation remained consistent over time, it may be concluded that GTA Baldwin was a consistent grader (i.e., TB graded all the students fairly). However, since the average *difference of scores* shows an increasing trend over time, GTA Baldwin was *not* an accurate grader and may have implemented the rubric with deviations or errors. The overall upward trend in difference of scores between GTA Baldwin and expert may be considered evidence of either negative or negligible impact of the training provided. Although the option to attend back-reading sessions was recommended to GTA Baldwin on several occasions, non-participation, or diminished interest in utilizing the training could have contributed to this trend. This also speaks to our empirical hypothesis from individual back-reading observations that GTA Baldwin might have required more support through individual back-reading, which could have influenced his grading differently.

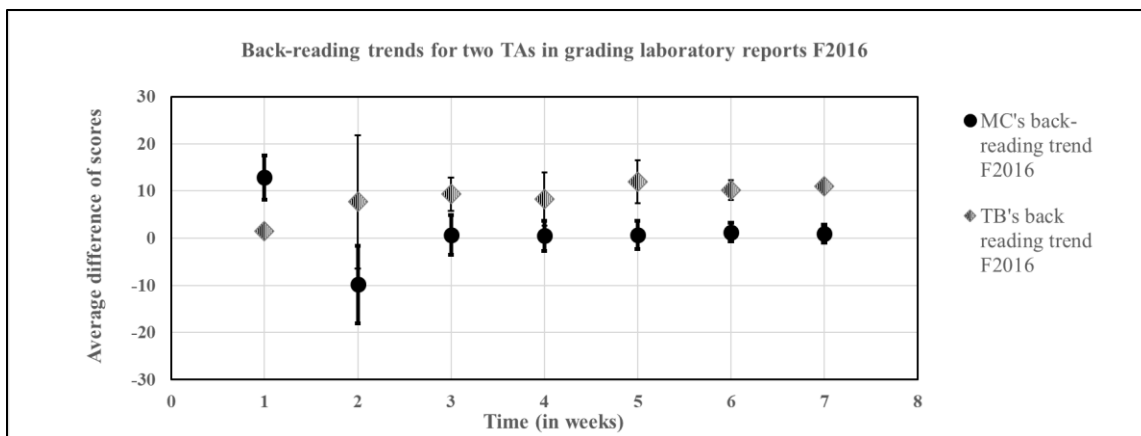


Figure 19:GTA MC's and GTA TB's data for the Fall term of Year 2 back-reading study.

2.6 Analysis & Discussion

2.6.1 Grading Accuracy

When interviewed about the back-reading experience, GTA Clifford reported that they were satisfied with their back-reading experience and were also able to address student concerns or questions regarding grading with more confidence. GTA Clifford also noted the growth in their own ability to grade any given laboratory report quickly and critically. Qualitative analysis of the feedback/comments provided by this GTA agreed with the expert's annotations on the corresponding example/post-check reports.

Similarly, GTA Baldwin reported being satisfied with their grading and back-reading experience during weeks 1-3. However, GTA Baldwin was unable to (or personally chose not to) attend further back-reading sessions despite recommendations based on random checks performed during week 3 and onward. One of the reasons for the recommendations was the observation that class average scores for TB's students were relatively and significantly higher than the overall class (n = 570).

This led us to consider two possibilities: either the students in TB's section were better than average writers and understood the chemistry content well as they drafted their laboratory reports or TB's grading overlooked many critical points as stated in the rubric, had some elements of bias (i.e., more leniency in grading) and was tailored to ensure students were "happy" with their grade, thereby making follow-up discussions about scores with GTA Baldwin unnecessary. Follow-up data analysis of the chemistry laboratory practical exam and chemistry lecture course exam scores (see Figure 20) indicated that TB's students were not above average compared to the larger class. Therefore, inaccurate grading does contribute heavily to improper assessment of students' learning and may also give them misleading grade expectations.

2.6.2 Grading Consistency

Grades for student work serve different purposes for various people. Students, for instance, benefit from grading by looking for feedback that is helpful, such as comments pointing out areas for improvement in understanding or presentation of their findings. Instructors are likely to consider assignment scores (or grades) as a measure of students' understanding and a reliable way to track students' learning curves. The average and standard deviation for a given assignment provides information on overall performance in the course on the score distribution. A small standard deviation could be indicative of either excellent student performance clustered around the average or of an extremely biased (lenient) GTA's grading. Both GTAs Clifford and Baldwin were quite consistent in their grading and have reasonable spread of scores. This could have been a reason for students not requiring

follow-ups with the GTA during office hours since they found their scores to be satisfactory (comparable to their peers).

2.6.3 Course Grade and Overall Learning Outcomes

At the end of the fall term, we provided more evidence to GTAs for the need to participate and assimilate training in grading. Our hypothesis at this stage was that GTAs impacted positively by the back-reading training would be accurate and consistent graders, providing their students with reliable assessments and constructive feedback on their laboratory reports. The students' performance in the laboratory practical exam was therefore likely to reflect the effect of GTAs teaching and grading.

Using overall class statistics, we further attempted to identify trends and outliers in accuracy and consistency of each GTA's grading patterns. Data is provided in Figure 20 for GTAs Clifford and Baldwin. The notable difference in TB's scoring compared to the expert scoring is cause for concern in more than one way. Clearly TB's grading patterns do not match the expert, indicating inaccurate use of the grading rubric. We also observed consistently higher scores for TB's students through weeks 4-7 which implied a lenient or inaccurate assessment of their laboratory report writing abilities. Consequently, TB's students did not receive accurate or realistic feedback on the quality of their work and may have gained a false sense of being successful in the course during the term. However, this also resulted in average or below average grades at the end of the term from the course instructor.

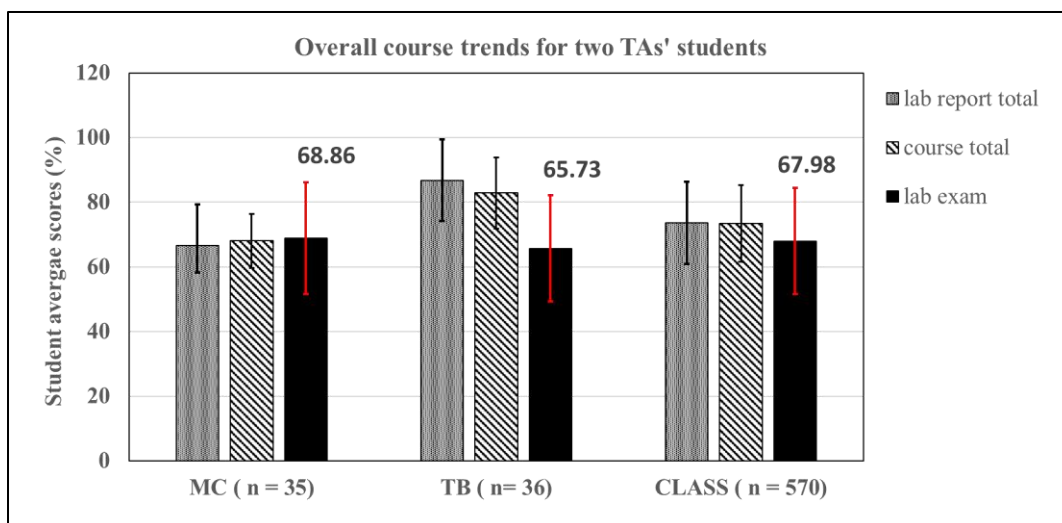


Figure 20: Comparison of overall course grades and laboratory practical exam performance for GTAs MC, TB with general chemistry laboratory course total student population

Having a lenient GTA like Baldwin during the first term of a three-term course is likely to set a misleading precedent for the students enrolling in subsequent terms of the chemistry course. Although, there is no guarantee that the students will have the same GTA again in the next term, the lack of relevant, constructive, and effective feedback in the beginning could impact student engagement in sequential terms. Admittedly, students will face differences in the assessment and feedback provided by their new GTA, if that were to be the case. Such contrasting messages lead students to feel conflicted and confused about the areas they could work on to improve their scientific writing skills. In the larger perspective, having TAs that grade very differently as well as provide significantly varying degrees of feedback may also negatively impact students' interest in pursuing a STEM major, thereby causing an undesirable attrition issue.

All these points were discussed with GTAs at the one-on-one end of term meeting between GTA and course instructors at the close of the term. In addition to conveying our gratitude to each GTAs for their time, service and participation in the course, instructors also

provided GTAs feedback on their evaluations from students, their own section's course performance and of course, grading skills. In turn, GTAs were also asked for their review or inputs for the back-reading training and many of them shared insightful experiences and suggestions verbally as well as in their survey responses.

2.6.4 GTA Feedback Comments

A snapshot of the comments from GTAs regarding the usefulness of the back-reading process is provided in Table 6 below. These are select responses from the pool of comments provided by GTAs in the end-of-term survey. Negative comments were limited to lower scores for Likert scale questions, and not available as open-responses. Based on these comments, we believe back-reading made a positive impact on GTA's approach to grading responsibilities and fine-tuned their laboratory report grading skills.

Table 6:GTA comments for open-response question on the usefulness of back-reading

■ <i>“Going over examples and grading rubrics were critical to understanding how to grade the lab reports fairly.”</i>
■ <i>“I think having a practice lab report to grade during the staff meeting was always useful, both [to] become adjusted to the rubric and to see what some common errors on the lab reports were.”</i>
■ <i>“Back reading session was very helpful because I was unclear on how the rubric should have been interpreted when I graded the [assignments] the first time around. “</i>
■ <i>“It helped me grade the open-ended questions asked of students where it wasn't a simple yes no answer.”</i>

Table 6 (continued): GTA comments for open-response question on the usefulness of back-reading

■ <i>“I felt that all of the work one-on-one with [the Head TA] were invaluable, and some of the small group discussions were helpful.”</i>
■ <i>“Helps to make sure you are interpreting the rubric as teacher/head TA intended. Makes it easier to grade more consistently because you can outline specifically what you are looking for.”</i>
■ <i>“Peer-grading helped me to see what other GTAs were removing points for.”</i>
■ <i>“Sessions taught me <u>what to avoid when grading</u> which is just as useful as knowing what to try.”</i>

2.7 Conclusion

2.7.1 Implications for GTA Training:

Training in grading and back-reading laboratory reports as a term-long, continuous GTA professional development process

Based on the data in our back-reading study and feedback from follow-up surveys with the GTAs, the necessity for training GTAs in grading has been substantiated. Although we recorded similar data for most GTAs and observed overall positive trends in grading accuracy and consistency, we have reported results for only two GTAs, Clifford and Baldwin, for the sake of brevity. We recommend that GTA training and professional development should not be restricted to only the “orientation” session at the start of them but provided as a term-long process. Monitoring GTAs’ grading and providing them with feedback on their grading practices throughout the term serves several purposes:

- 1) Students benefit by receiving reliable scores from GTAs, along with the reassurance of their scores being fair and accurate due to the back-reading process and sample grading of laboratory reports during the weekly staff meetings.
- 2) Students benefit by receiving accurate feedback on their weekly laboratory reports. This allows students to learn from their mistakes and improve their writing skills as well as their understanding of chemistry. A long-term effect of such reliable feedback to students is demonstrated trends of success and reduction of attrition in STEM disciplines.
- 3) GTAs understand and implement the grading rubric efficiently. Therefore, they are less likely to lag on grading and return graded assignments on time. Students are less likely to repeat their errors in writing if they receive graded reports in a timely manner.
- 4) GTAs and instructors become more aware of student learning and performance trends in by examining randomly sampled laboratory reports (qualitative) weekly averages and spread of scores (quantitative).

2.7.2 Limitations and Challenges in Back-reading Protocol

The back-reading process presented in this study also serves as needs-based platform for GTAs to share their concerns about grading, quality of student work, or the demands on their own time. For instance, if a GTA has many non-native speakers of English in their section, the quality of lab reports may not be comparable to another section with a higher number of domestic/ native English-speaking students. Performing back-reading with the GTA in this case allows for an honest discussion on how to provide more support for such students and assess their learning/understanding accurately and fairly. The impact of each

GTA's grading patterns on the students' learning and ability to write effective laboratory reports are unique. Exploration of the effect of each GTA's grading on students' abilities to draft a scientific report were planned as part of a later stage of this study.

As with any new remedial or intervention method proposed, there are several challenges associated with our back-reading process. One of the biggest hurdles is GTA participation and departmental/ instructor willingness to implement such a specific training program. The amount of effort and time required can be substantial depending on how many GTAs are involved in each course. The willingness and motivation of the Head GTA to organize and conduct grading-specific training is another factor to consider, since back-reading at this stage is intensive for any one individual and causes complex consistency issues if there are too many personnel involved. The effort on the part of the instructor or head GTA in working individually with each GTA to perform back-reading is also substantial, and at times may require a full-time grading expert or special assistant GTA position to be set up for term-long back-reading.

Training in grading not only impacts GTAs' grading practices, but also affects their teaching practices. Hence, it is also necessary to carefully select training examples that will orient GTAs to specific learning and assessment goals in that course. Other factors that could be challenges include buy-in from course instructors and department heads. GTAs are primarily graduate students pursuing their masters or doctoral degrees, and the time commitment for teaching and grading can sometimes be viewed as a hindrance in making progress in research^{9, 50, 65}. This may contribute to significant pushback on implementing back-reading even if GTAs acknowledge its utility.

To address one of the research questions for this study, incorporating training modules for GTAs to gain sufficient competency in grading should be an integral and substantial part of any GTA training/orientation program. At the very least, these tasks must be designed to provide GTAs with an opportunity to learn the characteristics of accurate and consistent grading through assessment of sample laboratory reports. If resources such as personnel, workspaces, and time for grading are available, the inclusion of back reading as a term long GTA training/GTA support module in grading can be a worthwhile addition to the program.

We believe back-reading is an effective method of ensuring accurate and consistent scoring among graders. Adapting this process and tailoring it to provide GTAs with training, resources and evidence of their successful work is important for fairness in grading the work of students, and for their professional development as future faculty. It is also critical that grading responsibilities be recognized to be on par with teaching responsibilities. There is a need for a shift in perspective regarding grading from “*grading, a time-consuming task*” to “*grading, an important task that is a major part of the job.*”

The positive influence of a back-reading component and weekly lab report grading on the scoring practices of UO GTAs is substantiated through anecdotal, statistical, and qualitative evidence presented in this paper. Most of the students enrolled in the first term of a general chemistry course sequence will require about four laboratory reports to begin to produce quality reports. It takes about four laboratory experiments for most GTAs to accurately grade laboratory reports. Future steps in this study are likely to focus on presenting detailed examples of the challenges in the implementation of back-reading methods with GTAs and gathering more quantifiable evidence of the positive influence of

this back-reading method on the grading practices of GTAs. We recommend grading-specific studies to further explore and validate the positive impact of a back reading protocol on GTA professional development and weekly grading of sample reports. Additional studies to explore how this enhancement of grading accuracy influences students' scientific writing abilities through laboratory reports is also recommended as a long-term goal.

CHAPTER III: QUALITATIVE ANALYSIS OF GRADED STUDENT LABORATORY REPORTS

3.1 Chapter Abstract

Design and implementation of a back-reading protocol reported earlier produced favorable results. One of the key goals of the back-reading study was to coach GTAs to accurately assess laboratory reports and provide necessary and effective feedback to their students. By ensuring such reliable and timely feedback that is, the overall effect of having accurate and reliable graders on student understanding of chemistry which is also a tangible impact of this study, would be available for further exploration. In this chapter we examine two specific GTAs' grading of student laboratory reports qualitatively, emphasizing areas where they successfully characterized good conceptual understanding and scientific writing by inclusion of comments or annotations. We also highlight areas of student reports where comments or annotations would have been helpful, but were missing or not necessarily provided in a desirable format. A discussion of the impact of the type of feedback comments and annotations while grading student laboratory reports follows the detailed analysis of these exemplars.

3.2 Introduction

A back-reading (BR) protocol used for training GTAs was implemented at UO, a university in the Pacific Northwest. Goals for the GTA training programs were: (1) to ensure inclusion of training in grading in the start-of-term TA training/orientation program and (2) provide TAs with resources and support to ensure back-reading is administered (or

made available as an option) throughout the duration of the course in addition to orientation week. Challenges to the establishment and continued implementation of such target-specific BR protocols exist. For example:

- a) Approval from university and departmental decision-makers who will want to weigh the necessity and time/labor investment into such efforts.
- b) GTA participation: Recognition and willingness to engage in professional development activities which may not provide them with immediate, tangible results.
- c) Training Personnel: As described previously, BR can be a very intense exercise for a single individual. If the course instructor decides to undertake BR in addition to other responsibilities it may have a detrimental effect on not only BR exercises, but also other commitments. A critical step is to identify, hire and/or invite suitable personnel who are trained and experienced with grading, willing to lead BR activities and can communicate findings, suggestions to GTAs constructively. A key responsibility of such personnel is also to provide necessary support and interventions for GTAs and can impact the success of a BR program significantly.

One of the essential factors that ensures continuity of BR -like programs is the pattern of demonstrated success. During Year 1 and 2 (fall, winter, and spring terms), we collected BR data from GTAs during staff meetings, individual BR, and post-check samples from GTAs. We also collected feedback comments every week, observations during BR meetings, and solicited end-of-term suggestions from GTAs using a survey. The gathered information helped build a clearer picture with each iteration of the BR method administered at UO. Instead of simply addressing the grading discrepancies between GTAs

and experts, we also sought to explore the subtler influences of BR exercises on the GTAs' *chemistry content* assessment of student laboratory reports. In this chapter, we examine qualitatively how GTAs (1) grade laboratory reports reliably and consistently and (2) provide relevant, constructive, and effective written comments/ feedback to their students. Here are some key definitions of the terms used in the context of back-reading.

- Reliable grading: The use of the rubric to assess written work or responses is accurate and reflects interpretation of rubric as intended by the instructor.
- Consistent grading: Grading of multiple student reports is not impacted by (a) any personal bias on the TAs part, (b) time lags between grading reports, (c) consistent awarding or deducting points.
- Relevant feedback: Comments or information within the context of the laboratory report content/ chemistry topic being graded.
- Constructive feedback: comments or annotations that not limited to simply pointing out shortcomings in written reports but also provide guidance on addressing them.
- Effective feedback: TAs comments or annotations that cause noticeable changes in students writing skills such as not repeating previous errors or including proper scientific formatting where applicable.

3.2.1 Qualitative Research Context

First-year undergraduate students completed a general chemistry course in three-quarter (fall, winter, spring) terms. The fall term commences with basic skills in the laboratory and progressively increases the content and skill sets provided to students over winter and

spring terms. For example, the fall term experiments are simple chemical reactions like galvanization of iron and determination of percent carbonates in compounds. Spring term experiments are more complex, such as acid-base titrations and kinetic studies.

The syllabus for the general chemistry laboratory course outlines the following learning goals:

1. Hands-on skills with basic laboratory equipment and procedures
2. An understanding of how to use common laboratory chemistry techniques to solve chemical problems
3. An understanding of how to carry out experiments by formulating questions, following instructions, and recording data
4. Knowledge of chemical and laboratory safety
5. The ability to think analytically from data collected by students and their classmates
6. The ability to work effectively as part of a collaborative team
7. The ability to write an effective, properly formatted scientific report

As is evident from learning goals #5-7, students should be able to use data collected during experiments, work collaboratively with other groups, and perform analyses to provide reasonable evidence to support their claims. The required format for student laboratory reports is based on a guided inquiry scientific writing heuristic approach^{15, 82}.

A qualitative examination of several student laboratory reports from the winter term in year 2 of this study was performed, and feedback comments provided by GTAs were analyzed. The findings are specific to identifying areas where GTAs provide relevant, effective, and

constructive feedback or deviate from this desired track. This in-depth analysis also provided the necessary input to amend the protocols for training GTAs in grading at various stages in our study. There were seventeen (17) total GTAs teaching twenty-six (26) sections, each with 18 students during the winter term. Considering that students turned in written laboratory reports for seven (7) out of the total ten (10) weeks, and that back-reading protocols resulted in examination of four graded reports on average from each GTA, a starting pool of nearly 480 reports was available for qualitative analysis. We examined six (6) randomly selected GTAs' graded reports for the winter term (that is, ~168 reports). We then narrowed our qualitative analysis to graded reports from two GTAs Molly and Klaus, examining a total of 15 reports for each GTA and included an in-depth analysis of sections from *four* laboratory reports. This selection was made on the basis of data that would be representative of GTA experience, gender and back-reading participation as well as quality of student work.

3.2.2 Qualitative Analysis Subjects

We report our analysis for assessments by two GTAs: Molly and Klaus. Each individual represents a unique demography or set of characteristics from the GTA pool.

GTA Molly

Molly (*name changed*) was a second-year female graduate student and second-year GTA in the general chemistry course. She successfully completed one year of teaching as a general chemistry laboratory GTA and based on observational and anecdotal evidence, was one of the positively engaged GTAs in the BR process and research study. This chapter focuses on the grading training and examples relevant to a winter term. Molly's GTA

responsibilities for this specific term included teaching and grading for ONE laboratory section of twenty students enrolled in the general chemistry laboratory course.

GTA Klaus

Klaus (*name changed*) was a male graduate student and a first-time TA for the general chemistry laboratory course in the winter term. Klaus had just completed the preliminary orientation for general chemistry GTAs, provided to all GTAs before the weekly laboratory sessions began for the academic term. BR training was provided to all GTAs at every start-of-term orientation irrespective of whether they had been through it earlier. There are no observations regarding Klaus' engagement (or otherwise) with the BR process and study, since this was his very first time teaching the general chemistry laboratory course.

3.2.3 Qualitative Data Sources

We focused on examples drawn from Molly and Klaus selected from a group of 17 GTAs in the winter term. As described previously, GTAs were requested to provide, at minimum, THREE graded laboratory reports that could be photocopied/scanned before being returned to the student(s). It was also suggested that GTAs might select such exemplars to [ideally] reflect high, medium, and low quality of student work after grading each week. These exemplars were used for post-checks in the BR study, assessed by the researcher and compared with GTA's scoring. The student would have received their graded work back by the time the post checks were completed. Any discrepancies in scoring were noted, but not necessarily communicated due to lack of opportunity to address them at this stage. However, if a consistent pattern of large discrepancies was observed for two or more weeks, the GTA was informed verbally and requested to consider attending BR meeting/s

again. As always, this was entirely optional and almost never taken up by any GTA due to limited time availability. Almost all GTAs who were part of the teaching crew in general chemistry obliged and provided such exemplars each week. GTAs responsible for two laboratory sections were very considerate and provided at least three exemplars per section i.e., a total of SIX graded laboratory reports. Some GTAs who were mindful and cognizant of the significance of the BR study also provided additional reports on occasion, specifying the noticeable quality of student work as the reason for such an addition to the weekly exemplars. Table 7 is a list of the experiments performed during this term.

Table 7: List of laboratory experiments performed during winter term.

Week	Title
1	Graphing
2	Emission Spectroscopy
4	Molecular Models
5	Gas Laws
6	Spectrophotometric Determination of Food Dyes
7	Forensic Analysis of Ink
8	Kinetics Exploration
9	Kinetic Studies: Bleaching of Allura Red
10	Lab Practical Exam

For purposes of brevity, we limit our results and discussion based on experiment # 6, ‘Spectrophotometric Determination of Food Dyes’. A detailed reproduction of this experiment from the chemistry laboratory manual ¹⁰⁵ is provided in Appendix C for further reference. A summary of this experiment is provided below.

3.2.4 The Food Dyes Laboratory Experiment

The purpose of the food dyes lab was to examine artificial food coloring used in edible commercial products such as energy drinks and calculate the concentration of the food dye used in them. This was achieved by measuring absorbance and using Beer's Law (Absorbance (A) = ϵbc (where ϵ is the molar absorptivity constant in $M^{-1}cm^{-1}$; b is the path length in cm; and c is the concentration in M) . Students were also expected to pool data for different colored dyes and understand how the concentrations of such additives may be harmful, if not kept in check. The laboratory experiment was designed to help students understand spectrophotometry as a chemistry technique for such analyses. Laboratory learning outcomes were primarily preparation of stock solution, dilutions, recording absorbance data and interpreting data using a calibration curve. Students were provided with a 1-component colored food dye (red, yellow, blue) to prepare stock solutions and dilutions and generate a calibration curve. Following this, an unknown concentration food dye, a two-component dye (green, orange, or purple) food dye was provided. Students were expected to prepare a set of standard dilutions and use absorbance data to determine concentration of each component in the unknown sample.

3.2.5 Examples of Student Work for Inductive Analysis

As described in the previous chapter, GTAs were requested to provide high, medium, and low-quality student exemplars for our BR study each week. Figures 21-34 in the following sections of this chapter are screen captures of scanned student reports with GTA annotations from the Food Dyes laboratory experiment. Each GTA provided these reports as representative of “high” and “low” quality exemplars for our BR post-check process.

Each figure is labelled as “GTA (source), Section of report, GTA Assigned quality” for each section of the exemplar i.e., “Introduction”; ‘Claims and Evidence’; ‘Reflection’ sections discussed. We focused on specifically these three sections because of the abundance of written content from students, which makes it relatively easier to identify any errors or misconceptions in student thinking or writing compared to a data table or graphs. Examining GTA comments/ annotations for any identified errors or misconceptions, in turn, helped us fine-tune BR training for GTAs addressing these areas of improvement. The laboratory report writing guidelines provided to students is included in Appendix D for readers’ reference.

Note on writing the introduction section of a laboratory report

In one of our GTA’s discussions with their students about writing laboratory reports, (recorded observations during a laboratory visit),

“Your (student’s) written laboratory report must be a complete document, that [if and when] given to another individual for experiment replication, should help them perform the entire experiment without any hitches. The presentation of observations, data, calculations, and analysis must also be comprehensive and enable the other individual to verify their own data and findings with reference to those reported in this document.”

From the student laboratory manual basic report writing pointers (pages xvi-xxi)¹⁰⁶ we see that the introduction section should include information about any relevant scientific theory, definitions, brief description of phenomena of interest and any chemical reactions or formulae that would be used in this experiment. In another GTA’s words,

“The introduction section is setting the stage for your laboratory report. You (student authoring the report) want to provide, as briefly as possible, any and all relevant information for the experiment that will help the reader understand the data and analysis that follows.”

Note on writing the claims and evidence section of a laboratory report

For UO student laboratory reports the claims and evidence section is expected to be brief and must answer any beginning questions presented in the introduction / earlier sections of the report. Ideal responses would be extremely ‘crisp’ with one-two sentences stating claims (or findings) and three-four sentences providing substantiating evidence including references to data, graphs or other results presented in the report.

Any experiment performed in the laboratory is bound to have data or trends that do not follow the expected track. When applicable, students are expected to analyze any errors, deviations or observation that point to alternate hypotheses in the ‘analysis and discussion’ section of the laboratory report. Alternatively, the laboratory manual will at times include a set of discrete questions that either (a) require students to use their experimental data/ results to provide a logical response or (b) use their understanding of laboratory techniques, experimental design ideas and chemistry content understanding gained during the experiment to explain a specific pattern or deviation. In such cases, the analysis and discussion section are often a paragraph-like response to each question. Students are still expected to substantiate their responses with references to data, results or graphs generated while performing the experiment.

Note on writing the reflection section of a laboratory report

The reflection section of the report is primarily to encourage students to think beyond the present experiment and provide evidence of a broader understanding of the chemistry content or phenomena. For every report, there are primarily five responses expected from students for the reflection section. Table 8 lists these prompts along with a brief explanation of the anticipated outcomes.

Table 8: Reflection section prompts for student laboratory reports

Prompt	Response features	Outcomes
Extend	Provide a reasonable, accurate and meaningful extension of the experiment performed and justify the idea	allows us to understand their ability to independently frame a novel idea/ research question to explore further as well as provide justification for its need and experimental design showing achievability.
Connect	Accurate articulation of related concepts/ideas from lecture course	visualize students' ability to navigate the macro-micro-symbolic spectrum in chemistry.
Apply	list and briefly explain relevant applications from real life	quality of the explanation shows us how the student utilizes their chemistry knowledge to understand, interpret and problem-solve in real-life scenarios.
Literature	Provide appropriate literature sources/citations that conform or dispute data or findings.	Demonstrates students' ability to examine other sources of data or results from other researchers; report and cite them in an acceptable format
Green Chemistry	List or elaborate relevant GC principles used in experiment	Demonstrates understanding of GC principles and ability to implement them correctly.

With this contextual information about the food dyes experiment and laboratory report expectations, we then proceed with our examination and analysis of exemplars from Molly

and Klaus. Figures 21-23 are screen captures of specific sections from a high-quality report from GTA Molly and Figure 24-26 from a high-quality report provided by GTA Klaus. This is followed by a qualitative analysis of low-quality reports (Figures 27-29 and 30-34 respectively) from GTAs Molly and Klaus. For purposes of clear demarcation, student responses reproduced as part of the text is highlighted as italicized font. Any suggestion or modifications to these responses (by way of comments or annotations from the researcher) are also included in italicized font.

3.3 GTA Molly's High Quality Student Report

3.3.1 Introduction Section

Figure 21 is a screen capture of the introduction section of a high-quality report from GTA Molly. From the student guidelines document (see Appendix D) we can see that this student has covered all the anticipated points for this section: basic theory, definitions/formulae, purpose of the experiment, beginning questions, brief procedure and safety protocols required. Let us now examine each response qualitatively. There is scope for improvement in the first paragraph which introduces this experiment. Usage of phrasing such as “*on top of that*” is definitely not scientific and deserves at least a cursory underlining or highlighting by the GTA to point out choice of phrase. Similarly, the problem statement about food dyes in health drinks is not clearly phrased, since just study of health effects is not necessarily cause for controversy. This could be rewritten (possibly) as, “*people are concerned about the amount of food dye in commercially available drinks because of potential health risks.*” Similarly, the purpose of the lab was to explore the use of spectrophotometry as a quantitative analysis method to determine the concentration of food

dyes in specific commercially available drinks and whether these concentrations were safe for human consumption.

Introduction

Food coloring is and has been an essential part of the drink industry. More often than not drink companies will add food coloring to improve the overall appeal of the drink by improving appearance of the drink as well as to give it a captivating name that makes consumers want it more; examples of this are electric yellow, blue raspberry, or racing red. On top of that, colors are now associated with certain flavors and traits. By being able to visually examine the color of a food, we can more accurately expect what flavor or trait they will have red means spicy, blue means cold. Controversy comes into play when the health effects of food additives are studied. People are concerned about whether or not the amount of food coloring present in the different drinks is safe to consume.

The purpose of this lab was not to examine whether or not the food coloring found in popular drinks is safe but rather to determine how much food coloring is used in popular drinks. The lab also introduces us to spectrophotometry as a form of quantitative analysis. Spectrophotometry allowed us to determine the concentration of a certain food dye by plotting absorbance versus concentration. — *calibration!* — 1

Spectrophotometry allows us to find the concentration because a spectrometer measures the absorbance of light. The basics behind the graph of a spectrometer is that it shows which light is absorbed and which is not absorbed, and the light that is visible to our eyes is the light that is not absorbed. To find the concentration, we used the Beer-Lambert Law. The Beer-Lambert Law gives us the equation $A = \epsilon bc$, which is absorbance equals molar absorptivity multiplied by path length multiplied by concentration. When absorbance and concentration are graphed together, the slope is ϵb .

To find the absorbance of a liquid, we had to perform the following steps. Our given food dye was FD&C Yellow No. 5. The first part of the lab had us create the solution. We first calculated, the necessary amount of food dye needed to make 500mL of solution. We then proceeded to weigh and measure the correct amount using weighing paper and a balance. Once we reached the necessary amount of food dye, we then created a solution by mixing the food dye and water to create 500mL of solution. In the next part of the lab, we had four 100mL beakers to create various concentrations of the liquid. Each beaker was labeled A, B, C, or D, and each had varying concentrations; A had 1mL of solution, B had 5mL, C had 10mL, and D had 15mL. We then filled the rest of each volumetric flask until we arrived at 100mL. In the third part of the lab, we used the spectrometer to find the different absorbance levels of each of the new solutions. In the final part of the lab, we were given a mixture of two food dyes in a popular drink and had to find the absorbance for that solution as well.

Personal beginning questions I had for this lab were: what association is there with color and taste, as well as how can we identify the food dye used in everyday foods? Questions that the class came up with together was; is the molar absorptivity constant different for different substances, along with how do you determine if they are different? The last question the class came up with was; how can absorbance/transmittance determine health risks?

Figure 21:GTA Molly; Introduction section from high-quality report

The student has successfully included relevant definitions and formulae for the Beer-Lambert law as part of background theory for this report. However, the use of “*which light is absorbed, and which light is not absorbed*” deserves a cursory annotation from the GTA because it is not scientifically phrased. A more appropriate way of saying this would be

“the spectrophotometer detects the portion of light transmitted from the sample.” Using the Beer Lambert Law, we can determine the concentration of the solution using absorbance data. Description of the experimental procedure is also satisfactory. The use of “beakers” can be misleading if this report were used to replicate the experiment because the use of volumetric flasks (not beakers) is required for making accurate dilutions from a stock solution. During the last part of the experiment, pairs/groups were assigned a food dye containing two components (e.g., green food dye is made of blue and yellow component dyes) and required to calculate the concentration of components in this food dye (not just record the absorbance) The student has also not mentioned a calibration curve which would have been essential for determining concentration of an unknown. The grading rubric allows for a maximum of ten points for the introduction section. GTA Molly awarded this response 9 out of 10, whereas based on the above analysis this response would merit 6 out of 10 at best. Based on the grading rubric, the scores awarded are provided below. We can see there are discrepancies in not only the scores but also the feedback or annotations which could help this student write an excellent scientific report.

Table 9: Summary of scores awarded for the introduction section in GTA Molly's high-quality report.

<i>Introduction</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on Qualitative analysis</i>
<i>Theory /Purpose</i>	3	2	1
<i>Beginning Question(s)</i>	2	2	2
<i>Safety statement (s)</i>	2	2	2
<i>Experimental Procedure</i>	3	3	1

3.3.2 Claims and Evidence Section

Similarly, an examination of the claims and evidence section (Figure 22) in the high-quality report is presented below. Here, the student states and answers their beginning questions. However, their quality of answers may not necessarily qualify as a ‘high-quality’ response. For example, in response to “how can we identify food dyes in everyday foods?” a reasonable scientific answer would be phrased as “*we can achieve this [goal] by simple visual identification and if we wanted to be extremely precise, by using spectrophotometric data. Recording the absorbance of the sample and comparing this value to known absorbance data reported in literature can help us identify a specific-colored food dye or ingredient.*”

Claims and Evidence

After finishing this lab, I was able to answer many of my personal beginning questions as well as the class' beginning questions. One of my personal questions that was answered through this lab was how can we identify the food dye used in everyday foods? We can identify the food dye used by examining what color it is and even going as far as using spectrophotometric analysis to be more specific about the concentration and absorbance of the dye. If we obtained a sample of a food and broke it down, we would most likely be able to use spectrophotometric analysis. The two class questions we were able to answer through this lab are; is the molar absorptivity constant different for different substances, along with how do you determine if they are different? Second, how can absorbance/transmittance determine health risks? The molar absorptivity constant is different for different substances because molar absorptivity is a measure of how well something absorbs light, and since in this lab we can see that different color absorbs different amounts of light then we can conclude that they are different. To find the molar absorptivity constant of a substance we would use the Beer-Lambert Law and rearrange it so that the product is molar absorptivity. Absorbance/ transmittance can help to show health risks of different chemicals by showing the maximum amount of a chemical that can be absorbed or transmitted and what the concentration of that chemical can be. In other words, absorbance/transmittance allows us to determine a threshold that we cannot pass or else we may be in risk of disease or sickness.

When analyzing the graph of the green juice, obtained through the use of the spectrometer, we were able to find that the green juice is made up of a mixture of blue and yellow dye as expected. By examining the graph we could also tell that more yellow dye was used than blue dye because yellow had a higher absorbance and wavelength meaning that more light was absorbed.

not really help find one.

need to look at conc, not just abs.

Figure 22: GTA Molly; Claims-evidence section from high- quality report

For the second question, “*is the molar absorptivity coefficient different for different substances?*” an expected response would confirm this fact and include actual data

collected during the experiment, such as: “Yes, the molar absorptivity coefficient (ϵ) for different colored food dyes is different. This is because every color absorbs and transmits a definite quantity of incident light, and according to Beer’s law $A = \epsilon bc$ (where A is the absorbance; c is the concentration of the dye and b is the path length). From our class data and calculations, we recorded the following values for different food dyes.” Finally, for the last beginning question, “how can absorbance or transmittance determine health risks?” a well-phrased scientific response would reference the calculations section and highlight the concentrations of the food dye calculated in a commercial drink.

Further using the LD₅₀ values known from literature (Table 10), the claim regarding health risks could be supported or disputed based on the comparative values of concentration from the actual data/calculations. The student makes a good attempt to explain the components in the green dye (blue and yellow), however none of their statements are substantiated with data or results from the report.

Table 10: Recorded values for different food dyes

Color	Yellow	Red	Blue	Green
ϵ (L mol ⁻¹ cm ⁻¹)	2.73x10 ⁴	2.13x10 ⁴	1.38 x 10 ⁴	4.30 x 10 ⁴
Wavelength (nm)	427	504	631	620

The grading rubric allows for a total of six points for this section and based on our analysis of the best possible responses described above, the claims evidence section would be awarded 2 or 3 points. GTA Molly awarded this response five out of 6 points.

3.3.3 Reflection Section

Lastly, let us closely examine at Figure 23 (below) which is a screen capture of the reflection section from GTA Molly's high-quality student report.

A reasonable extension of the experiment allows for a new research question to be answered (i.e., not repeating the exact same experiment) with different or newer variables. For example, instead of phrasing the extension as "*I would also like to test my own drinks, and drinks that I have in my fridge and drink regularly because I'm interested in seeing the amount of dye used in my personal favorites.*" the student could have proposed to examine colored cereal or candy which often contains food dyes. By examining a solid product (instead of liquids), not only has the sample changed, but it also requires a specific design of experiments where the sample and diluted aliquot preparation is different, and other edible ingredients could play a role in the health impacts of such foods.

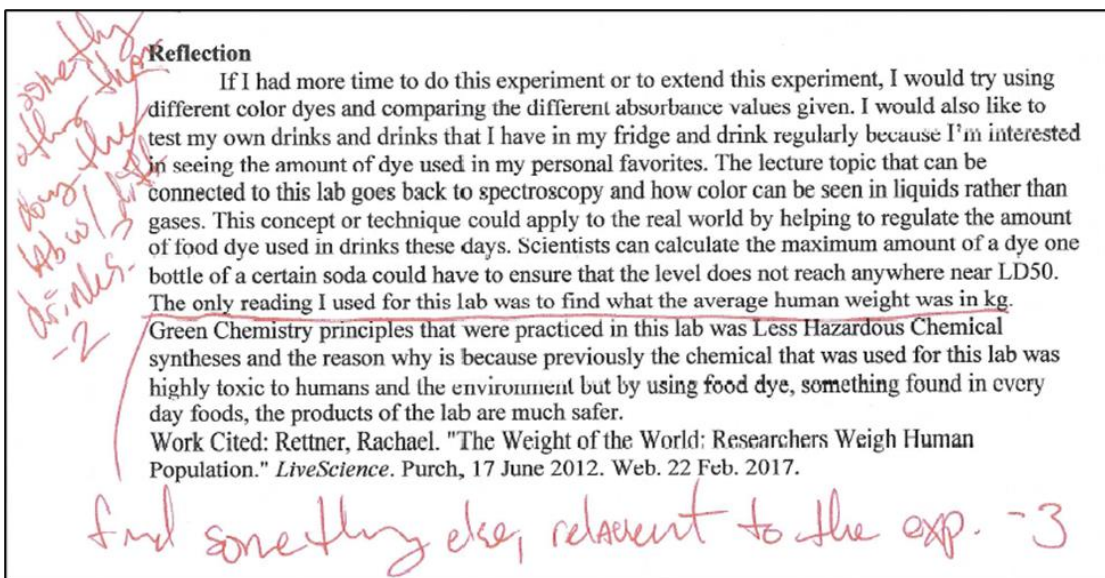


Figure 23: GTA Molly; Reflection section from high-quality report

The lecture connection response “...goes back to spectroscopy and how color can be seen in liquids rather than gases.” is indicative that this student may not have accurately understood the topics connected to this experiment. Here, an acceptable response would be phrased in the context of materials absorbing specific wavelengths of incident light and transmitting others, thereby giving an object the color, it is associated with.

For the Application response, the student states, “...could apply to the real world by helping to regulate the amount of food dye used in drinks these days. Scientists can calculate the maximum amount of dye one bottle of a certain soda could have to ensure that the level does not reach anywhere near LD50.” Although this response is reasonable, the student definitely missed the opportunity to use stating this application in tandem with a literature or news article citation to help drive the point home. For example, “...in this 2007³ article, McCann, Barrett et.al. undertook a randomized, double-blinded, placebo-controlled, crossover trial to test whether intake of artificial food color and additives (AFCA) affected childhood behavior. Artificial colors or a sodium benzoate preservative (or both) in the diet result in increased hyperactivity in 3-year-old and 8/9-year-old children in the general population.” Performing spectrophotometric analysis or using relevant data for ingested dyes was part of the experimental protocol. Of course, a corresponding citation³ of this source in the references section is also expected. The related reading (literature) response includes GTA’s comment about looking for something relevant to the experiment. In hindsight, if the student had used the example above or something along the same lines,

³ McCann, Donna, et al. "Food Additives and Hyperactive Behavior in 3-year-old and 8/9-year-old Children in the Community: A Randomized, Double-blinded, Placebo-controlled Trial." *The Lancet (British Edition)* 370.9598 (2007): 1560-567.

the ‘application’ and ‘related reading’ response would have been a case of addressing two birds with one stone.

Finally, the ‘green chemistry’ response lists the relevant principle (less hazardous chemicals) but provides a somewhat garbled explanation of how this is applied. Stating “... *the chemical that was used for this lab was highly toxic to humans and the environment...*” leaves the reader hanging, not knowing what exactly was so toxic that using food dyes for this experiment was much safer and in alignment with desired green chemistry principles.

Table 11: Summary of scores awarded by GTA and researcher for reflection section

<i>Prompt</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on qualitative analysis</i>
<i>Extend</i>	2	0	0
<i>Connect</i>	2	2	0
<i>Apply</i>	2	2	1
<i>Related reading</i>	2	0	0
<i>Green Chemistry</i>	2	2	1

3.3.4 Summary

Based on our qualitative examination, we observe that GTA Molly picked out key responses where the student could improve their scientific writing skills. We did not observe much of an emphasis on identifying and /or annotating conceptual errors from the student (such as incorrect phrasing or omission of actual data to substantiate a claim). Our qualitative analysis also estimated the scores that would be awarded, and as seen in Table 12 below, the scores do have some discrepancies. From our overall report scores for this student (Table 12) GTA Molly awarded this report 90/100. The BR score from the researcher was 81/100 for this report. In other words, a 9-point difference overall, and if we

factor in the scores based on the qualitative analysis performed above, the score awarded by the researcher was 80/100 that is a difference of nearly 10 points between GTA and researcher. Notably, the report was back-read once by the researcher for post-checks and found to have a difference of nine points and again for the present qualitative analysis and found to have nearly the same difference. This can be considered evidence of the effect of back-reading, demonstrating the researcher's accuracy and consistency of grading the exact same lab report and awarding nearly the same score with a 2-year gap between backreading and qualitative examination. GTA Molly's assessment of the quality of this report as "high" when providing it for post-checks is indicative of a different problem that needs to be addressed.

Table 12: Total scores for sections in Molly's high-quality report

<i>Section</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on qualitative analysis</i>
<i>Introduction</i>	<i>10</i>	<i>9</i>	<i>6</i>
<i>Claims/Evidence</i>	<i>6</i>	<i>6</i>	<i>3</i>
<i>Reflection</i>	<i>12</i>	<i>6</i>	<i>2</i>

A high-quality report should fulfil the characteristics of having high-quality responses along with grader annotations that show minimal chemistry content corrections thereby reflecting sound understanding on the student's part. However, our qualitative examination provides evidence of a lack of accurate assessment of student understanding and possibly the effect of GTA Molly's inherent bias about quality of student work based on whose report they are grading. That is to say, allowing previous expectations of this specific student's reports to allow the continued assumption that the quality was high with no

fluctuations or errors. In this case; a simple refresher of the BR technique may have helped GTA Molly at this stage and continued her trend of accurate and reliable grading.

3.4 GTA Klaus' High Quality Student Report

3.4.1 Introduction section

The introduction section of GTA Klaus' high quality report (Fig. 24) shows a good start of introducing the topic that "... *almost all foods have a food dye in them*" and clear purpose of this experiment stated as, "... *to observe the relationship between concentration and absorbance, The goal: to determine the concentration of food dyes in different drinks using spectrophotometers.*" This is followed by a detailed description of the spectrophotometer, its working mechanism and an explanation of how concentration can be calculated using Beer-Lambert's law. There are some concerning phrasing choices in the description of the working of the spectrophotometer. E.g., "...*hits a dispersion device*" (could be rephrased as: is incident on a prism for dispersion) or "*then only specific wavelengths are allowed to pass through and hit the sample in the cuvette*" (how exactly is this accomplished?). The most critical statement that should have been addressed by GTA here is: "*An absorbance spectrum shows the amount of absorbance over a series of wavelengths.*" Now, an absorbance spectrum is recorded at a single wavelength (λ_{max}) not a series of wavelengths. Secondly, '*amount of absorbance*' is not the accurate scientific phrase quantifying the ratio of incident light to transmitted light and could have been rephrased as "*the intensity of the absorbance curve*". Beginning questions listed are not didactic (i.e., not yes or no response type questions) and therefore, acceptable research questions.

Today, almost all foods have some food dye in them, even oranges! In this experiment, we tested the spectrophotometric properties of multiple food dyes, in different concentrations and in different drinks. The purpose of this experiment, the Spectrophotometric Determination of a Food Dye, was to observe the relationship between concentration and absorbance. The goal: to determine the concentration of these dyes in different drinks using spectrophotometers. The Spectrophotometer Process: light from a light source enters the spectrophotometer through a small slit and hits a dispersion device, which causes the different wavelengths to refract at different angles, splitting white light into a spectrum. Then, only specific wavelengths are allowed to pass through and hit the sample in the cuvette. Light that passes through the sample then hits the detector. The detector measures the intensity of the light that passes through the sample. Light that does not pass through the sample has been absorbed. An absorbance spectrum shows the amount of absorbance over a series of wavelengths. Lambda max (λ_{max}) is the wavelength at which there is the highest absorbance.

With A as absorbance, ϵ as the molar absorptivity coefficient ($M^{-1}cm^{-1}$), b as the path length the light has to travel (cm) and c the concentration of the dye (M), the Beer-Lambert law states that:

$$A = \epsilon bc + \text{constant}$$

This law can be rearranged to calculate the concentration of a solution. Epsilon is found by looking at the Beer-Lambert Law as a linear relationship, and the constant (the y intercept) should be zero.

Beginning questions for this lab were: Why use lambda max? What effect does concentration have on absorbance peaks?

Though food dyes are fairly safe, safety precautions were taken during the lab as though the substances used were unknown. Gloves were worn at all times when handling dyes. Caution was used when handling dye powder to avoid inhalation.

Handwritten notes in red: "use symbol" (pointing to ϵ), "considering" (pointing to "looking at"), "calibration" (written below "linear relationship"), "e" (written above "epsilon").

Handwritten notes in blue: "General safety" (written at the bottom).

Figure 24: GTA Klaus; Introduction section from high quality report

As we can see in the screen capture in Figure 24, the GTA has not provided any comments or annotations in this paragraph meaning either this was not noticed at all or if noticed, it did not merit any corrective feedback. Safety comments are satisfactory. However, a brief description of the experimental procedure is missing. Table 13 is a summary of the point breakdown for scores awarded to these responses.

Table 13: GTA Klaus; Scores awarded to introduction section in high-quality report

<i>Introduction</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on Qualitative analysis</i>
<i>Theory /Purpose</i>	3	2	2
<i>Beginning Question(s)</i>	2	2	1
<i>Safety statement (s)</i>	2	1	1
<i>Experimental Procedure</i>	3	3	3

3.4.2 Claims and Evidence

The claims and evidence section from Klaus' high-quality report is reproduced in figure 25. As described earlier, in this section answers to the beginning questions (BQs) are expected along with supporting evidence from data, results and graphs.

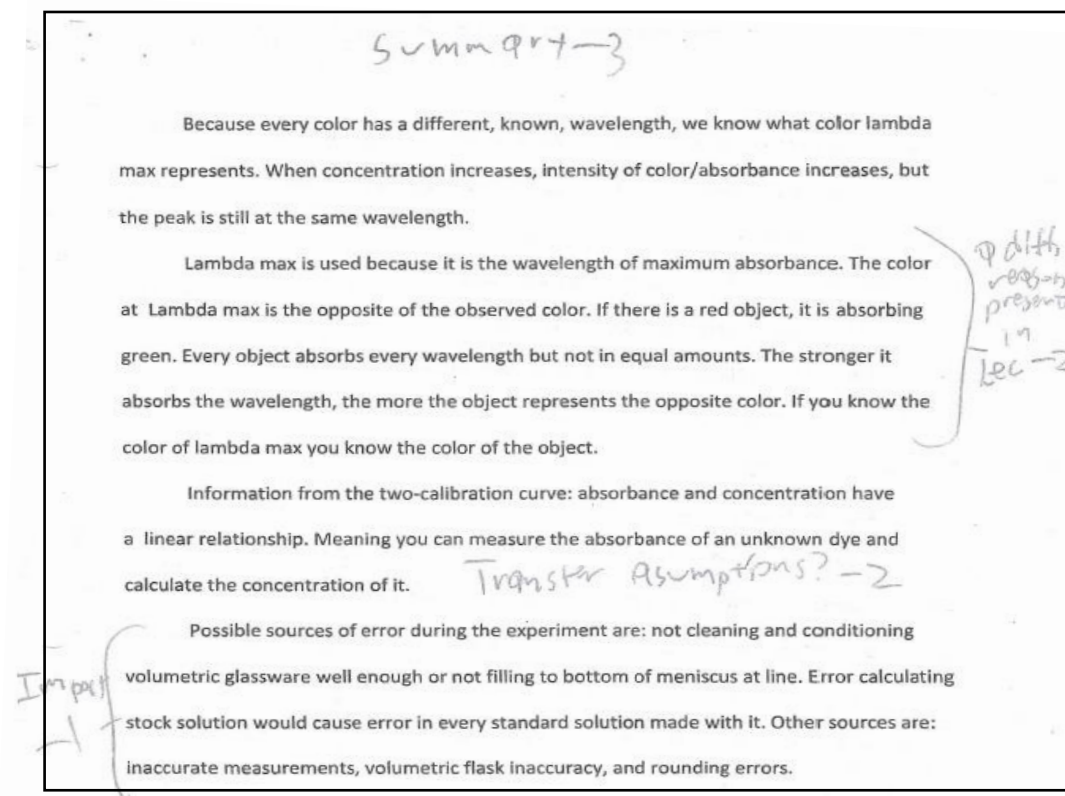


Figure 25: GTA Klaus; Claims and evidence section from high-quality report

In response to the student's first beginning question, "*why use lambda max?*" we see a clear and accurate claim statement, "*Lambda max is used because it is the wavelength of maximum absorbance.*" The student expands on the evidence for this by saying "*If there is a red object, it is absorbing green....*" However, there is a lack of using data from the experiment to substantiate this.

If they included something on the lines of "*We see from our data for red food dye, that lambda max is at -nm, and the highest absorbance is recorded for the most concentrated solution. Therefore, the most concentrated red solution, absorbs the green color component and transmits only the red color for us to see.*" The response would have been scientifically sound.

Similarly, for BQ2, "*What effect does concentration have on absorbance peaks?*" If the student had phrased their response as stated above, this question would have been answered along with the first BQ. However, even their actual response is barely sufficient or even understandable as 'evidence': "*Information from the two-calibration curve: absorbance and concentration have a linear relationship. Meaning you can measure the absorbance of an unknown dye and calculate the concentration of it.*" This response does not reference the attached graph (or graph data included in the report) and does not answer the actual question at all. Here, an acceptable response would have been on the lines of "*From the attached graphs for absorbance of various dilutions we can see that with higher concentration the absorbance recorded is also higher. Therefore, there is a linear relationship between concentration and absorbance.*" GTA Klaus' comments on this section are not very legible and from what we do read, they are not helpful to the student in addressing the large gaps in a decently written claims-evidence section.

The last paragraph of “possible sources of error” is intended to be part of the discussion section of this report. We are able to report that this student has not answered any of the analysis questions in a separate discussion, and only the “possible sources of error” is available for consideration as one response available for grading the analysis/discussion section. Even here, the student identifies some sources of error but fails to identify them as determinate or indeterminate error, or even the impact these errors have on the actual findings or calculated results in the experiment. The GTA’s comment “impact (-1)” is sufficient at this point to help the student identify what is missing but they could have included legible and constructive annotations to cause any tangible change in students’ writing. Klaus awarded three points out of a possible six points for the claims and evidence section. The score based on our analysis is 2 of 6 points. It is important to note here that although the numerical scores do not vary significantly, there striking differences in *why* these points were deducted, *what* the GTA was looking for/assessing and *which* key conceptual gaps were not at all addressed in GTA annotations.

3.4.3 Reflection section

As seen in Figure 26, the reflection section begins with Student’s response for the “extend” prompt is, “...foods could be liquified and their juices could be tested. Because orange peels contain food dye to give them a brighter color, if the juice was extracted from it, it could be potentially tested for the molarity of the food dye it contains.” For this response, the GTA rightly annotates the key missing portion of response with “why would we do this?” For the related reading and applications prompts, the student possibly tries to answer them simultaneously using the chlorophyll example: “Chlorophyll amounts in bodies of

water are measured using spectrophotometry. Measuring the amount of chlorophyll in the water helps scientists determine the quality of the water. In an experiment done by the royal society of chemistry [chemistry], the expected wavelength for lambda max of yellow dye was 380 nm-450nm. This held true with our collected data, with a lambda max of 406 nm.”

For further experimentation of food dyes, foods could be liquefied and their juices could be tested. Because orange peels contain food dye to give them a brighter color, if the juice was extracted from it, it could potentially be tested for the molarity of food dye it contains.

Chlorophyll amounts in bodies of water are measured using spectrophotometry. Measuring the amount of chlorophyll in the water helps scientist determine the quality of the water. In an experiment done by Royal Society of Chemistry, the expected wavelength for lambda max of yellow dye was 380 nm - 450 nm. This held true with our collected data, with a lambda max of 406 nm.

Classically, toxic heavy metals are used for this experiment. Due to heavy metals, negative effects on plants, animals, and most living things, it was not used in this spectrophotometric experiment. Instead of heavy metals, food dyes were tested, which are non-toxic and okay to dispose of in the drain along with water.

which green principle-1

whT would it do this?

contin on next our results -2

Figure 26 : GTA Klaus; reflection section from high-quality report

For the green chemistry response, the student says, “Classically, toxic heavy metals are used for this experiment. Due to heavy metals, negative effects on plants, animals, and most living things, it was not used in this spectrophotometric experiment. Instead of heavy

metals, food dyes were tested, which are non-toxic and okay to dispose of [off] in the drain along with water.” There is no specific response for the lecture topics connection and related reading prompts here. Table 15 summarizes the numerical scores from the GTA and researcher for this section.

Table 14: GTA Klaus; scores for Reflection section in high-quality report

<i>Prompt</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on qualitative analysis</i>
<i>Extend</i>	<i>2</i>	<i>1</i>	<i>1</i>
<i>Connect</i>	<i>2</i>	<i>2</i>	<i>0</i>
<i>Apply</i>	<i>2</i>	<i>2</i>	<i>1</i>
<i>Related reading</i>	<i>2</i>	<i>2</i>	<i>1</i>
<i>Green Chemistry</i>	<i>2</i>	<i>1</i>	<i>1</i>

3.4.4 Summary

Based on our qualitative examination, we observe that GTA Klaus picked out key responses where the student could improve their scientific writing skills. Again, there are few instances of observing legible, relevant, constructive, and effective feedback in the GTAs’ annotations and comments. Our qualitative analysis also estimated the scores that would be awarded, and as seen in the table below, the scores do have some discrepancies. From scanned reports we see that the GTA awarded this report a numerical score of 65/100. Back-reading score for this report was also recorded as 62/100, hence the overall grading scores agree. However, when we compare the areas where the GTA and the researcher deduct or award points, there is a stark difference in how the rubric was implemented. For GTA Klaus, since this was the first time teaching this course and we identified several areas where the GTA training could help him improve. Strong recommendations to

backread each week and consider the type of feedback they were giving students on written reports could have influenced GTA Klaus' approach to grading. Such constant monitoring and support would have helped both Klaus and his students tremendously.

Table 15: Total scores for GTA Klaus' high quality report

<i>Section</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on qualitative analysis</i>
<i>Introduction</i>	<i>10</i>	<i>8</i>	<i>7</i>
<i>Claims/Evidence</i>	<i>6</i>	<i>3</i>	<i>2</i>
<i>Reflection</i>	<i>12</i>	<i>8</i>	<i>4</i>

3.5 GTA Molly's Low Quality Student Report

3.5.1 Introduction section

Figure 27 is a screen capture of the introduction section of the Food dyes laboratory report from a student in Molly's class. The GTA rated this exemplar as a "low" quality report.

The student correctly defines spectrophotometry as "...measurement of how much something absorbs light by using the intensity of light as a ray of light goes through a solution." And includes the relevant formula used for this experiment, "The equation for this law is $A = \epsilon bc$, where A is the absorbance, ϵ is the molar absorptivity coefficient, b is the pathlength and c is the concentration."

Introduction

Most of the foods we eat today contain some sort of food dye, but does that make them safe? Although in this lab we are not determining the safety, we did determine the amount that is in the drinks provided. One method of quantitative analysis is spectrophotometry.

Spectrophotometry is the measurement of how much something absorbs light by using the intensity of light as a ray of light goes through the solution. An important law to know when performing this experiment is Beer-Lambert Law. The equation for this law is $A = \epsilon bc$, where A is the absorbance, ϵ is the molar absorptivity coefficient, b is the pathlength, and c is the concentration. To find the concentration of an unknown a standard curve is made by measuring the A of standard solutions of the unknown concentration. The first step of the procedure is to calculate the mass of your assigned food dye that is necessary to produce 500 mL of a solution of the target molarity. Then, make sure your hands are clean and dry before you place a square of weighing paper on the analytical balance and tare the balance, carefully weigh to the nearest 0.001g. Quantitatively transfer the food dye to the 500 mL volumetric flask, place the weighing paper back on the scale and record the mass. Fill the flask with deionized water to the line and mix well. Dribble a few drops of water on to the weighing paper where the food dye was, record your observations. Clean and dry four containers of at least 100 mL capacity and label them A, B, C, and D. Prepare standard solution A by pipetting the quantity of stock solution given in the lab manual. Dilute to volume and mix well. Transfer to container A. Repeat for solutions B, C, and D. Then set up the logger pro. First, calibrate the spectrometer. Fill a cuvette about three fourths of the way full with standard solution D, place cuvette in holder and click the green Collect button. Click the red Stop button to end data collection, make sure to Store Latest Run. Repeat for solutions C, B, and A. Once you have four absorbance spectra, change all the lines to black. To find the wavelength of maximum absorbance, select analyze then examine.

Concentration of Food Dyes in Commercial Drinks requires you to work with your partner and together as a table. Obtain 30 mL of a commercial drink that is the color of your food dye and determine whether or not the absorbance of the drink is in the range of colors you found earlier. Measure and record the absorbance of the drink. Finally, working as a table, use a 50 mL beaker to obtain the drink that contains both of the food dyes used. Repeat the process to determine the concentrations of both food dyes. Safety precautions that need to be observed are that we will be using lots of new equipment, wear gloves, all chemical waste can go down the drain, and do not eat or drink anything. The beginning questions we chose to investigate are “how can absorptivity/transmittance determine health risks?” and “how can you determine concentration of a mixed solution using Beer's Law?”

is long!

use paragraphs! -1

Figure 27: GTA Molly; Introduction section from low-quality report

The procedure write up is most certainly long but includes a very detailed description. However, the student could have written it as: “*First, we prepare a 500mL stock solution of the assigned food dye by weighing dry powder accurately and transferring precisely to a volumetric flask. Next, we prepare four 100mL standard solutions (A, B, C and D) from this stock solution using the formula $M_1V_1 = M_2V_2$. Then, calibrate the spectrophotometer*

and record the absorbance of the standard solutions to generate a calibration curve. Lastly, record the absorbance of the 'unknown' solution, diluting it if necessary and calculate the concentration of the food dye components using Beer-Lambert's Law." In this case, length of the write up would have matched the expectations and also covered all aspects of the experimental procedure.

Not specifying relevant green chemistry principles and simply stating "*...all chemical waste can go down the drain*" show the need for improvement in writing choices and "*...using lots of new equipment*" is not a safety statement. The student does list beginning questions that are relevant in the context of this experiment. However, the phrasing of the question "*how can absorptivity/transmittance determine health risks?*" could have been better, because the goal of the lab was to calculate the concentration of food dye and determine if it had potential health risks. The GTA awarded this introduction section with seven out of a possible ten points. Based on our qualitative analysis, a numerical score of 4 out of ten is more accurate here.

3.5.2 Claims and Evidence section

In response to the first beginning question (Fig 28), "*how can absorptivity/transmittance determine health risks?*" the student provides a generalized response, "*... the higher the absorbance, the more food dye that is present. When there is more food dye, there is a higher health risk because they are not natural and good for consumption in large amounts.*" Now this claim statement is not backed up by any evidence either from the calculated results of concentrations or report values of LD 50 which actually help us determine if a food dye concentration is a health risk or not. Simply stating a linear

relationship between absorbance and concentration also does not justify the claim as it is referenced to actual data and definitely does not answer the beginning question specifically.

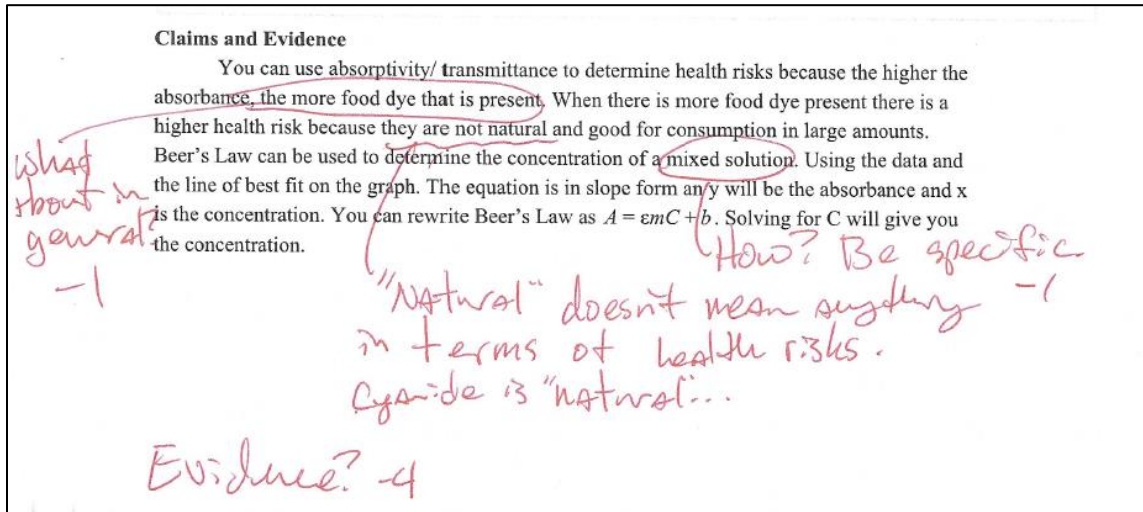


Figure 28: GTA Molly; Claims and evidence section from low-quality report

To answer the second beginning question, “*how can you determine the concentration of a mixed solution using Beer’s law?*” the student describes the calculation using Beer-Lambert’s law in words which is not accurate for “evidence.” Instead, if the student had written, “*With reference to the calculations shown in the experimental data section of this report, we determined the concentration of the yellow color to be $3.79 \times 10^{-5} M$ and blue color to be $3.46 \times 10^{-6} M$ in the unknown “green” solution. From calculations, we see that the LD 50 values for yellow no 5 is 12,750 mg/kg which corresponds to 41,753.4 glasses of the yellow drink. Therefore, we can say that the concentrations of food dye in the green drink were not a health risk.*” This would have been the expected and much more accurate scientific response.

The GTA awarded this section with zero out of a possible six points. Based on our qualitative analysis, a numerical score of zero is warranted here because the claim

statements do not make any sense and there is no evidence provided to support any of these statements sufficiently.

3.5.3 Reflection section

As described earlier, there are five primary prompts that are expected in the reflection section: Extend, connect, application, related reading, and green chemistry. In the low-quality report from Molly's class (Figure 29), we see very poorly framed responses overall, and entirely missing responses to the related reading and green chemistry prompts. The student's response to the extend prompt is reasonable enough, "*Using liquids that are more dense than water could affect the concentration and absorbance levels,*" makes for an exploration -worthy claim statement.

Reflection

To further this experiment we could add the variable of density to the calculations. Using liquids that are more dense than water could affect the concentration and the absorbance levels. In the real world, an example of spectrophotometric analysis can be found in the field of agriculture. Spectrophotometers can monitor the phosphorous and nitrogen levels in fertilizers to try to determine whether or not it is improving the soil. During my General Chemistry lecture we used Beer's law to solve for the molarity of a given substance.

based on what?
Be specific how this applies! -1

Additional reading? -4
Green Chem? -2

Figure 29: GTA Molly; Reflection section from low-quality report

Stating, "*Spectrophotometers can monitor the phosphorus and nitrogen levels in fertilizers to try to determine whether or not it is improving the soil.*" is a good example of a real-life

application. However, with no citation or elaboration of this example to confirm or dispute reported findings, it becomes an incomplete response.

Similarly, there is no context or connection provided to the experiment in the “connect to lecture topics response” where the student says, “*During my general chemistry lecture we used Beers Law to solve for the molarity of a given substance.*” Instead, this could have been rephrased as “*We have used the Beers law formula to solve for concentration in general chemistry lecture examples. This experiment helped us actually prepare solutions and use precise measurements as well as a standard calibration curve to calculate the concentration of the food dye using the same formula.*”

The GTA awarded this reflection section with five out of a possible 12 points. A numerical score of 4 out of 12 was awarded based on our qualitative analysis here.

3.5.4 Summary

For the low-quality report from Molly’s student, we see several notations throughout the report. These are mostly constructive and supported with underlined statements/ encircled words to help the student navigate various areas where improvement in writing and presentation is required. Based on our qualitative analysis, and back-reading data available to us the numerical scores for these sections in the low-quality report are tabulated below (Table 16). In comparison to annotations and markings made on the high-quality report we can see a significant rise in Molly’s multiple annotations to provide the student as much feedback as possible.

Table 16: Total scores for low-quality report; GTA Molly

<i>Section</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on qualitative analysis</i>
<i>Introduction</i>	<i>10</i>	<i>7</i>	<i>7</i>
<i>Claims/Evidence</i>	<i>6</i>	<i>0</i>	<i>2</i>
<i>Reflection</i>	<i>12</i>	<i>5</i>	<i>4</i>

Overall, the GTA awarded the low-quality report 67/100 and the back-read score for this report was 62/100 which is in somewhat reasonable agreement for the back-reading goals. However, two perspectives that should be considered arise at this stage: (a) are there lesser number of annotations on the high-quality report due to any inherent bias/assumptions on the GTAs part? (b) If the number of markings and annotations on low-quality reports remains the same (in Molly’s case it does) do students who receive lower scores on reports need a different format of feedback focused on helping them write better reports each week? We discuss these in more depth in the concluding section of this chapter.

3.6 GTA Klaus’ Low Quality Student Report

3.6.1 Introduction section

Figure 30 shows a screen capture of the introduction section of the low-quality report from GTA Klaus. The student begins with a clear purpose statement, “...*I used absorption spectroscopy to determine the concentration of food dyes in two commercial beverages: one with one dye and one with two dyes.*” Here the GTA circling several words does not convey any meaningful annotation to the student. Do they imply correct words or incorrect ones? Are the circled because the GTA marked down some points because of them? The

student is likely to be confused about the feedback on their graded report from the very beginning!

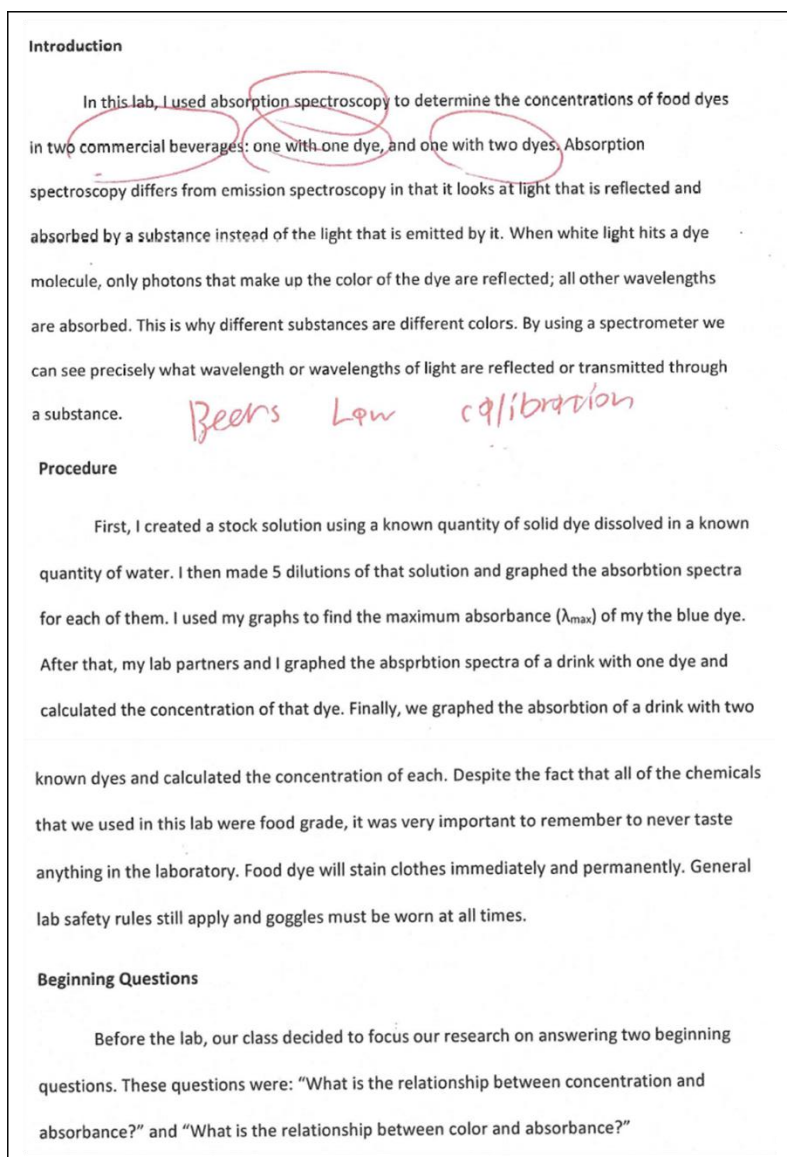


Figure 30: GTA Klaus; Introduction section from low-quality report

The student provides excellent detail on the phenomenon of spectroscopy and thus, a glimpse into their understanding of absorption versus emission spectroscopy, "...looks at light that is reflected and absorbed instead of light that is emitted from [the material]" which tell us that this student understood the previous experiment 'emission spectroscopy'

and uses that context to build the introduction to this experiment. Further, “*When white lights hits a dye molecule, only the photons that make up the color of the dye are reflected, all other wavelengths are absorbed*” shows us at minimum the ability to explain why dyes are colored. There is room for more accurate phrasing, e.g., *photons versus wavelengths used interchangeably*, but the attempt is commendable.

The experimental procedure is listed under a separate heading, and satisfactory in description of details. The beginning questions are didactic and oversimplified, “*what is the relationship between concentration (or color) and absorbance*” and would merit a one-sentence response based on how they are framed. These questions are not aimed at addressing the overall goal of this experiment, which is application of the Beer-Lambert Law to determine the concentration of a food dye in a commercial drink, and subsequently, any potential health risks. There are no marking or annotations in the procedure and beginning questions from the GTA. Based on our qualitative examination, numerical score awarded by the GTA is 6.5 and 7 out of a possible ten from the researcher. An estimated breakdown (since it is not very clear in the graded report from GTA) is included in Table 17 below.

Table 17: Scores for introduction section from GTA Klaus' low-quality report

<i>Introduction</i>	<i>Points Possible</i>	<i>GTA</i>	<i>Score based on Qualitative analysis</i>
<i>Theory /Purpose</i>	<i>3</i>	<i>1.5</i>	<i>2</i>
<i>Beginning Question(s)</i>	<i>2</i>	<i>2</i>	<i>1</i>
<i>Safety statement (s)</i>	<i>2</i>	<i>2</i>	<i>1</i>
<i>Experimental Procedure</i>	<i>3</i>	<i>1</i>	<i>3</i>

3.6.2 Claims and Evidence section

Examination of the claims and evidence in Figure 31 shows us that the student addresses their beginning questions specifically. In response to “*what is the relationship between concentration and absorbance?*” the student claims, “*absorbance and concentration are directly proportional for both dyes.*” which is an accurate statement.

However, we see a huge misconception in the continued response, “*...the red dye has significantly less absorbance per unit mass.*” and “*.... the slope of the blue calibration curve is steeper than that of the red curve absorbance increases more rapidly with concentration. Therefore, the blue dye is more potent than the red.*”

Now, the goal of the lab is not determining the “*potency*” of any colored dye. Therefore, these statements are inaccurate to begin with. Second, where is the student getting the information to infer that “*slope is steeper*”?

Claims and Evidence

Based on the evidence that I gathered in lab, absorbance and concentration are directly proportional for both dyes. The red dye has significantly less absorbance per unit mass. My graph of absorbance vs concentration for blue dye has a linear trend, represented by the equation $y = 113249x + 0.0442$ where y =absorbance and x =concentration (mol/L). The red dye also follows a linear trend with the equation $y = 22702x - 0.033$. Since the slope of the blue calibration curve (Absorbance vs Concentration) is steeper then that of the red curve, absorbance increases more rapidly with concentration. Therefore the blue dye is more potent than the red.

claims about data? -S

Figure 31: GTA Klaus; Claims and Evidence section from low-quality report

The student's two-component drink was a purple Kool-Aid i.e., a mixture red and blue dye, but that is not evident in the beginning questions stated earlier.

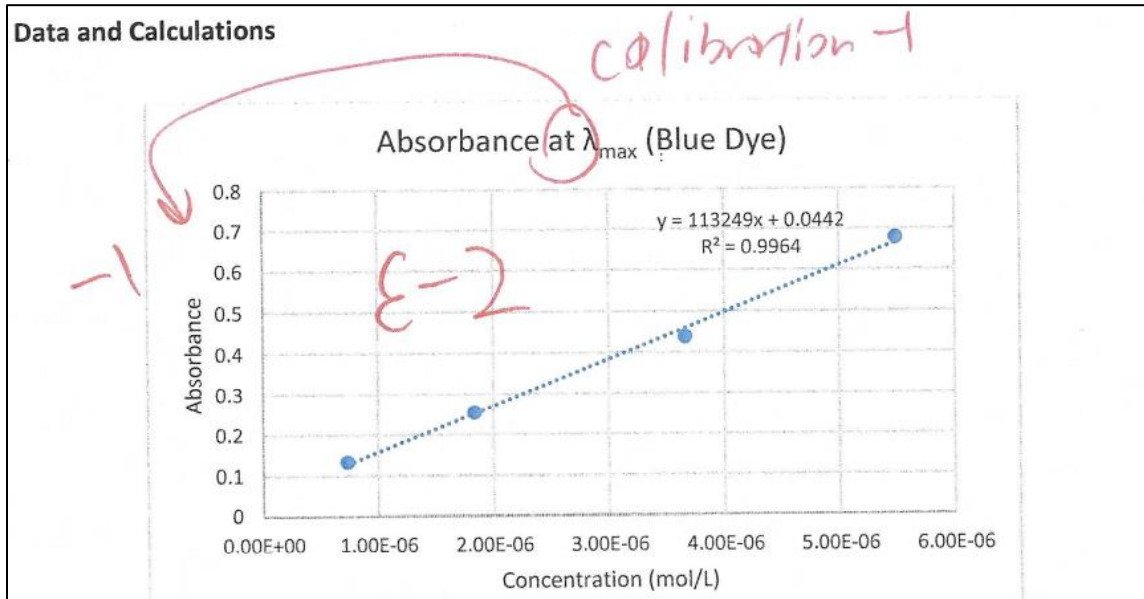


Figure 32: Graph output of calibration curve blue dye (GTA Klaus, low-quality report)

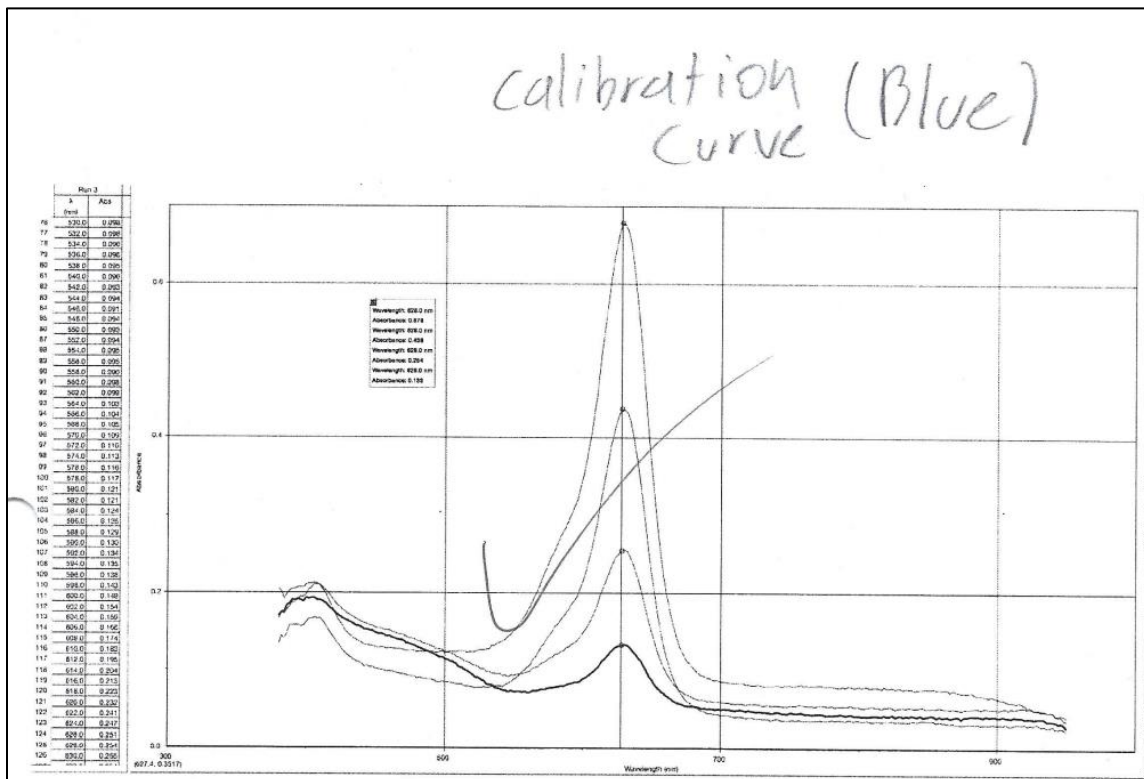


Figure 33: Graph output of calibration curve for blue dye (GTA Klaus, low-quality report)

Therefore, we see a claims-evidence paragraph laden with several misconceptions and haphazard evidence provided, but no annotations or markings from the GTA addressing these.

Based on the students framing of the question and their absorbance data or graphs (see Figures 32 and 33), an appropriate answer would have been, *“the relationship between concentration and absorbance is linear. As seen from the attached graph, as concentration of the sample increases, the intensity of the absorbance peak also increases.”* Therefore, the claims and evidence section would receive one out of a possible 6 points on the basis of our qualitative examination. The GTA also awarded this section 1 out of a possible six but for very different reasons based on their annotations.

3.6.3 Reflection section

A close analysis of Figure 34 shows another haphazardly organized response to the prompts of ‘extend, connect, apply, related reading and green chemistry.’ *“...the context of our general chemistry education was the application of Beer’s Law”* is the only sensible statement addressing the “connect to lecture concepts” prompt. The student is clearly inaccurate in stating *“Beer’s Law ($M_1V_1 = M_2V_2$)”* equates the relationships between molarities and volumes of two different concentrations of a substance.” $M_1V_1 = M_2V_2$ is the dilution equation, not Beer’s law.

The responses to application and related reading are possibly condensed together in *“...used in toxicology to determine how much of a substance is required to kill a person. Spectrophotometry is a commonly used technique in biochemistry where it is applied to*

determine the concentration of different proteins in a solution. Since the absorbance of a substance is tied to its molecular shape, and every protein has a unique shape, then each protein has a unique absorption.”

Reflection

The central purpose for this lab in the context of our general chemistry education was the application of Beer's Law. Beer's Law ($M_1V_1=M_2V_2$) equates the relationships between the molarities and volumes of two different concentrations of a substance. As demonstrated in question number 2, this can be used in toxicology to determine how much of a particular substance is required to kill a person. Spectrophotometry is a commonly used laboratory technique in biochemistry, where it is applied to determine the concentration of different proteins in a solution. Since the absorbance of a substance is tied to its molecular shape, and every protein has a unique shape, then each protein had a unique absorption.

Bibliography

"The weight of nations: an estimation of adult human biomass." BMC Public Health. BioMed Central, 18 June 2012. Web. 23 Feb. 2017.

This is not a related reading

Figure 34:GTA Klaus, Reflection section from low-quality report

Here, the student fails to provide good examples of applications and related literature that dispute or confirm their own findings in the laboratory. Also, they do not provide a citation for the “*since the absorbance is tied to molecular shape*” statement making it a not-so-well-phrased response.

The article citation provided is tied into the toxicology component where adult biomass is used to calculate LD50 and is a relevant inclusion with respect to the ‘related reading’ response. However, the GTA marks this as “... *not a related reading*” for no specific reason stated. This reflection section scored two out of a possible from the qualitative examination perspective and three out for 12 from the GTA.

3.6.4 Summary

For the low-quality report from GTA Klaus, we see a scarcity of notations throughout the report. These are mostly not legible, and seldom constructive feedback to help the student improve in their next laboratory report. Underlined statements/ encircled words alone do not convey to the student any mistakes or commendations on the grader’s part. This report received nearly similar scores based on our qualitative analysis, and back-reading data available to us.

However, there are so many points of disagreement as to why these points were taken off to begin with. The GTA also fails to address obvious misconceptions in the student’s written responses and leaves with a fairly strong assumption that the quality of this student future reports did not improve over time.

Table 18: Total scores for GTA Klaus' low-quality report

<i>Section</i>	<i>Points possible</i>	<i>GTA</i>	<i>Score based on qualitative analysis</i>
<i>Introduction</i>	<i>10</i>	<i>6.5</i>	<i>7</i>
<i>Claims/Evidence</i>	<i>6</i>	<i>1</i>	<i>1</i>
<i>Reflection</i>	<i>12</i>	<i>2</i>	<i>3</i>

Overall, the GTA awarded the low-quality report 59/100 and the back-read score for this report was 57/100 which is in reasonable agreement for the back-reading goals.

However, for GTA Klaus, we have observations that a high-quality report scored a 65/100 and a low-quality report was awarded 59/100. In this situation, the ‘range’ of scores differentiating high and low quality is not very wide and raises doubts about the reasoning on the GTA’s part in providing these reports with the quality they did state. Could it be that this GTA’s grading bias was so high that the quality of the work did not have discrete criteria for being high- or low-quality work? Did the GTA’s grading approach cause so much a negative impact on the students, that they simply stopped trying to write good reports or improve on their work because they did not receive adequate and effective feedback? The Discussion section of this chapter provides an insight into possible reasons for GTA Klaus’ grading approach being problematic and thus, creating room for improvement in the training-in-grading research.

3.7 Conceptual Analysis Rubric (CAR) And Statistical Data

To add a statistical element to our qualitative analysis, we qualitatively re-examined fifteen laboratory reports for the winter term from GTAs Molly and Klaus. Following the in-depth analysis presented in previous sections of this chapter, we used a uniquely designed and validated conceptual analysis rubric (Figure 35) to evaluate specific sections of thirty laboratory reports.

3.7.1 Objectivity in Using the CAR Rubric

Every effort was made to examine all back-read or qualitatively analyzed reports in this research project with impartial, unbiased, and an objective researcher’s lens. This included

using alphanumeric coding (e.g., M1, K1 etc.) and white-out markers to blind all identifying information about the student or the GTA during the analysis. Any anecdotal information known to the researcher, which may have had the potential to influence the analytic procedure were ignored completely.

Introduction section		
Expected (from student report guidelines)	HI /Yes -3; med (partial) -2; low -1 ; missing -0	Max possible
conceptually sound statements about the overall topic	introductory comments on expt.	2
	theory/phenomena	2
	formulae/relevant equations	1
purpose of experiment	clearly identified goal; conceptual soundness: does the statement reflect the purpose of the experiment correctly?	3
BQs	are the questions framed to be explorable?	2
brief description of procedure (3-4 sentences)	Can the experiment be replicated using ONLY the students description of procedure?	3
safety statements	appropriate stements/cautions	2
INTRODUCTION TOTAL		15
Claims/Evidence section		
Expected (from student report guidelines)	HI /Yes -3; med (partial) -2; low -1 ; missing -0	Max possible
Claims- answers to BQs	BQs specifically, clearly answered in the claims section	3
Claims-2	Extent of conceptual relevance of the claims w.r.t experiment	3
Evidence -1.	Does student address the "burden of proof" adequately? (3
Evidence -2	i.e., reasonably stated evidence + explained as needed (Evidence based on individual data/calculations)	3
CLAIMS/EVIDENCE TOTAL		12
Reflection section		
Expected (from student report guidelines)	HI /Yes -3; med (partial) -2; low -1 ; missing -0	Max possible
EXTEND	reasonable, accurate and meaningful extension	3
CONNECT	lecture course content related to expt: evaluate quality of response	3
APPLY	Applications of findings/concepts learned in expt: relevant applications from real life	3
	quality of explanation of relationship to experiment/results/conclusive findings	3
RELATED READING	Quality of response: another reliable source to confirm or dispute data/findings	3
GREEN CHEM	identification of relevant principles used in experiment	3
REFLECTION TOTAL		18

Figure 35: Conceptual analysis rubric used for qualitative examination of thirty selected reports from GTA Molly and Klaus

For example, known information about students copying down the TAs exact words/phrases during the laboratory session and/or using it in the laboratory report, which the researcher recognized during the analysis, as being the TAs own phrases/words, were disregarded while examining the exemplars during our inductive analysis. The “CAR rubric” developed from this approach was specifically developed to examine chemistry laboratory reports for evidence of students’ conceptual understanding in chemistry topics. This rubric was independently examined and implemented by two other researchers on exemplars from other TAs as a test of rubric validity (*does it answer the questions it asks?*) and rubric reliability (*do we get the same answers if different individuals apply the same rubric?*).

3.7.2 CAR Data Analysis and Graphs

Selected sections (introduction, claims evidence and reflection) from fifteen laboratory reports for each GTA were scored using the conceptual analysis rubric (CAR) and normalized to 100. These were then compared to corresponding GTA and back-read numerical scores from previously available data. Figure 36 shows a graph of normalized CAR scores versus normalized back-read (BR) scores. The numerical scores for GTA Molly (M in blue) and Klaus (K in orange) show us the distinction of Molly’s student reports into High, medium, and low quality very clearly. However, the same cannot be inferred for Klaus’ assessments (i.e., most report scores are clustered in the medium quality range) and reinforces the need for intensive back-reading or further training here.

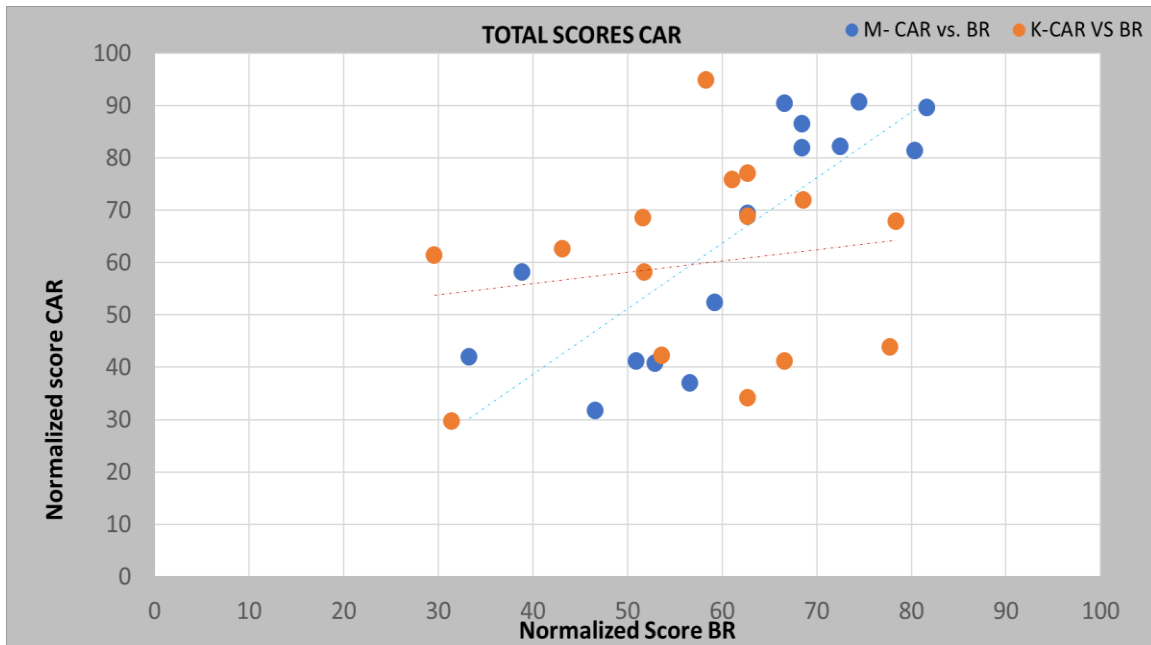


Figure 36: Plot of normalized CAR scores versus BR scores for GTAs Molly(blue) and Klaus (orange)

Similarly, we also examined the back-read scores for both GTAs against their students' laboratory final exam scores. If the GTA scoring was accurate, the laboratory final exam scores would reflect a pattern in reasonable agreement. That is to say, if the GTA assessed reports as average or below average, the laboratory final exam scores would also show that most students were below average in their performance. This would help us infer that the GTAs assessment approach was reliable, and the back-reading likely had a positive impact here. Figures 37 and 38 show us the boxplots: GTA's numerical scores (blue), researcher's back-read scores (orange) and the final exam scores (grey), all normalized to 100 for ease of comparison for fifteen laboratory reports from each GTA Molly and Klaus.

We infer that Molly's students were assessed more accurately with relative spread (Min. = 31.6, Max. = 90.6; and Range = 59.0) showing distinctions between high- and low-quality work ($mean = 64.9 \pm 5.7$). Molly's feedback can be considered to be more constructive and

effective based on the overall improvement in the average performance on the final exams ($mean = 68.2 \pm 3.58$).

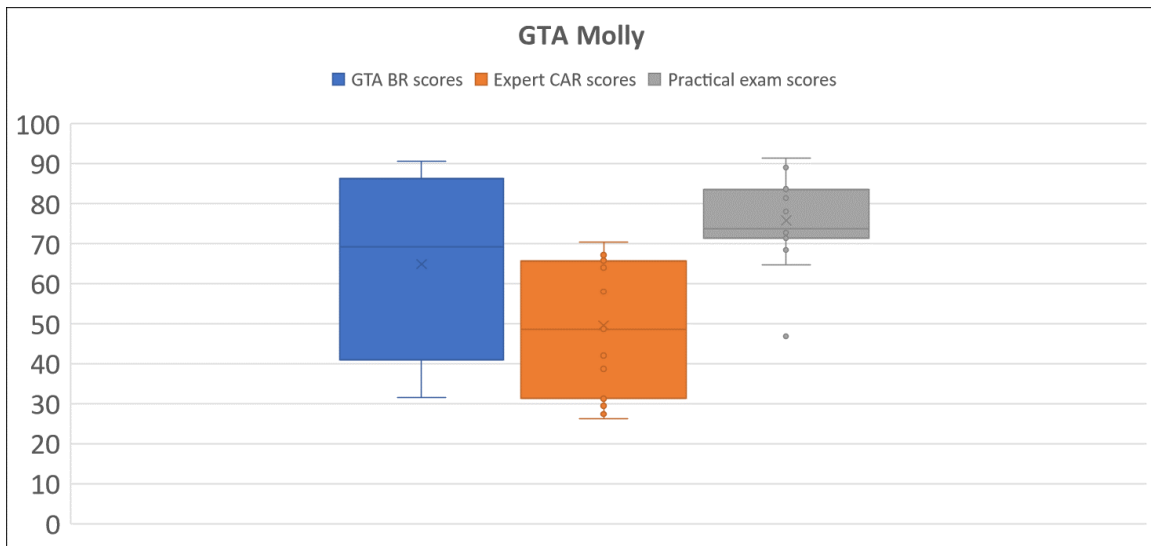


Figure 37: Box-whisker plots of CAR scores, BR scores and final exam scores for Molly's students ($n = 15$)

Klaus' grading however provides a picture that his majority of his students were producing similar quality of work over the entire term ($Min. = 29.5$, $Max. = 94.73$; and $Range = 65.2$, $mean = 61.0 \pm 4.51$). However, their final exam performance shows a remarkable improvement ($mean = 82.1 \pm 2.51$). This might be attributed to students' ability to understand, perform, and present their work in a scientific format was *actually* much better than what was being suggested by the GTAs assessment of their laboratory reports.

In such cases, critically examining GTA grading and providing more support and resources is key and should help the GTA as well students in having an overall positive experience in the course.

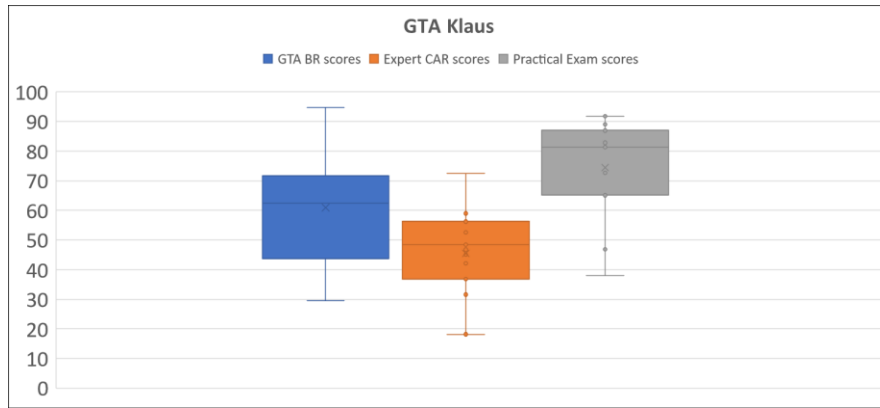


Figure 38: Box-whisker plots of CAR scores, BR scores and final exam scores for Klaus' students (n = 15)

Based on this analysis, we conclude that there is sufficient evidence to warrant improvements in the back-reading approach for training GTAs to (a) examine conceptual understanding in written responses (b) address any misconceptions identified and (c) exploring and assimilating examples of constructive and effective feedback to avoid less helpful comments or annotations on student reports during grading. The biggest challenge of course, is to be able to address and achieve all these goals with the limited amount of time, resources and GTA interest and availability to implement such changes in their teaching and grading.

3.8 Discussion

Qualitatively examining a few reports provides us with an insight on some differences in the grading approaches of GTAs Molly and Klaus. From Figures 21-34, we see how Molly approaches grading low-quality student reports with more focused comments, annotations

and marking. This is rightly so, as the students who need more support and feedback are likely scoring lower in the course. Her comments are clear, precise, and sprinkled with suggestions where possible, which tells us that this grading approach is favorable. We have anecdotal data from observing Molly's laboratory section and office hours that students were comfortable interacting with Molly and requesting clarifications, support, and guidance about their work. Although Molly had a fair share of grading-related grievances, we have significant evidence of this issue being much more intense for GTA Klaus' students. When students focus on asking for clarifications or revisions to their work on the previous assignments, it is difficult for them to prepare and dedicate time to making the upcoming assignment better. The excerpts from student correspondence about GTA Klaus' grading below demonstrate the struggles that occur when student do not receive constructive, timely, reliable, and effective feedback on their written reports. These excerpts were collated from personal communications to the Head GTA during conversations or correspondence pertaining to students' issues with graded reports and scores obtained.

When I got my first lab report for this term back from him, I noticed the low score and immediately felt determined to improve. I wanted to know exactly what I did wrong and how I could fix those problems in the future. I went into his office hours and got very unclear answers, and felt that his grading was inconsistent. A specific example of his inconsistent grading is with green chemistry principles: I never once specified the number of the green chemistry principle we followed in my reports, but I always specified exactly which principles we followed and how we followed them. But suddenly on the Kinetics Exploration report, he docked me points for not noting the exact number of the principle. With every report, I went into his office hours and asked questions about why I lost points in certain areas and how I could improve, but I got flat answers such as “it’s wrong, you are missing what the rubric asks for.” What’s frustrating for me about that answer is that it confirms what I already know, that I was missing something, but does not actually tell me how I can improve on that in future reports. Sometimes, [GTA] would explain further about why I lost points, but even his explanations did not make much sense. For example, in my TLC report, I provided a numerical summary in answering my beginning questions and claims which involved comparing my R_f values and referring to my data tables, but [GTA] took points off for not showing more than one R_f value calculation and for not providing a numerical summary in the manner he expected. I was completely unaware we needed to show the R_f value calculation multiple times nor that there is a specific manner for expressing numerical results [...]

Textbox 1: Excerpt from student 1 regarding GTA Klaus' grading of laboratory reports

My largest concern with [REDACTED] is not only his accuracy (using the rubric correctly) but also his consistency (grading the same way over time). Over the course of this term I have received an A, B, C, and D on labs I have done equal and identical work in. The hope for every teacher should be that your students evolves and improves, so that by the end of the term she/he has walked away with more knowledge than before. This did not seem to be his agenda. Though requirements change with every lab, a great portion of the points are on the same sections. I don't feel that going from an A to D is not logical but weighs on me and lowers my self esteem. I feel as if I could have done better, or done something different even though I do not understand what I should do to improve.

His judgemental attitude is extremely detrimental to the success of his students. It creates a toxic environment that discourages learning and mistakes. I felt extremely stressed and anxious every time I would have to confront him, and sometimes I avoided it altogether which kept me from earning points back. Every part of me wanted to avoid my work and avoid office hours, which was not how I felt last term in chemistry. Not only do we feel discouraged to seek help, we also feel less desire to work hard. *If my work is going to be devalued then what is the point of working hard?* I do not believe this attitude is what the chemistry department wants for their students.

Textbox 2: Excerpt from student 2 regarding GTA Klaus' grading of laboratory report

As far as grading and reports, my lab report grades were extremely varied. One week I would get a D and the next I would get a C and then back to a D. I got one A on a lab, and that was in between a D and a C. It was discouraging because I never saw any improvement on my lab reports. Compared to last term I saw a steady increase in my grade as I made changes and improved. It almost made me want to give up entirely on my reports. I worked extremely hard every week for hours and hours putting together my reports, only to receive a D almost every time. This was hard for me to accept, but I just kept pushing through and doing my best. Having [REDACTED] as my TA this term was really discouraging and I don't feel like I really learned or improved much on my lab report writing because of it. The only person that was really helpful when writing my reports was the head TA [REDACTED]

Textbox 3: Excerpt from student 3 about GTA Klaus' grading of laboratory reports

General issues with grading of laboratory reports (summary)

The grading of reports was a complete disaster. The number of points that were being taken off seemed unfair and not in junction with what was written in my lab reports. If I had the right idea but stated something wrong or didn't give the correct answer right away but lead up to it later in my lab, I would get all the points taken away. In some labs, I would use one wrong word in a sentence and get a point off. There were many points that half credit should have been given but instead I wasn't given any. In the pre-lab, I was docked points for parts that I didn't even know we had to include in our pre-lab or wasn't given enough time to finish. Since [redacted] had the rubric for the pre-lab before the start of class I believe it is only fair for him to let us know what is expected so we have a fair chance at getting all the points, instead of being blindsided with parts we didn't know we needed. Graphs were also a point of disagreement, where I would have a perfectly fine graph that had the same title as someone else or look the exact same, but I would get almost no credit for it. After getting back a couple lab reports back and not passing, I knew I had to step up and do better. I started spending more time, tried to put more information and just overall increase the quality of the lab. This did not increase my grade however. I stayed in the 50-70 range. So I began going to his lab hours and asking questions and trying to get a better understanding of what was expected and how to better answer questions that I needed to answer in my report. This still did not show an increase in grades for my reports. There was an overall inconsistency in grading, lack of direct feedback, messy handwriting making it difficult to understand comments and withholding of necessary information

Personally (confidence or motivation affected)

Last quarter, I got mid 50's on my first report. From there, I got better and better at writing the reports throughout the quarter which showed in my increasing of grades. My TA worked hard, answered questions and gave us all the information we needed to get the best grade possible. She did her part and I did mine. This quarter, I have spent more time on my labs then I did all last term but I am not getting anywhere near the same payoff. I understand that hard work does not equal good grades but I do believe that my work has increased throughout the semester which is not represented in the grades that I was getting. I felt unconfident in what I was writing and was frustrated that I was spending hours on something I knew I wasn't going to get a good grade on or even an improving grade. There was a lab write up that I did at 2'oclock in the morning in an hour and a half that I got a 52 on. I believe that this represented how well I wrote the lab and the time that I spent on it. However, the next lab, I spend over 4 hours on and went ot office hours to try and get a better understanding of the material and I received a 54. This crushed my motivation. Why work hard or try and talk to the TA if it has not effect on the grade I am going to be getting?

Academically (learning curve with no improvement, going from above average to a much lower performance level, impact on scholarship/ financial aid due to a low grade)

This lab is crushing me. I had a B+ last term and this term I am going to be lucky to end up with a C+. I need to keep a 3.00 for my scholarship and I need a B- for graduate school. With this C+ it makes it more difficult to keep that 3.00GPA since lab is not one of my harder classes. I also believe that I increased throughout the term with no grade acknowledgment. I think that there is an issue when the lowest grade in the class in a 30 and the highest is only a 73 with an average of a 61. I do not believe that nobody in class deserves at least a B for the work that they have done this quarter.

Textbox 4: Excerpt from student 4 regarding GTA Klaus' grading of laboratory reports

From the student comments reproduced in textboxes 1-4 we can see an example of how unreliable and inaccurate grading impacts students' ability to continue focusing on their academic performance. Such a negative impact on student morale is also detrimental a departmental program level since it has tangible effects such as poor student attrition. This feedback also drives home the necessity to monitor GTAs grading using back-reading to ensure reliable and accurate grading provided with constructive feedback.

3.9 Conclusion

In summary, qualitative examination of two high-quality and two low-quality laboratory reports (as designated by GTAs) shows us several areas of discrepancies that were not necessarily addressed by back-reading process alone.

Molly, A GTA who participated in the backreading program and self-reported its positive impact on her grading practices provided accurate, constructive, and effective feedback to her students and maintained a positive growth environment overall for development of good scientific writing skills.

Klaus on the hand, demonstrated the use of aggressive and at times, non-legible feedback that increasingly contributed to the students' frustration at understanding how and where to improve in their writing, and resulted in an overall negative experience in the general chemistry course.

Furthermore, using a conceptual analysis rubric specifically designed to examine the "conceptual mastery" in student responses shows that both GTAs may have graded their students either too harshly or leniently, and with time, lost touch with basics of grading as

provided in back-reading training. Also, since we used the CAR rubric to grade for conceptual understanding and not just the presence or absence of responses to specific prompts, we also identified several gaps in the back-reading adapted approach that need to be addressed beginning with rubric design.

CHAPTER IV: GROWTH MODELS AS A UNIQUE APPROACH FOR EVALUATION OF GRADING TRAINING

4.1 Chapter Abstract

This chapter covers the first-known report of a growth model approach to data from a TG A professional development program focused on grading chemistry laboratory reports. Design and implementation of a back-reading protocol was qualitatively and quantitatively analyzed to elicit implications of such training as described earlier. Training programs have historically been evaluated on the basis of qualitative feedback from participants and sometimes, by simple inductive analysis of data from videotapes, focus groups or semi-structured interviews. The transience of such data being the cause of unreliable assessments of training program persistence and success is hypothesized. A theoretically robust approach using time, and time-invariant predictor variables such as gender and experience to generate growth trajectories for individuals and groups is explored and interpreted to demonstrate its potential.

4.2 Introduction

4.2.1 Literature Review: Evaluation of Teaching Assistant Training Programs

Departmental and campus wide GTA training programs are almost always followed up with a participant assessment/ evaluation protocol, such as paper or online questionnaires. The intent of these surveys is to collect participant views on and responses to questions

about existing components of the training program. Additionally, they provide an opportunity for participants to identify any gaps/ shortcomings they may have experienced individually.

Program assessment protocols establish whether the training program serves the overall goals of the department, such as (a) to prepare teaching assistants adequately for their assignment, (b) provision of training, teaching material, and (c) hands-on experiences/ training scenarios which [provide feedback on GTAs' teaching skills. Departments and universities utilize such training program evaluations to make decisions ranging from sustainability of current training resources, future training requirements, personnel and sometimes, even budgets.

The number and nature of GTA training programs has increased rapidly over the last 50 years. Evaluation of GTA training programs has largely focused on the *changes* in the GTAs' individual approach to teaching. These are denoted in a variety of ways in literature: such as "Interactional"¹⁰⁹; or "behavioral" changes and sometimes, "attitude analysis".²⁹,

35, 110

Crooks (1980)⁵¹ is probably one of the earliest reported literature examining the training program established by the Office of Instructional Resources (OIR) at the University of Illinois, Urbana-Champaign. The paper discusses *three* different surveys – addressed to departments, teaching/supervising faculty and participant GTAs for the training program. A detailed discussion on the reactions of participants, faculty and departments using actual response items is provided for each survey. One of the features of this training program was to videotape the GTAs while they were teaching followed by a one-on-one with a program staff member, providing constructive feedback to the GTA. These responses are

particularly significant, considering the amount of time and effort it would have taken in a time when technology was not as advanced as it is today.

Carrol's (1977) research with psychology GTAs is one of the earliest demonstrating a true comparison experiment with GTA training¹⁰⁹. The treatment group of GTAs were predicted to (a) use teaching objectives more actively, (b) show student-centered teaching approaches rather than teacher-centered ones, (c) ask deeper cognitive questions compared to control group, (d) have more student Talk 'levels' (term used by author) in the class instead of single-instructor delivery, and (e) obtain higher evaluations than the control group. All these hypotheses were observed and proven to be significantly true for the GTAs that underwent GTA training.

Pescosolido and Milikie's (1995)¹¹¹ summary presents a good example of the gradual growth of several GTA teaching seminar programs offered in sociology departments at U.S. and Canadian universities. The authors claim to be impressed with one of the outcomes: that most institutions offered a term/semester long training program for their GTAs/ graduate instructors. In 2015, Blouin and Moss¹¹² revisited the topic and found that formal teacher /GTA training efforts have increased across the board between 1995 and 2015. Some of the quotable statistics are an increase from 48 to 55 percent for GTA training and more significantly for graduate instructors training (stand-alone courses) from 55 to 68 percent.

An analysis of the components and long-term success of established GTA training programs is by Thornburg, Wood and Davis(2000)¹¹³. This article identifies key elements required for the 'survival and maturing' of a training program. The factors identified are further supported with a detail-rich case study of the success of the GTA training program

in the department of chemistry at UC Davis. A parallel article (2000)¹¹⁴ examines the key aspects on which the success of a formal GTA training program is based: faculty involvement; participant contribution to program design/content; range of teaching experience among participants; accommodation of time constraints (participants' graduate work versus time devoted to the program) and Tangible recognition (participants were able to utilize program-related projects or data for publications , conferences etc.; as well as assured completion certificates and a teaching program awards ceremony. This data was collected at UC Davis for five years (1995-2000).

Emerson et.al (1996)¹¹⁵ present the development and success of the “Penn State course in college teaching”. This paper is a remarkable example of how a teaching seminar-style training is not necessarily beneficial to apprentice GTAs alone. Their findings indicate that faculty from various levels of experience, novices to tenured, also participated in the course which combined teaching practicum with pedagogical theory and reflective thinking approaches to their own teaching.

A paper reporting the results for a National Survey of Teaching Assistants¹¹ provides information about TAs from across eight diverse universities in the U.S. The results are inferences from the analysis of survey responses based on five broad question groups: (i)Demographics, (ii)teaching responsibilities and (iii)preparation resources, and (iv)questions specific to international TAs followed by (v) recommendations for training program supervisors.

Rothfuss and Gray (1989), also report on the use and training of GTAs in Higher Education Universities¹¹⁶. The survey responses are grouped into two categories: GTA-level and administrative-level (department or campus-wide). Based on outcome variables such as

GTAs' satisfaction with training, evaluation of GTAs' teaching efficacy, survey responses of (a) faculty involved with training, and (b) administrative personnel overseeing training, the authors conclude that most respondents are committed to continued training and improvement in training programs.

4.2.2 Training Programs Focused on International TAs (ITAs)

Universities may consider students with non-immigrant visas who hold teaching assistantships as “international TAs” regardless of their native language³⁹. There are instances of considering students who are non-native speakers of the English language being grouped as ITAs as well those students who completed their undergraduate studies outside of the United States. Irrespective of how ITAs are defined, the issues and factors influencing their experiences are much more complex, pedagogically and culturally^{42, 44, 117}, requiring different perspectives in the design, implementation and continuation of training programs¹¹⁷. Considering the rapid growth in college enrolments (and corresponding hiring of more GTAs) in the decades from 1980-2000⁴⁶, training programs specifically targeting the orientation and evaluation of international TAs (ITAs) are more widely reported.^{26, 36, 56, 118, 119} Studies also address issues ITAs face in acclimatizing to new culture^{119, 120} and academic classroom settings^{37, 43, 44}. Coupled with the responsibilities that ITAs handle as part of their teaching assistantships, the factors of interest vary from language fluency or communication skills⁴¹ to classroom behavior management, it is evident that the scope ITA training program evaluation studies is significant.^{30, 43, 45, 47, 49, 50, 118, 120, 121}. Extensive literature is available identifying and illustrating various techniques for working with ITAs. The most common issue with ITAs are language barriers, fluency,

and communication skills. The second most researched issue is Teacher Anxiety^{30, 34, 36, 45, 89, 122} which affects the ability to independently conduct/address a classroom of students when the GTA/ITA^{42, 44} has had no prior stand-alone teaching experience.

SPEAK-TEACH (1988)¹²³ which describes the “Taped Evaluation of Assistants Classroom Handling” method of training ITAS to lead independent classroom/recitation sections has now become an established program at Iowa State University. This method of ITA training has even undergone updates to the extent of applying a systems-approach to renewing the curriculum for ITA training.¹²⁴ As expected, most of these programs are evaluated based on the survey responses or interviews and observations of ITAs.

4.2.3 Training Programs for TAs In Inquiry Teaching

The idea behind inquiry based learning is teaching students how to think, as opposed to, what to think¹²⁵. For example, when building electronic circuits, students must be able to think for themselves about how to design and troubleshoot. As a result, laboratory demonstrators should not help students by giving them the answer or doing the experiment themselves; instead, they need to question the students strategically so they can procure their own answer or process^{70, 126}. It has also been found that inquiry-based training improves the effectiveness of demonstrators^{22, 23, 127}.

4.2.4 Training Programs Focused on Student Outcomes

GTA training program evaluations with more traction examine the impact of GTA training on student success or behaviors. These outcomes have matured over time, from examining simplistic trends like student course evaluations, to more complex ones such as attitudes

toward discipline-based learning, and student attrition in specific programs or courses. As of now the current interest is largely focused on student development of guided-inquiry abilities.^{16, 128}

The various GTA training programs that have been studied and reported are likely to be (a) department-specific (b) include a small or moderate number of participants and (c) provide a newly developed method of exploring variables of interest such as pre-post assessment protocol or a classroom /laboratory observation protocol. Most of them are evaluated on the basis of qualitative feedback from the participants collected as surveys, interview responses or analysis of reflection or videotaped activities during training.

Survey responses consist of individual feedback which are susceptible to contextual biases, based on the teaching assistants' experiences *at the time*. This makes the need for a reliable evaluation method more crucial than ever.

Classroom or laboratory observation protocols are as expandable as needed, limited only by the variety and complexity of outcome variables defined by the researchers. Although such protocols analyze characteristics of classroom teaching and teacher-belief systems that influence instructional efficacy, they are essentially simple cause-and-effect systems: focused on identifying on-going teaching practices and addressing gaps and/or providing supportive interventions to address any shortcomings.

Since undergraduate programs represent an ever-growing and thriving section of university population, the necessity for well-prepared teaching assistants cannot be emphasized enough.

4.3 Research Context

4.3.1 The Issue with Existing Program Evaluation Techniques Measures, And Approaches

As Pelton ⁸⁹ says, “As more programs aim to train graduate students in the role of both researcher and teacher, assessments strategies are needed that will ensure that these courses effectively prepare graduate students for their new status in the classroom. The ultimate measure of success would be improvements in student learning as a direct result of enhanced teaching effectiveness.” This is very much in alignment with Park’s view “Training may be defined as bringing the teaching assistant to an agreed standard of proficiency by practice and instruction” ³³ and results in a chicken-and-egg loop wherein TA training influences student success and student performance helps improve and address TA training needs. Figure 39 shows a summary histogram of literature sources reporting on training program evaluations. The criteria for differentiating these programs are the specific modes of data collection used for training program evaluation. Eight data collection modes were identified and coded as used (yes) or not used (no) for literature sources between 1975-2019. These modes are (a) pre/post survey or questionnaire (b) student evaluations of teaching (c) classroom or laboratory observations (d) TA interviews/ focus groups (e) Student interviews / focus groups (f) Faculty /trainer interview (g) Other sources – majority reported recorded digital videotapes as sources for either self-evaluation by GTAs along with the researcher or independent evaluation and coding for specific themes.

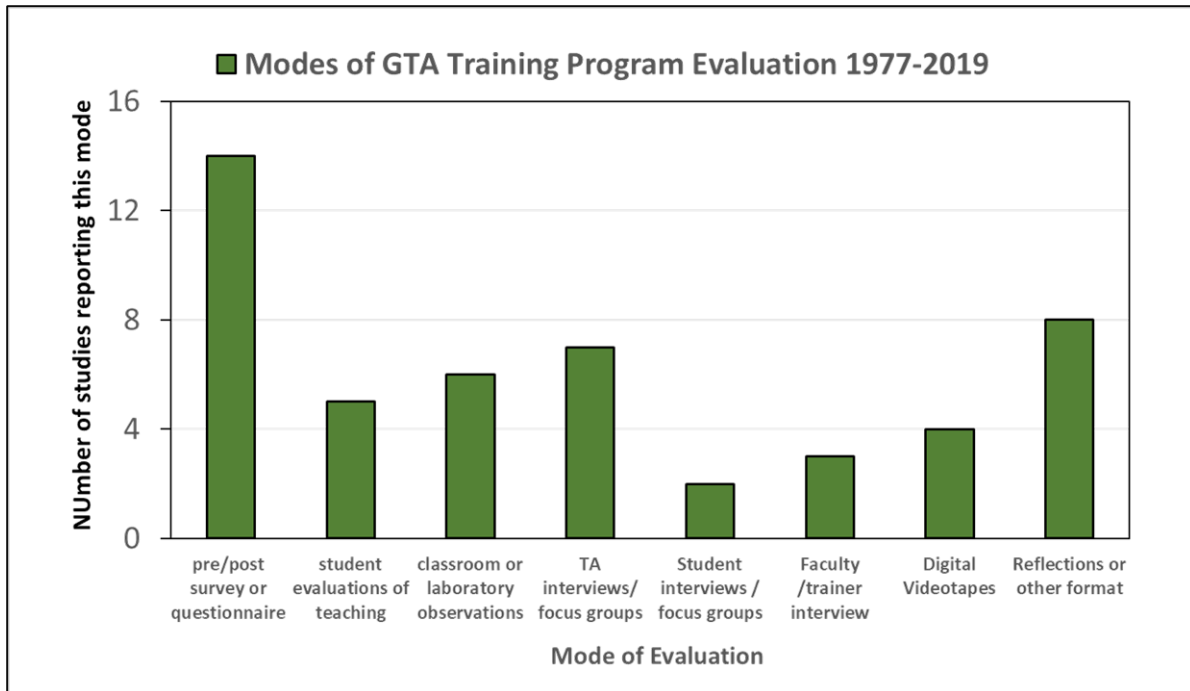


Figure 39: Histogram showing different modes of data collection used for training program evaluations reported in literature.

By and large, surveys/ questionnaires and videotapes are the most consistently used data sources for evaluation of TA training programs during this period. Although these modes are excellent data sources for evaluating TA training programs and may be accurate at the time of collection, the effect of training is not quantifiable or robust like numerical measures. For example, A GTA may self-report the effectiveness and extreme agreement with the usefulness of the TA training program in their first-year as a graduate student and a novice TA. However, their views or beliefs may change over time, and be invalid a few years later, and therefore utilizing a data source that has a short “shelf-life” may prove unreliable over time. However, recording a measurable outcome such as a difference of scores between GTA and expert has a finite error associated with it, and can be verified by

repeating these measures over and over, if necessary. Thus, a quantifying or analytic approach devoid of qualitative bias or transience is justified.

4.3.2 Justification of Interest in Growth Models for Assessing TA Training Programs

Studying the nature of change with time is integral to educational and behavioral research, almost as common as the study of chemical reaction kinetics in chemistry. A necessity of any such study is a model, one that can guide the inquiry process, and result in a deeper understanding of the phenomenon of interest. According to Raudenbush and Bryk (2002), “studies of change typically use instruments that were developed to discriminate among individuals [or observations] at a fixed point in time. Many studies collect data at only two [pre, post] points in time. The practice of scaling instruments [or measures] to have a constant variance over time can be fatal to studying change and determinants of change.”

A major objective of developmental science is to describe how, when and why individuals’ behaviors change over time¹²⁹. A growth model is a longitudinal analysis of growth (or change in measured outcomes) over time.

Contemporary use of the term ‘growth curve model’¹³⁰ typically refers to statistical methods that allow for the estimation of inter-individual variability (differences between individuals) in intra-individual patterns of change (changes within an individual) over time. One of the simplest explanations for the ubiquity of growth models is that this model accounts for the changes in variance as well as changes in variables of interest.

In the present study, our interest is (1) to achieve accuracy and consistency in grading general chemistry laboratory reports and exams and (2) provide professional development

training for GTAs to achieve that goal. Research and teaching experiences are mandatory requirements in the graduate level chemistry curriculum. This indicates that the majority of graduate level programs have a visionary interest in training each graduate student to become a meticulous researcher as well as developing them as a competent teacher. The terms ‘training’ and ‘development’ imply the need for monitoring change (1) with training and (2) over time. The milestones (coursework, qualifying exams, candidacy exams etc.) and time required for completion of a graduate PhD program are strong indicators it is not intended to be a “graduate school treatment” with evidence of change using two reference points: before and after graduate school.

The objective of growth curve modelling is to describe a set of time-ordered, within-person observations using only a few parameters. Therefore, there is potential in using this approach for tracking the efficacy of training GTAs in grading laboratory reports and a promising analytic approach to add to existing program assessment techniques.

Therefore, we attempt to answer the following research question based on our growth model approach.

- How do growth trajectories for individual GTAs or groups of GTAs progress with time?
- Which factor (or predictor variables) can best explain any significant variations in the grading-related growth trajectories of GTAs in the back reading study?

4.3.3 Literature Review on The Use of Growth Models in Academic or Intervention Studies

Growth models have been extensively used in early childhood education and particularly language fluency studies.¹³¹⁻¹³⁴ The attraction of a longitudinal approach to examining data lies in the ability to utilize it to simulate predictions of future trends, and or use predictors to explain certain trajectories in the observed trends. A great example of a linear growth model is the study of children's reading growth during the first two years of school. They report that children average 1.67 points of reading growth per month. Student-level variables such as socioeconomic status were significant for reading growth. Similarly, school-level variables show that the socioeconomic status of school clientele and not instructional differences or school resource allocations, explain the variation in results. These findings provide evidence of the importance of early learning intervention programs for lower socioeconomic status children. A key piece of this study is the use of linear piecewise growth model to explain the transitions in reading growth when school is in session and during summer break¹³¹.

Another example of using a quadratic growth model is the characterization of vocabulary growth rates for children below age 3. This study finds vocabulary growth to best an exponential model (represented by a quadratic equation)¹³⁵

A much more complex analysis is reported by King et al.¹³² using a technique known as latent growth curve analysis (LCGA). Compared to hierarchical linear modeling (HLM) discussed in this chapter, LGCA is more flexible and can provide unique curves for each individual or groups of individuals, represented as *deviations from the average function*, in addition to testing hypothesis about trajectories of interest.

In their study examining 2618 students' self-concepts in English and Math, the researchers used an unconditional growth model (time as a predictor) to examine any systematic variation in the self-concept measures. By adding predictors such as gender and school-band (high, medium, or low ability schools in Hong Kong), they were able to determine that girls showed higher self-concepts in English, while boys demonstrated these in Math. Also, students from low-ability schools had lower self-concepts relative to those in higher-ability schools based on the model results. There are numerous such studies utilizing growth models or longitudinal analysis to examine the effect of interventions¹³⁶ or new programs on the target audience(s), changes in self-reported measures of success, anxiety, beliefs, and behavior. (Owens & Shaw, 2003; ^{133, 137}. The fields implementing growth model approaches range from ecological research¹³⁸ to oncology,^{129, 139, 140} psychometric treatment¹⁴⁰, reading interventions^{131, 133, 141} and even large scale studies such as traffic management¹⁴² and tourism¹⁴³.

Considering the scope and interest in using longitudinal analysis for studies from such a variety of areas, it makes sense to explore whether it has been implemented in the field STEM education or more specifically, chemistry education. There is only one documented study thus far¹⁴⁴, examining students' cognitive and affective expectations and experiences within the context of performing experiments in their chemistry laboratory courses. These measures were collected using the Meaningful Learning in the Laboratory Instrument (MLLI) administered to general and organic chemistry students from 15 colleges and universities across the United States. However, the analysis in this paper is neither that of a hierarchical or latent growth model, therefore, making the present study a first unique

report of utilizing a growth model approach to evaluate the professional training in grading for Chemistry GTAs.

4.4 Growth Model Theory

4.4.1 Explanation of A Growth Model Using Individual GTA Grading Data

Our GTA training in grading focuses on achieving agreement and consistency in scoring laboratory reports. Therefore, our outcome variable is the difference in the scores (Δ score OR DIFSCR) provided by the GTA and the expert for the same laboratory report.

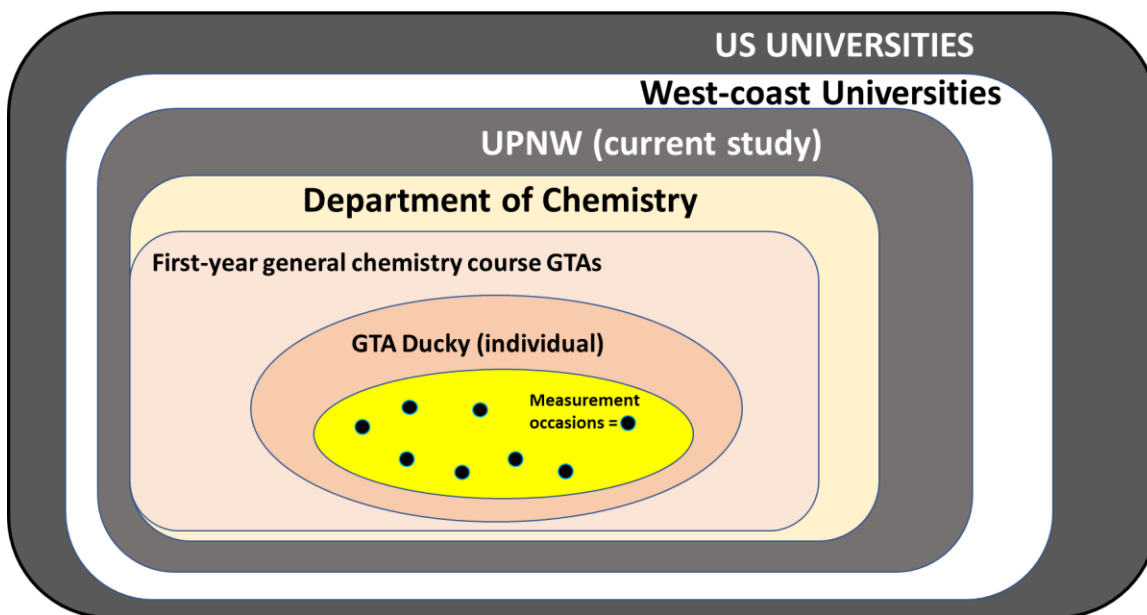


Figure 40: Measurements recorded for an individual at various points in time

An individual is a part of many nested groups, as can be seen in Figure 40. Therefore, when we collect measures of a variable of interest for multiple individuals, we can also use the various “nesting” markers to group the data and build complex layers into our approach.

Consider a set of outcomes measured for an individual GTA, Ducky, at different time points: $t = 0, 1, 2, 3, \dots$ and so on (Figure 40) ⁴.

Hypothetically assuming an increasing trend in the difference in scores for GTA Ducky and the expert the outcome variable (abbreviated as (DIFSCR as the term progresses, a graph of the outcome variable versus time will probably resemble Figure 41⁵. This is the most basic inference from a visual examination of a growth trajectory (illustrative only, not actual data).

Analysis of such an increasing trend over time (simple linear model fit) would mathematically be represented as:

$$\text{Dependent Variable (DIFSCR)} = (\text{slope}) * (\text{time}) + \text{intercept}$$

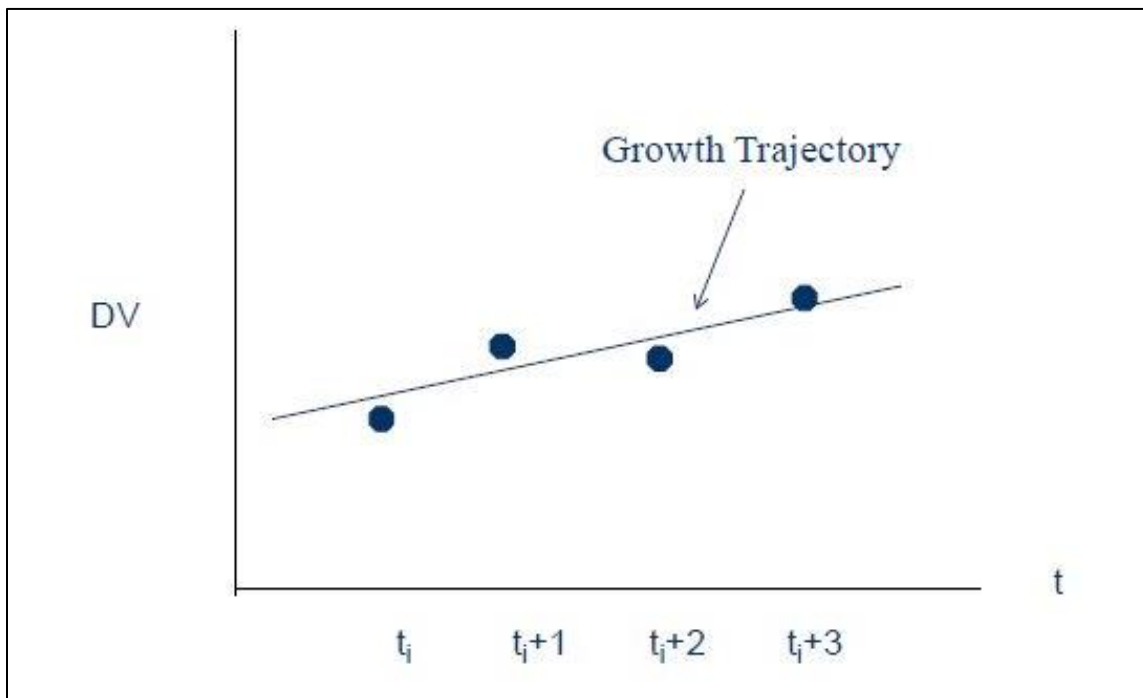


Figure 41: Graph of dependent variable vs. time, the trendline represents the growth trajectory for this individual

⁴IMAGE SOURCE (34) EDLD 650 (spring 2016) Course slides from

⁵IMAGE SOURCE (35) EDLD 650 (spring 2016) Course slides

We now have two features of interest in Ducky’s growth trajectory graph: the intercept and the slope. The intercept represents the predicted value of the individual’s outcome status at the beginning of measurement, also known as initial status. The slope represents the average amount of change in the outcome per unit change in time. This growth model is often referred to as a level-one growth model and requires only an identification variable, clock variable, and outcome variable as input data. This model is mathematically represented as shown in Figure 42⁶.

Level-1 Within-Person Model:

$$Y_{ti} = \pi_{0i} + \pi_{1i}(Time_{ti}) + e_{ti}$$

Y_{ti} is the outcome at time t for individual i ;
 π_{0i} is the status of individual i when time = 0
 π_{1i} is the growth rate for individual i over the data collection period (the expected change during a fixed unit of time)
 e_{ti} is a error term representing variation from the latent growth trajectory at time t

Figure 42: Statistical representation and terms in an unconditional, level- 1growth model

4.4.2 Examining Multiple Individuals in A Group for Overall Growth Trends

Our back-reading data is a measure of discrepancies (DIFSCR) between the GTAs scoring and an expert’s scoring and used a strictly defined range of ± 1 unit of score difference among raters. An ideal result for each back-reading measurement, after successful GTA training, would be 100% agreement between the GTA and expert. In other words, for each report which was graded, there would be no difference between the two raters’ scores.

⁶ EDLD 650 (spring 2016) Course slide from Instructor

$$[(\text{GTA score}) \text{ minus } (\text{expert score})] = \text{zero}$$

Understandably, this is an extremely ideal scenario with relatively low probability.

For data pertinent to multiple individuals on multiple occasions (i.e., repeated measures) variation in the outcome variable is expected. The point of using a growth model is to *partition* this variation. That is to say, to explore whether there is a systematic variation in the outcome variable (DIFSCR), and whether it is significant enough to merit further exploration. Additionally, if there is detectable significant systematic variation, we can identify specific predictor variables that help us explain the cause of such variation. For example, participation in back-reading function as predictor to explain the shift in intercept or rate of change. If it is systematic, then the positive impact of the back-reading treatment can be interpreted by detailed analysis of the growth model output. The other component is random variation (such as random error) which inherently exists with measurements as with any dataset.

Consider GTA Ducky's peers in the general chemistry course, Molly, Milo, Klaus, and May. Please refer to graphs and examples provided in the data and results section of this report for more details. As a group, we now have a few variables that could influence the trends we see:

Are the GTAs novices, or experienced? Have they undergone training in grading or not participated at all? Do male GTAs' growth trajectories differ significantly from those of female GTA's?

The first step of examining longitudinal data is to build an unconditional model. That is, to simply examine only the outcome variable for any significant variances in slope and initial

status. By analyzing any systematic variation in *just* the outcome variable, we can determine if the next are worth pursuing. If significant variance is detected in the outcome variable, the addition of a level-one variable (e.g., time) results in a level-one growth model. This model provides information about the effect of time on the outcome variable. If the systematic variation can be explained by the addition of the time variable, then the slope and intercepts values would be statistically significant. That is to say, some of the variance in the outcome variable is explained by the addition of time or a modified form such as second order (time squared) or third order derivative (time cubed) of time.

Lastly, using relevant coded data such as participation status in the training program, gender, discipline or prior experience status as predictors provides us with a conditional model. A level-two growth model requires an additional layer of information apart from ID, time, and DIFSCR (or outcome variable). A grouping variable is the key factor for a level-two growth model. Are we grouping GTAs as experienced or novices? Participants or non-participants? Male or female? A binary coding system is often used for the level-two variable (male-female; present-absent etc.) However, the level-two variable is not limited to just binary data-coding. There can be as many “boxes” in the level-two variable as required. The essential aspect is that the ID variable linking the outcome variable to the “boxes” is consistent. This ensures level-one and level two growth model data are correctly tied-in so that the comparison is accurate. This nesting variable approach (using coded predictors in level two models) can be used for as many levels as required and generates what is termed “hierarchical multilevel modeling.” The effects of nested predictor variables (such as participants versus non-participants; male vs female) provides information about the changes over time including the effect of the level two variables (participation, gender

etc.) and attempts to explain any systematic variation or pattern changes due to these predictors. Thus, to examine the effect of a training program or intervention, a growth model with predictors can be an extremely powerful tool.

Level-2 Between-Person Model:

$$\pi_{0i} = \beta_{00} + \beta_{01}(X_i) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(X_i) + r_{1i}$$

$$\text{Var}(r_{0i}) = \tau_{00} \qquad \text{Var}(r_{1i}) = \tau_{11}$$

$$\text{Cov}(r_{0i}, r_{1i}) = \tau_{01}$$

β_{00} represents the ‘conditional’ mean status of individuals;
 β_{10} represents the ‘conditional’ mean growth rate of individuals;
 X_i are individual-level predictor variables
 β_{01} represents the average “effect” of X_i on individual status
 β_{11} represents the average “effect” of X_i on individual growth
 r_{0i} capture the ‘conditional’ differences between an individual’s status and mean status
 r_{1i} capture the ‘conditional’ differences between an individual’s growth rate and mean growth

Figure 43: Statistical representation and terms in an unconditional, level-2 growth model

A mathematical representation for the level-two or conditional (between persons) growth model consists of various terms as seen in Figure 43⁷. When we add a predictor variable to explore differences among groups of individuals over time, we generate a conditional, growth model (level two, between individuals). This model compares the average difference of intercepts and slopes between groups and tests for statistically significant differences between groups.

⁷ EDDL 650 (spring 2016) Course slide from Instructor

For example, in the current study for training GTAs in grading, we wanted to explore the differences between various GTAs by grouping them according to their experience level. In our study, we used a binary-coding system, where novices = 0 and experienced (trained) GTAs =1. GTAs Mickey, Mimi and Milo were novices in our study coded as “0” while Molly had prior experience in grading and back-reading and was coded as “1”. The “prior experience level” thus becomes a variable (known as a predictor variable) for analyzing the nature of changes in measured outcome as a function of time. For purposes of brevity in this chapter, we have included findings for using participation and gender as predictor variables. Results and discussions for these models are presented in later sections of this chapter. Of course, the overarching goal of this exercise is not just to provide definitive answers to these questions but also substantiate them with evidence from observed trends or statistical data from the growth model.

4.4.3 Missing Data

Often, longitudinal analysts encounter missing data for some participants at some points of time for various reasons. The Hierarchical Linear Modeling (HLM) software allows for extrapolation / interpolation to account for missing data in one of two ways: either delete the missing data during the initial model set-up or include missing data extra/interpolation during analysis only. The option of excluding missing data while preparing the input file for HLM modelling versus excluding missing data while running the model analyses is extremely useful for researchers using this technique.

4.4.4 Other Useful Results from A Growth Model Output

We have already described unconditional growth models (level one, within individual) and conditional growth models (level two, between individuals).

Another, rather the most basic model is called a null model which examines just the variation in the measured outcome (no predictors at level one or level two). This model is called an unconditional or null model, represented as:

$$Y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

In the first equation, Y_{ij} is the outcome for the i^{th} individual at level-1, nested in the level-2-unit j . i.e., it is the level-1 intercept β_{0j} + random unexplained variation (or error), r_{ij} . This represents a level -1 of the null model (for within individual variance)

In the second equation, β_{0j} is the outcome at level-2 of the null model, and is the sum of γ_{00} , which is the average initial status across all individuals (level-2 intercept), + u_{0j} , the random variation (or error) associated with β_{0j} due to individual variations. This null model is the equivalent of a one-way ANOVA and is used to establish a baseline model for comparisons with higher-level models.

(i) ICC (ii) Deviance and (iii) Pseudo R^2

Intraclass Correlation Coefficient (ICC)

If a substantial proportion of the total variance in the measured outcome (DIFSCR) can be explained by the amount of variance at level two (between individuals, or the value of u_{0j}), then the most suitable statistical model is likely to be different from a simple regression model.

This proportion is termed as the intraclass correlation coefficient (ICC), calculated using the formula $(\tau_{00}) / (\sigma^2 + \tau_{00})$

where $\tau_{00} = u_{0j}$ = variance at level 2 and $\sigma^2 = r$ = variance at level 1.

The value of ICC ranges from 0 to 1. According to literature sources¹⁴⁵, ICC values > 0.1 can be a valid justification for the use of a multi-level growth model rather than simple regression models.

Pseudo R²

To examine how much variance is accounted for as predictors are added to a model, we can calculate Pseudo R².

Pseudo R² = $[(\sigma^2(\text{unconditional}) - \sigma^2(\text{conditional})) / [\sigma^2(\text{unconditional})]]$.

For example, a PseudoR² value of 0.5 calculated after gender is added as a predictor to a model would be interpreted as 50% variance between individuals is explained by gender.

Deviance Statistic

Lastly, we are interested in the “fit” of our growth model!

Because the null model is like a baseline for later models, a deviance statistic represents the primary fit in growth model analysis. It is mathematically stated as:

Deviance statistic = -2 (natural log of likelihood ratio)

We begin by examining the deviance statistic of the null model. After entering predictors in the unconditional growth model (level one, within individual) we compare deviance statistics for this model with those of the null model by hypothesis testing. Performing iterative deviance tests before adding higher-level predictors into an existing growth model ensures that the previous models already have the best possible fit.

A smaller deviance statistic indicates a better fitting model.

A large deviance statistic indicates a poorly fitting model.

A model with predictors and a similar deviance statistic is not an improvement over the unconditional or previous model.

4.4.5 Summary

Growth models allow us to track development or changes in desired outcome measures as a function of time. In the case of GTA training, the measured outcome is the difference between the scores provided by the GTA and the expert on the same laboratory report (DIFSCR). Since every academic term and course comprises of a varied demographic among GTAs, variables such as GTA experience, participation in the study, gender or other teaching-related beliefs are likely to be informative level two predictors.

A quasi-experimental design does not involve random assignment to a control or treatment group for data collection and analysis purposes. In our back-reading study, we worked with several constraining factors such as non-mandatory participation, alignment with department-laid rules and policies outlined with the Graduate teaching Fellow Federation (GTFF) for work hours and expectation. Data collected during the back-reading had several factors contributing to missing data. However, we are looking at growth models as an exploratory method to assess training program efficacy at this stage. This approach adds to the uniqueness of studies such as back-reading and adding a novel element to program assessment in the larger context of chemistry education research.

4.4.6 Research Questions

With the background information about growth models, we recap with our research questions for the reader's benefit:

- How does the trend in individual trajectory for a GTA change with time?
- What inferences can we draw from the growth trajectory about the impact of training in grading? (Predictor variable: participation)
- Considering other predictor variables such as 'experience' or 'gender' for a level-two growth model which factors explain the extent of differences in growth trajectories /trends for GTAs?
- What information can a multiple predictor growth model provide in the context of the backreading study?

4.5 Methods, Data and Results

4.5.1 Data Collection

Back-reading training and implementation were carried out as described in the earlier chapter(s). We now summarize the data sourcing for developing a level-one and level two growth model below.

All participant GTAs signed IRB forms permitting the use of graded lab reports and any written comments for inclusion in this research project. All identifying information was redacted before using scanned images or screen captures of graded student work. GTA names were substituted with pseudonyms and numeric codes on datasets.

GTAAs were invited to attend a back-reading session with the course instructor or Head GTA each week. During these sessions, three randomly selected laboratory reports from the GTAAs’ current grading pile were photocopied and graded by both the GTA and the expert. Back-reading sessions lasted about 30-40 minute for all three laboratory reports unless the GTA or expert had reasons for further grading/ discussion. These were our “back-read reports.” As a follow-up to the back-reading process, GTAAs were also requested to draw three additional laboratory reports from their *graded* pile(s) for photocopying before returning to the students. These additional samples may be termed as “post-check” laboratory reports since back-reading on these laboratory reports was performed *after* the reports had been graded, and relevant feedback about grading was provided only to the GTA(s) and any score revisions were not available to students. In case the GTA(s) chose *not* to attend a back-reading session with an expert, the only available data was from the “post-check” laboratory reports (3 per week), and these were used for building the data sets for growth model analyses.

Since our BR study at UO was implemented for two years, we consider the very first TA orientation in the laboratory as our initial time-point (t=0). Since we have data from each week hereon, time coding for the growth model was spans 0-89 weeks in total. We obtained grading data from sample staff meeting grading, individual back-reading, and post- check back-reading reports for this duration. The time coding for each “chunk” of our BR study is shown in Table 20.

Table 19: Time-coding for longitudinal data in two-year back-reading study

Term	Fall Y1	Winter Y1	Spring Y1	Fall Y2	Winter Y2	Spring Y2
Coding (Week)	0-10	14-24	26-36	52-62	67-77	79-89

4.5.2 Data Preparation

Our variable of interest for building the growth model was the difference of scores (DIFSCR) between those provided by the GTA and expert. For purposes of eliminating the effect of *direction* of discrepancy (i.e., GTA score higher or lower than the expert), only the *magnitude* of this difference was recorded as raw data. Since more than one graded laboratory report was available for data recorded each week per GTA, the *average* and *standard deviation* values of DIFSCR were also considered as potential sources of data for building the growth model. Therefore, in all our growth model simulations, our independent variable is time (in weeks), and our dependent variable or outcome variable is the modulus of difference of scores (DIFSCR) which is a unitless quantity. Based on necessity of our analyses, we also considered a manipulated form of DIFSCR such as average per week as the dependent variable. For level-two predictor variables, we used (a) participation and (b) gender. The coding for both predictor variables was intentionally simplified to a binary format to keep our growth model interpretations simple and easy to understand.

Table 20: Predictor Coding by Participation and Gender

Predictor Coding	Participation	Gender
0	Non-Participant	Male
1	Participant	Female

4.5.3 Example of GM Data for A Single GTA's Grading Pattern

We have discussed GTA Molly's grading of laboratory reports extensively in the previous chapter. Therefore, our individual GTA growth model analysis is presented in the next section using GTA Molly's back-reading data available for the time duration as shown in Table 22. We collated data from Molly's back-reading meetings and post-check laboratory reports. Molly was a GTA for Fall, winter, and spring terms in Year 1 of our study and Fall, winter terms in year 2. For a total of five terms, we have longitudinal data coded by week and student ID.

Table 21: Data for GTA Molly; coded for weeks in the back-reading study

Term	Fall Y1	Winter Y1	Spring Y1	Fall Y2	Winter Y2
Coding (Week)	0-10	14-24	26-36	52-62	67-77

We begin with an unconditional means model (or commonly known as a null model) which has no predictors at either level to examine whether there is systematic variation in the outcome worth exploring. If yes, does it exist within person(s) or between two or more individuals?

For GTA Molly, the unconditional means model output and graph is as shown in Figure 44. This model shows us an intercept value (γ_{00}) = 1.31 ± 0.12 , which is statistically significant (t -ratio = 10.374, df = 35, $p < 0.001$).

Mixed Model

$$DIFSCR_{ij} = \gamma_{00} + \gamma_{10} * CLOCK_{ij} + u_{0j} + r_{ij}$$

Final Results - Iteration 197

Iterations stopped due to small change in likelihood function

$$\sigma^2 = 1.25476$$

τ
INTRCPT1, β_0 0.04583

Random level-1 coefficient	Reliability estimate
INTRCPT1, β_0	0.127

The value of the log-likelihood function at iteration 197 = -2.280733E+02

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	1.917850	0.188941	10.151	35	<0.001
For CLOCK slope, β_1					
INTRCPT2, γ_{10}	-0.019276	0.004058	-4.750	107	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	d.f.	χ^2	p-value
INTRCPT1, u_0	0.21408	0.04583	35	39.09129	0.291
level-1, r	1.12016	1.25476			

Statistics for current covariance components model

Deviance = 456.146643

Number of estimated parameters = 2

Figure 44: Model output for GTA Molly's longitudinal data, unconditional means model

$$DIFSCR_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

Final Results - Iteration 2

Iterations stopped due to small change in likelihood function

$$\sigma^2 = 1.29630$$

τ

INTRCPT1, β_0 0.20847

Random level-1 coefficient	Reliability estimate
INTRCPT1, β_0	0.391

The value of the log-likelihood function at iteration 2 = -2.317212E+02

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	1.305556	0.119924	10.886	35	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	d.f.	χ^2	p-value
INTRCPT1, u_0	0.45658	0.20847	35	57.51429	0.010
level-1, r	1.13855	1.29630			

Statistics for current covariance components model

Deviance = 463.442443

Number of estimated parameters = 2

Figure 45: Level-1 model output for GTA Molly's back-reading data

When we proceed with generating a level-1 model for GTA Molly's back-reading data, the model fit improves slightly based on deviance test results as described previously. That is to say, adding the "time" as a variable allows for a better model fit, and is likely to help explain the variance in DIFSCR better than the unconditional means model. Figure 45 shows the model output for the level-1 model.

The intercept value (γ_{00}) = 1.91 ± 0.19 , is analyzed in the model and determined to be statistically significant (t-ratio = 10.151, df = 35, $p < 0.001$).

The slope of the level model is -0.019 ± 0.004 indicating a significant decreasing trend over time (see Figure 46), which is a desirable outcome for our study. (t-ratio = -4.750, df = 107, $p < 0.001$). We are also limiting our growth model fit to only linear, instead of exploring a quadratic, cubic or other polynomial fit at this stage.

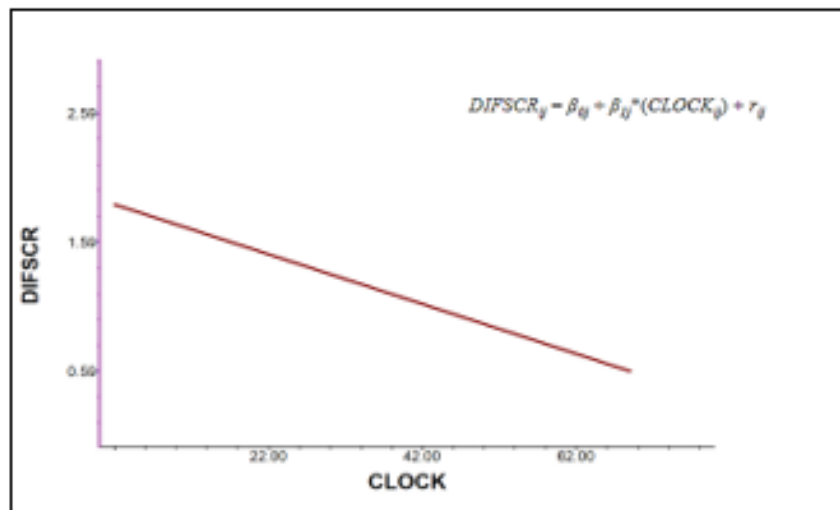


Figure 46: Level-1 model graph for GTA Molly's back-reading data

Model Interpretation Summary

Table 23 highlights the estimated parameters of interest from the model outputs along with a brief explanation of the technical definition and interpretation.

Table 22: Estimated parameters for GTA Molly, level-1 model

	Technical notation and definition	Interpretation	Values from model output
Level-1 intercept (initial status)	γ_{00} = Average true initial status for GTA Molly	Average DIFSCR value estimated at t=0	1.91*±0.19
Level1 slope (rate of change)	γ_{10} = Average true rate of change	Average change in DIFSCR over time	-0.019* ± 0.004 (per week)
Within person variance	μ_0 = within person variance	Amount of variation within individuals over time	0.0458 (not statistically significant)
Equation: DIFSCR_(Molly) = 1.91 -0.019(Clock) + error			

For Molly’s back-reading data from weeks 0 through 77, an initial status of DIFSCR = 1.91±0.19 which is statically significant compared to 0($t = 10.151$, $df = 35$, $p < 0.001$). This is indicative of the GTAs receptiveness to the training provided at the start of the term. With regards to slope, a decreasing trend in the outcome variable over time is indicative of GTA’s grading was progressing in a desirable direction and tending to zero (or 100% agreement with expert scores). This rate of change is further quantified as (-)0.019 ± 0.004 units/week and found to be statistically significant as well. Deviance for the null model is 463.44 and for the level-1 model it changes to 456.15. As described earlier, a lower deviance is a better fit.

Based on this output, we can say that GTA Molly’s individual growth curve follows a promising trajectory as a reliable grader.

4.5.4 Example of GM Data Over Two Years of Back-Reading for Groups of GTAs

Let us now consider multiple individuals using back-reading data from four GTAs (names changed) Molly, Mickey, Mimi, and Milo. collected over two years. This dataset contains the ID, clock and DIFSCR variable but also codes for back-reading participation (Treatment compliance abbr. TXCOMPLI), term as well year for examining any changes over time. The TXCOMPLI variable coding was performed based on observational and empirical evidence from back-reading meetings. If a GTA attended three or more back-reading meetings, their TXCOMPLI status was coded as a participant or numerically, as one. Otherwise, it was coded as a non-participant or numerically as 0. (participant = 1, non-participant =0). Therefore, a GTA who graded inaccurately or unreliably even after attending individual back-reading was still coded as a participant. Consequently, it was anticipated that their growth trajectory would reflect the negligible impact of training despite their participation. A nested model or level-2 model with one predictor variable provides us additional information such as the effect of that predictor e.g., participation and therefore the data for participant GTAs with negligible training impact would be reflected in the nested growth models comparing groups such as participants sand non-participants.

Model Interpretation Summary

Figure 47 shows the model output for a level-2 model using GTA codes (1,2,3, and 4) and treatment compliance (TXCOMPLI). Figure 48 is the graph output corresponding to this model and can be interpreted using the tabulated information below.

Using four GTAs (Molly, Mickey, Mimi, and Milo) back-reading data from weeks 0 through 90, we observe an initial status of $DIFSCR = 4.65 \pm 0.30$ which is statically significant compared to 0 ($t = 15.31$, $df = 108$, $p < 0.001$). There is missing data for GTAs during the weeks 0-14 and therefore we see a gap in the model graph. This tells us that there is significant difference between the estimates for true initial status for participant GTAs versus non-participants. The former group has an average 2.48 *lower initial* DIFSCR (from γ_{01}) compared to non-participants. The mean difference between GTAs initial status is statistically significant and tells us that back-reading training has a measurable impact. With regards to slope, a decreasing trend in the outcome variable over time is indicative of GTA's grading was progressing in a desirable direction and tending to zero (or 100% agreement with expert scores). This rate of change is further quantified as $(-)0.019 \pm 0.004$ units/week and found to be statistically significant for the two groups of GTAs as well. Deviance for the level-1 using time predictor is 1433.74 and for the level-2 model with the TXCOMPLI predictor, is lowered to 1352.80. Therefore, the addition of participation as predictor variable is substantive, and helps to explain the variation between trajectories for groups of GTAs. The individual trajectories for the four GTAs are as shown in Figure 49. At this stage, we can provide sufficient evidence that participation in back-reading training follows a linear decreasing trend following the trajectory similar to an expert grader and impacts grader reliability positively. A model that uses two or more predictor variables provides useful information about the combined effect of both variables on the outcome. In the case of using the participation variable and individual GTA codes, model results continued to be statistically significant.

$DIFSCR_{ij} = \gamma_{00} + \gamma_{01} * TXCOMPLI_{ij} + \gamma_{10} * CLOCK_{ij} + u_{0j} + r_{ij}$

Run-time deletion has reduced the number of level-1 records to 427

Final Results - Iteration 6

Iterations stopped due to small change in likelihood function

$\sigma^2 = 1.04993$

τ
INTRCPT1, β_0 0.46538

Random level-1 coefficient	Reliability estimate
INTRCPT1, β_0	0.635

The value of the log-likelihood function at iteration 6 = -6.764010E+02

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. df.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	4.657550	0.304213	15.310	106	<0.001
TXCOMPLI, γ_{01}	-2.483208	0.209287	-11.865	106	<0.001
For CLOCK slope, β_1					
INTRCPT2, γ_{10}	-0.024976	0.003965	-6.300	318	<0.001

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. df.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	4.657550	0.207093	22.490	106	<0.001
TXCOMPLI, γ_{01}	-2.483208	0.162221	-15.308	106	<0.001
For CLOCK slope, β_1					
INTRCPT2, γ_{10}	-0.024976	0.003331	-7.498	318	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	df.	χ^2	p-value
INTRCPT1, u_0	0.68218	0.46538	106	288.26280	<0.001
level-1, r	1.02466	1.04993			

Statistics for current covariance components model

Deviance = 1352.802057
Number of estimated parameters = 2

Figure 47: Growth model output for Four individuals using TXCOMPLI (participation) as predictor variable

Table 23: Estimated parameters for FOUR GTAs (Molly, Mickey, Mimi, and Molly), level-2 model with participation as predictor

	Technical notation and definition	Interpretation	Values from model output
Levl-1 intercept (initial status)	γ_{00} = mean of level-1 intercepts for individuals with predictor variable = 0	True initial status for non-participants.	4.65±0.30*
	γ_{01} = Mean difference in level-1 intercepts for a 1-unit difference in level-2 predictor	True initial status difference between non-participants and participants.	-2.48±0.21*
Level1 slope (rate of change)	$\beta_1 = \gamma_{10}$ = Mean difference in level-1 slopes for a 1-unit difference in level-2 predictor.	True rate of change for non-participants.	-0.025±0.003*
Within person variance	σ_0^2 = level-2 residual variance in true intercept across all individuals	Residual variance in intercept controlling for participation	--
Between person variance	σ_1^2 = level-2 residual variance in true slope across all individuals	Residual variances in true rate of change, controlling for participation	0.465* ($\chi^2 = 288.26, df = 107, p < 0.001$)

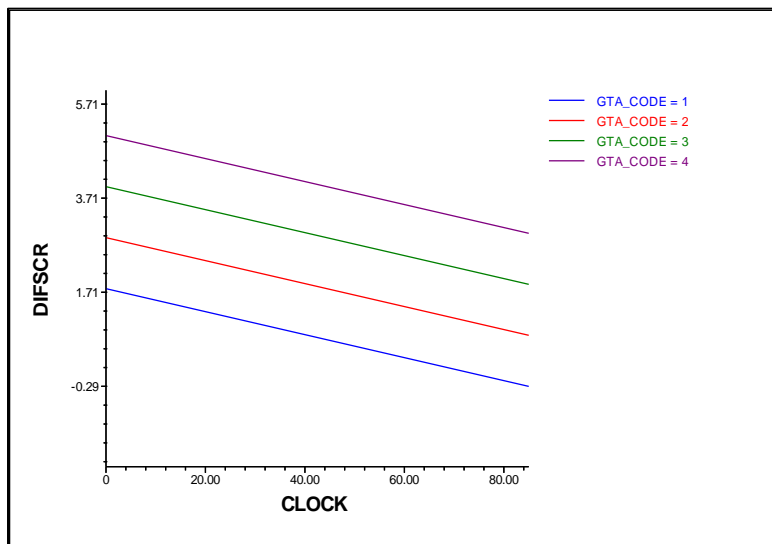


Figure 48: Graph output for level-2 growth model using data from four GTAs, TXCOMPLI and Individual GTA code predictor variables

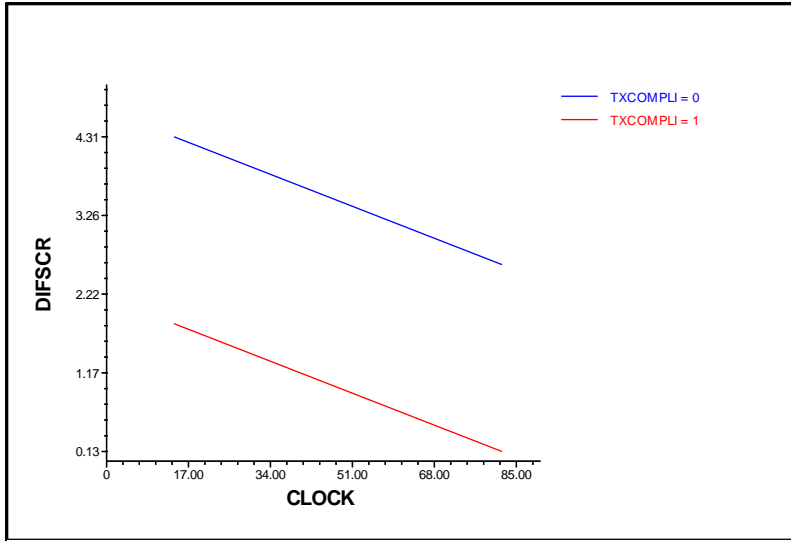


Figure 49: Graph output for level-2 growth model using data from four GTAs, TXCOMPLI predictor variable

Note On the Linearity And “Clean Data” Model Graphs

We believe the equidistant and parallel trendlines are due to missing data combined with the limitations of using a linear fit. We absolutely acknowledge the possibility that better fitting non-linear trajectories would result if we explored quadratic or other higher order variables in our datasets. However, due to limited time and beginner-level knowledge of this approach and technical proficiency, we limited our exploration to a linear fit.

4.5.5 Example of Two Years GM Data of Back-Reading for All GTAs

We also used the participation (TXCOMPLI) predictor to examine data for 108 GTAs over 2 years. Our input data was the average DIFSCR recorded on four discrete occasions in each term for each GTA. Data was coded for GTA ID (n =108), Clock (weeks), Participation (TXCOMPLI) and gender. As discussed extensively, not all GTAs were participants. Complete dataset for each GTA were collected as much as possible, and coding was informed by observational and empirical data from training, staff meeting and individual backreading meetings. The model output for a level-1 growth model using participation (TXCOMPLI) status of 108 GTAs as a predictor variable is shown in Figure 50. We provide a brief discussion of model estimates in Table 25. The model graph is shown in Figure 51.

Table 24: Estimated parameters for all GTAs (n =108), level-1 model.

	Technical notation and definition	Interpretation	Values from model output
Level-1 intercept (initial status)	γ_{00} = Average true initial status for large group of GTAs (n =108)	Average DIFSCR value estimated at time = 0	1.89±0.27 * (t = 7.034, df =107, p<0.001)
Level1 slope (rate of change)	γ_{10} = Average true rate of change	Average change in DIFSCR over time	-0.0004 ± 0.004 (per week)
Within person variance	μ_0 = within person variance	Amount of variation within individuals over time	1.37* (χ^2 = 629.06, df=107, p<0.001)
Equation: DIFSCR_(Molly) = 1.88 -0.00004(weeks) + 1.36 + 1.08			

Mixed Model

$$DIFSCR_{it} = \beta_{00} + \beta_{10} * WEEKS_{it} + r_{0i} + e_{it}$$

Run-time deletion has reduced the number of level-1 records to 424

Final Results - Iteration 6

Iterations stopped due to small change in likelihood function

$$\sigma^2 = 1.08831$$

τ
INTRCPT1, π_0 1.36750

Random level-1 coefficient	Reliability estimate
INTRCPT1, π_0	0.828

The value of the log-likelihood function at iteration 6 = -7.199388E+02

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. df.	p-value
For INTRCPT1, π_0					
INTRCPT2, β_{00}	1.880996	0.267420	7.034	107	<0.001
For WEEKS slope, π_1					
INTRCPT2, β_{10}	-0.000471	0.004788	-0.098	315	0.922

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. df.	p-value
For INTRCPT1, π_0					
INTRCPT2, β_{00}	1.880996	0.185861	10.120	107	<0.001
For WEEKS slope, π_1					
INTRCPT2, β_{10}	-0.000471	0.004447	-0.106	315	0.916

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	df.	χ^2	p-value
INTRCPT1, r_0	1.16940	1.36750	107	629.06071	<0.001
level-1, e	1.04322	1.08831			

Statistics for current covariance components model

Deviance = 1439.877666
Number of estimated parameters = 2

Figure 50: Level-1 model output for 108 GTAs

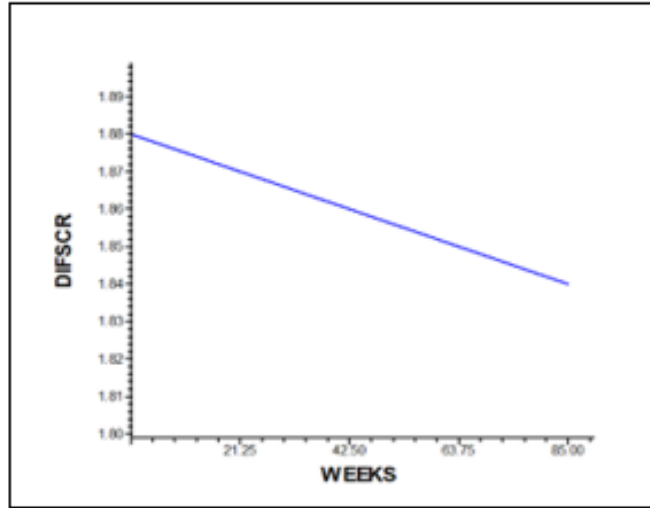


Figure 51: Graph output for level-1 model for 108 GTAs

Model Interpretation Summary

The overall group of GTAs over two years showed a decreasing trend in DIFSCR measures. This was found to be statistically significant. The rate of change in DIFSCR, although decreasing, is not statistically significant. The error term in the level-1 model output shows us that the time predictor is substantive in explaining the variance in DIFSCR. That is, addition of the time predictor variable helps explain the positive impact of back-reading training over two years. The graph output below agrees with our findings for this model. Finally, we examine the data for all 108 GTAs with 2 predictors at level 2: Participation and gender (male or female). The model output for a level-2 growth model using participation (TXCOMPLI) status of 108 GTAs as a predictor variable is shown in Figure 52. This is followed by a summary discussion based on the model estimates (Table 26) and model graph (Figure 52).

$$DIFSCR_{it} = \beta_{00} + \beta_{01} * TXCOMPLI_i + \beta_{02} * GENDER_i + \beta_{10} * WEEKS_{it} + r_{0i} + e_{it}$$

Run-time deletion has reduced the number of level-1 records to 424

Final Results - Iteration 7

Iterations stopped due to small change in likelihood function

$\sigma^2 = 1.05336$

τ
INTRCPT1, π_0 0.43031

Random level-1 coefficient	Reliability estimate
INTRCPT1, π_0	0.613

The value of the log-likelihood function at iteration 7 = -6.701312E+02

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. df.	p-value
For INTRCPT1, π_0					
INTRCPT2, β_{00}	4.181326	0.346645	12.062	105	<0.001
TXCOMPLI, β_{01}	-2.583515	0.208012	-12.420	105	<0.001
GENDER, β_{02}	0.442158	0.166150	2.661	105	0.009
For WEEKS slope, π_1					
INTRCPT2, β_{10}	-0.026852	0.003953	-6.793	315	<0.001

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. df.	p-value
For INTRCPT1, π_0					
INTRCPT2, β_{00}	4.181326	0.276000	15.150	105	<0.001
TXCOMPLI, β_{01}	-2.583515	0.156018	-16.559	105	<0.001
GENDER, β_{02}	0.442158	0.159771	2.767	105	0.007
For WEEKS slope, π_1					
INTRCPT2, β_{10}	-0.026852	0.003600	-7.459	315	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	df.	χ^2	p-value
INTRCPT1, r_0	0.65598	0.43031	105	270.01675	<0.001
level-1, e	1.02633	1.05336			

Statistics for current covariance components model

Deviance = 1340.262307
Number of estimated parameters = 2

Figure 52: Level-2 model output for 108 GTAs using participation and gender as predictor variables

Table 25: Estimated parameters for 108 GTAS level-2 model with participation and gender as predictors.

	Technical notation and definition	Interpretation	Values from model output
Level-1 intercept (initial status)	γ_{00} = mean of level-1 intercepts for individuals with predictor variable = 1	True initial status for male GTAS	4.18±0.346* (t =12.062, df =105, p<0.001)
	γ_{01} = Mean difference in level-1 intercepts for a 1-unit difference in level-2 predictor	True initial status difference between male and female GTAs	0.44 (not statistically significant)
Level1 slope (rate of change)	$\beta_1 = \gamma_{10}$ = Mean difference in level-1 slopes for a 1-unit difference in level-2 predictor.	True rate of change for male GTAs	-0.027±0.003*
Within person variance	σ_0^2 = level-2 residual variance in true intercept across all individuals	Residual variance in intercept controlling for gender	1.05
Between person variance	σ_1^2 = level-2 residual variance in true slope across all individuals	Residual variances in true rate of change, controlling for gender	0.43* ($\chi^2 = 270.02$, df = 105 p<0.001)

From the tabulated output estimates (Table 26) and graph in Figure 53 we can see that backreading does have a measurable impact on the outcome variable. Male GTAs who participated in backreading have significantly lower initial status compared to those that did not participate actively. Female participant GTAs also follow a similar trend. There are no detectable significant differences between male and female GTAs, but definitely exist between participants and non-participants. Once again, the limitations of using only a linear fit result in parallel-looking model fits, which appear equally spaced. At this stage, there is no specific interpretation for these trendlines. The only reliable comment on the slope or true rate of change is that it is statistically significant over time. The inclusion on

participation and gender terms allows for a better model fit as seen from the deviance and the statistically significant estimate of the variance in the data for 108 GTAs.

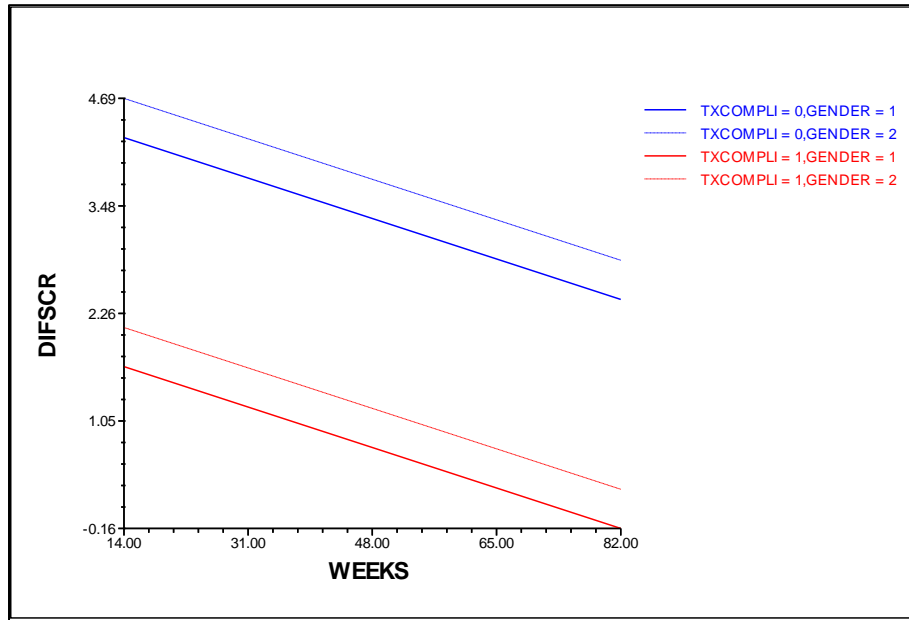


Figure 53: Graph output for level-2 model for 108 GTAs using participation and gender as predictor variables

4.6 Discussion

In response to the research questions stated in this chapter: The individual trajectory as shown for GTA Molly and as seen from other level-2 model graphs and estimates shows a downward trend in DIFSCR. This is a desirable outcome for our backreading study.

The rate of change in DIFSCR was found to vary significantly over time. This merited further inclusion of predictor variables and examination of other influencing factors. Predictor variables such as participation and gender add substantive robustness to the model and help us explain the systematic variance in DIFSCR trends. Considering multiple predictors in a single growth model also provided us with illustrative evidence of the

positive influence of backreading training and which predictors are relatively more reliable for explaining the variance in trends.

Our growth model approach is at a preliminary or exploratory stage for this type data and study. Limitations of our growth model approach include missing data, absence of informative predictor variables such as educational major/background, years of experience in teaching and self-reported teaching efficacy measures. The biggest limitation, though, is the use of only a linear fit to attempt an explanation of the trend and estimates. Utilizing higher order derivatives of data and including them in the growth model would undoubtedly provide us with rich, informative, and statistically robust interpretations that could be used to further fine-tune the backreading process and make it effective using evidence-based designs.

The use of a growth model approach is most certainly a notch above the conventional qualitative feedback methods. Further, the statistical estimation, ability to utilize large data sets with highly simplistic measurement criteria (ID, clock, outcome, predictors) makes it an attractive tool to explore. Comparatively, quantifying, inductively reasoning with and coding data obtained from focus groups, surveys, interviews etc., is highly time-consuming, often impacted by researcher bias and subject to non-continuity i.e., the feedback may not necessarily be valid at a later time. A growth model accounts for not only the measures but also the variance in measures and allows for a deeper analysis with ready-to-use tools, thus making program assessment a robust, informative process with a relatively faster turnaround time on interpretation and implications.

CHAPTER V: CONCLUSION AND FUTURE WORK

We conclude by summarizing our research findings from the backreading study and identifying potential areas of extending its scope for future researchers. The demography and population of graduate students pursuing higher education has changed drastically over the last 70 years (Figure 54)¹⁴⁶.

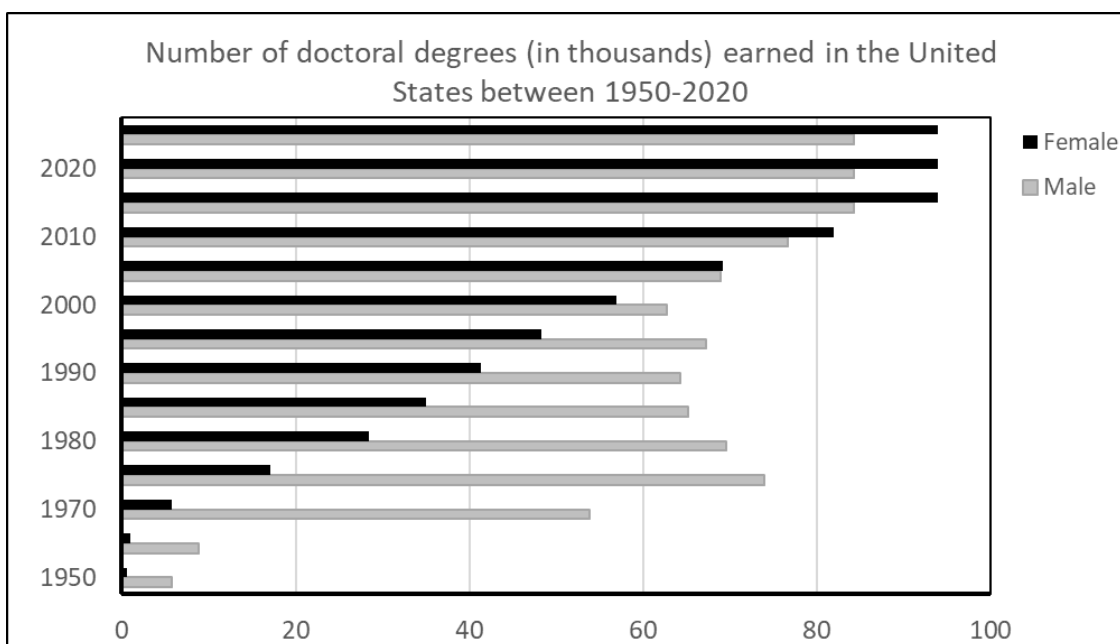


Figure 54: Adapted data from Statista for doctoral degrees awarded between 1950-2020

PhD graduates with proficiency in specific disciplines or materials are sought competitively in industry and other research collaborations. As trained professionals, they encounter new knowledge and methods to navigate expectations of performance for career growth. There is also an element of mentorship associated with these career trajectories. As a mentee, receiving timely and effective feedback from supervisors or peers is important to deliver the best possible results within specified timeframes.

As a mentor, being able to assess the current success status, rate and provide input on furthering it becomes critical. As part of their graduate training, doctoral students undergo regular evaluations at weekly or monthly meetings with academic advisers. They are also expected to successfully complete annual formal evaluations and presentations at conferences etc. to demonstrate their academic and research progress. Reliable understanding of the subject matter and consistent demonstration of progress as mentees are benchmarks for these evaluations. The frequency and high standards of these evaluations substantiate the necessity of appropriate training programs for graduate students.

Teaching experience is often a small element relative to the discipline-specific training and research components in a graduate students academic and professional development. However, it is possibly the only exposure to being a ‘mentor’ to another individual and working to ensure *their [student’s] success*. As Dewey rightly says, “Education in order to accomplish its ends both for the individual learner and for society must be based upon experience—which is always the actual life-experience of some individual”¹⁴⁷. Many graduate students pursuing doctoral degrees choose to pursue careers as academic scholars and tenured teaching faculty where the tenure portfolio is broadly partitioned as “research-teaching-service” with ratio of each vary by type of institution. For example, a Tier 1 research institution would specify an assistant professor’s academic profile as 40% Research, 40% Teaching and 20% Service with university specific expectations of each. A two-year community college on the other hand is likely to specify 70% Teaching, 20% Research and 10% Service.^{6, 148}. Thus, future faculty members would be expected to draw from their experiences as a GTA to perform their teaching responsibilities in academia.

This further reinforces the need for training in assessment or grading as a key skill for GTAs.

This dissertation provides a detailed account of development of backreading as a training protocol. Exploration of areas where further elements of training are required is evidenced through qualitative and quantitative analysis of grading. Most importantly, we conclude this work with a first-known implementation of growth models as a longitudinal program evaluation method.

5.1 Conclusions (Summary of Results from Previous Chapters)

5.1.1 Design and Development of Back-Reading for Training GTAs In Grading

Training in grading for GTAs must be a continuous professional development process. That is, training should not be restricted to only the “orientation” session at the start of term or orientation week, but throughout the academic term or year. Back-reading process for training GTAs is an adaptation of an established evaluation protocol used for grading Advanced Placement (AP) Chemistry student responses. Thus, validation of this method itself is not in question. The effect of back-reading training is observable after grading 3-4 laboratory reports initially, and this number decreases with time based on empirical data. GTAs response and feedback on BR training indicates mostly positive responses where they recognize the value of being able to grade accurately and consistently. BR process also serves as needs-based platform for GTAs to share their concerns about grading, quality of student work, or the demands on their own time. Observations specific to international GTAs (non-native speakers of English) indicate that they valued such training and

opportunities for interaction more than domestic peers because it helped them overcome cultural and language barriers when communicating with their students.

Monitoring GTAs' grading using back-reading and providing them with feedback on their grading practices throughout the term serves multiple purposes:

- Understanding rubric design and accurately implementing rubric criteria. (GTAs input during staff meeting grading exercises were included to ensure optimal rubric design and clearly phrased criteria).
- Consistent grades across the board mean reduced complaints from students about lower grades.
- Reduction in grade inflation issues due to monitoring and feedback to GTAs via individual BR meetings.

GTAs who incorporate BR training into their teaching practices are found to not just be accurate and consistent graders, they are also likely return assignments to students in a timely manner. By receiving accurate feedback on their weekly laboratory reports, students learn from their mistakes, improve their writing skills as well as their understanding of chemistry. These factors impact the overall success rate in general chemistry courses and STEM discipline attrition. BR Training program design and analysis of the results provide insight into the factors that potentially impact the success and sustainability of such initiatives. University and departmental-level policies, time and personnel constraints are all essential variables to consider with BR studies. Based on our preliminary data at the beginning of the study, we believe orientation week training should, unquestionably address grading as a key GTA responsibility. Clearly outlining expectations of accurate

and consistent grading along with hands-on opportunity to grade and understand these expectations are critical for GTAs. There is tremendous scope for a shift in perspective from “grading, a time-consuming task” to “grading, an important skill in a professional’s toolkit” and training in grading should be designed to be a continuous, heuristic process to reinforce this view.

5.1.2 Qualitative Analysis of Graded Student Laboratory Reports

Qualitative examination of two high-quality and two low-quality laboratory reports (as designated by GTAs) were reported to highlight the areas where GTA feedback or annotations were determined as relevant and effective (or otherwise) for the student.

Molly was a back-reading participant and reliable GTA based on laboratory observations, constructive and effective feedback to her students. We observed data for Molly’s class as having an overall positive environment and cumulative development of scientific writing skills.

Klaus’ grading shows numerous instances of irrelevant and at times, non-legible feedback which invariably created an overall negative experience for his students. Klaus did not participate in back-reading meetings and only underwent the required start-of-term orientations. However, we also believe that Klaus and his students would have greatly benefited by his continued participation and utilization of resources in back-reading training.

The implementation of an organically designed conceptual analysis rubric (CAR) designed to examine the degree (high, medium, or low) of chemistry-specific proficiency in student responses shows that discrepancies in GTAs grading continue to exist despite provision of

training. This highlights two major issues for consideration: two independent rubric designs (back-reading and conceptual analysis) resulted in similar scores for the researcher but several deviating scores for GTAs. Granted these analyses were performed independently, however, the significant DIFSCR gaps between expert and GTAs tell us that the back-reading study lacks a chemistry concept focused component of training. Inclusion of such a component will not only reduce grading discrepancies, but also help GTAs evaluate and provide comments specific to chemistry proficiency to their students, beginning with rubric design. A chemistry concepts component to training would have GTAs participate in active learning methods like think-pair-share or flipped classrooms, wherein they would play the role of students during staff meeting and visualize how students may think or interpret their findings. Such active learning elements work to attune GTAs to look for critical concepts and the extent students actually understand them during laboratory experiments. GTAs would then couple the chemistry concepts task they performed during staff meeting with their own observations of students in the laboratory during the session and use all this information holistically to comment, annotate, and guide students on conceptual understanding. It will be interesting to examine such qualitatively rich data by itself. Additionally, if there is scope of coding or quantifying it, then the longitudinal analysis of students' chemistry conceptual mastery within a 4-year majors' discipline is an exciting long-term project to consider.

5.1.3 Growth Models as A Unique Approach for Evaluation of Grading Training

Our BR longitudinal data spans two years and is the first known report of implementing a growth model approach to a training program for GTAs.

Our explorations of using a growth model to explain any systematic variations in outcome variable (DIFSCR) the difference of scores between GTAs and expert were limited to a linear fit and yielded promising results. We determined that GTAs' grading skills generally improve over time and follow a favorable trajectory having minimal deviations from an expert grading trajectory. However, many factors can and will influence these trends and they have not been explored to their full potential.

The individual trajectory as shown for GTA Molly and as seen from other level-2 model graphs and estimates shows a downward trend and a favorable outcome for our backreading study. Most GTAs begin at an initial DIFSCR of 4 or 5 points, and gradually understand and assimilate grading like an expert tending to DIFSCR of 1 or zero. The rate of change of DIFSCR is very small in most of our model simulation but estimated parameters show us they are statistically significant. Inclusion of predictor variables and examination of other influencing factors added substantive value and evidence to explain the systematic variance in DIFSCR trends. There were no statistically significant differences between Male and Female GTAs. However, Participation as a predictor variable resulted in statistically significant differences between BR participants and non-participants, thereby supporting our hypothesis that BR training is effective and impacts GTA grading practices positively. The growth model approach presented is preliminary or exploratory at this stage for this type data and study. Limitations of our growth model approach include missing data, absence of informative predictor variables such as educational major/background, years of experience in teaching and self-reported teaching efficacy measures. We also acknowledge the inherent limitations caused by use of only a linear fit to attempt an explaining trend from parameter estimates. Utilizing higher order derivatives of data and

including them in the growth model could certainly provide us statistically robust interpretations. However, these were not pursued due to researcher's limited technical knowledge and availability of time at this stage.

5.2 Future Work Ideas

5.2.1 Furthering the TA Training Protocol: Expectancy Value Theory

We have utilized andragogy⁹⁴ to develop our framework for present research. This adult-learner-centric on the adult learner, continues to be process-oriented i.e., it seeks out information to build a '*needs-input-feedback*' loop that analyze and enhance adult learning programs. Our backreading study is a successful example of applying the assumptions in andragogy to GTAs and providing professional training in grading laboratory reports. We have also successfully provided evidence of the influence of such training by measuring an unbiased outcome variable and information-rich longitudinal data. There is tremendous scope for further exploration of GTA adult learners' specific motivation and values associated with the task of grading. Expectancy value theory (EVT) is a development lens which would be helpful in examining GTAs' beliefs and values towards grading. This theory from the early 2000s by Eccles et al.¹⁴⁹ postulates that achievement-related decisions are motivated by a combination of individuals' *expectations for success* and *subjective task value* in particular domains. For example, children are more likely to pursue an activity if they expect to do well and they value the activity. The EVT model further identifies four major components in task values: *attainment value* (i.e., importance of performing task well), *intrinsic value* (i.e., personal enjoyment), *utility value* (i.e., perceived usefulness for future tasks), and *cost* (i.e., competition with other

goals). According to the expectancy-value model, expectations for success and task value are shaped by a combination of factors. These include child characteristics (abilities, previous experiences, goals, self-concepts, beliefs, expectations, interpretations) and environmental influences such as cultural milieu, socializers' beliefs and behaviors ¹⁵⁰. Variables or patterns of interest for furthering the backreading study would be (a) pre-and-post training values GTAs assign to grading-related tasks and (b) quantifying or numerically coding these measures as predictor variables in a growth model. The working hypotheses may be stated as:

- (i) GTAs with higher positive expectation values would demonstrate more persistent effects of backreading training and have higher rate of growth as expert graders.
- (ii) GTAs with lower expectation values would demonstrate a significant shift in their pre-versus post training trajectories possibly attribute-able to the effect of training and recognition of actual value added by grading-related activities to their regular responsibilities.

By implementing the combination of andragogy, EVT and growth modeling methods, we could also examine the bigger picture with questions about GTAs' self-concepts as teachers, their beliefs about their own teaching and grading; values they place on the outcomes of their grading of student assignments. Findings from these studies would also inform resources designed to support students in the general chemistry course as well.

5.2.2 A Unique Rewards program: GTA Digital Badges

Figure 55 is a visualization of the number of articles published in the last decade about “digital badges. Leading up to and during the COVID-19 pandemic there is a significant growth in these studies.

A digital badge is a form of recognition for learning achievements outside the traditional academic records. Digital badges validate the accomplishment, skill, or competencies earned in typically online learning environments. In other words, Digital badges are virtual visual representations, accomplishments, skills, or awards that present the characteristics of physical merit badges or awards but go farther in providing validation to viewers in that they are linked to metadata or artifacts. This new technology is capable of transforming accreditation and evaluation processes for educational and corporate contexts using information participants performance in training or work related activities.¹⁵¹

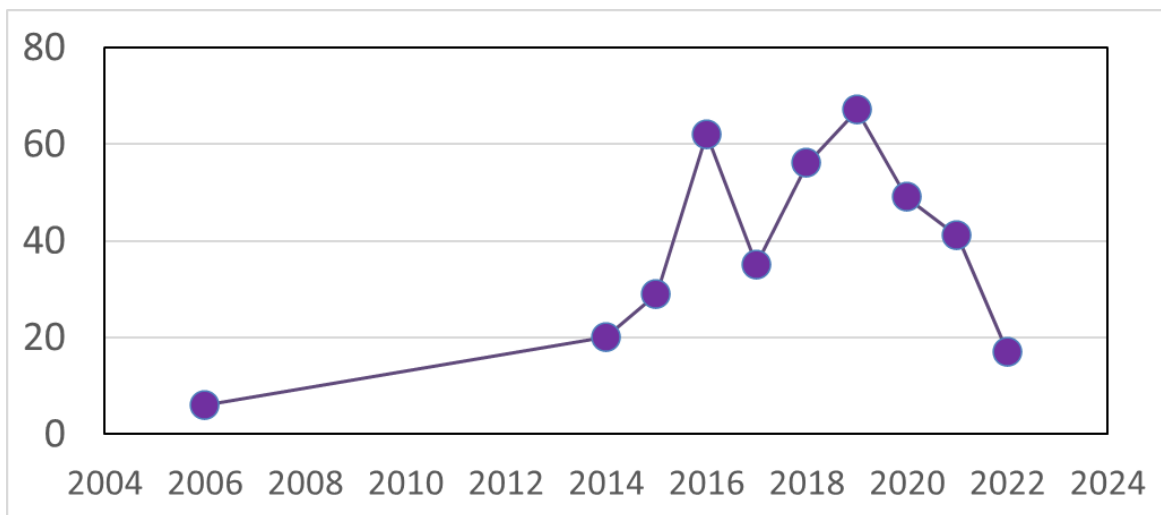









Figure 55: Scatter plot showing the rise of publications on "Digital Badges" between 2004-2022

Digital badges have been extensively reviewed between 2018-2021 and are incorporated for purposes ranging from motivation, encouragement, participation, recognition^{152, 153} to

awareness behavior change self-regulated learning, and achievement, skill accomplishment or final assessment.¹⁵⁴ There is empirical evidence of digital badging positively influencing learner motivation even without affecting grades.¹⁵⁵ Utilizing digital badges for pre-service teachers and students in programs for education licensure provide a wealth of qualitative data on the perspectives of participants and how they assign a utility value to their digital badge history. Some examples of grading -related digital badges for TAs are outlined below. These are designed based on existing digital badges in teaching/education.¹⁵⁶ Therefore, there is substantial evidence that digital badges can be a promising add-on to goal-specific training programs such as back-reading. Table 27 is an example of some proposed ideas for GTA digital badges in the backreading training program.

Digital Badge Design is informed by at least three core features: specific function of badges, the structure of badge systems, and the different types of design and interaction features used with badges.¹⁵⁷ Though users can be motivated by the social status accompanying badges, designers also have to consider that the more users earn a particular badge, the more its value diminishes¹⁵⁸. Other design considerations for incorporating rank-based rewards include whether to award badges for meeting absolute or relative levels of effort and the accessibility of winner information¹⁵⁹ Without highlighting appropriate information¹⁶⁰ implies that, badges may serve more as “personal affirmation rather than status” which may impact engagement and motivation significantly.

Table 26: Proposed ideas for GTA Digital Badges in grading

		<p>Back-reading Participation (successful completion of start-of-term training)</p>
		<p>Grading Accuracy badge at regular intervals (weeks, quarter, or semesters) (Demonstrated accuracy for a specific duration)</p>
		<p>Consistency badge (Demonstrated consistency ($\leq \pm 1$ DIFSCR for one or more academic years))</p>
		<p>Grading leader/ Expert Grader (Appointed as a grading lead to guide or assist novice graders or peers)</p>

There are a few instances of incorporating digital badges in high-school chemistry. digital badges have been developed for standard solution, volumetric pipetting, and titration.^{161, 162}). For first year undergraduate chemistry courses, digital badges for pipetting,¹⁶³ burets and volumetric flasks,¹⁶⁴ calibration, titration, distillation and standard solution

preparation,¹⁶¹; chemical hygiene and safety.¹⁶⁵ and transferable skills badges such as thinking and problem solving; use of tools, technology, and software; oral communication etc.¹⁶⁶ developed for post-secondary chemistry and other science courses have been reported. In the Hill et al. study¹⁶⁶, students preferred the badges to be awarded on the basis of merit rather than completion. Some badges are voluntary assignments while others were required and worth ~1% of students' grades²⁷.

5.2.3 Vision for The Growth Model Framework to Analyze GTA Training

Programs

Although collecting data for back-reading is definitely a challenge in terms of time-investment and effort. the biggest factor impacting training in grading and subsequent longitudinal growth modeling is GTA participation. Between course program outcomes, fair hours and wages issues, TA union policies and university protocols, GTA participation can become critical for this type of analysis. However, there is hope for the back-reading protocol to deliver and generate favorable growth trends if it can be implemented at a larger scale and involve a wider variety of participants and/or variables. We revisit an image (reproduced as Figure 56) from the previous chapter as a visual aid to framing this vision for further implementation of the growth model approach.

Figure 56 shows a nesting model considered for an individual GTA (Ducky) at UO. If we were to expand this pool of participants to multiple courses at varying undergraduate levels (freshman, sophomore, junior, senior) in multiple disciplines such as Physics, Biology etc. we would have a significantly large dataset only from UO. Nesting growth models deliver information about factors that cause positive or negative effects in TA training with grading. Chapter 4 discusses a simple two-level growth model using compliance

(participant or non-participant) as a factor. The future vision is to include multiple factors such as age, gender, teaching experience, major, year of graduate study, self-concept and measures based on expectancy value theory – each of which could provide an intricate picture of GTAs professional growth as individuals, groups and even generalize findings to a larger population.

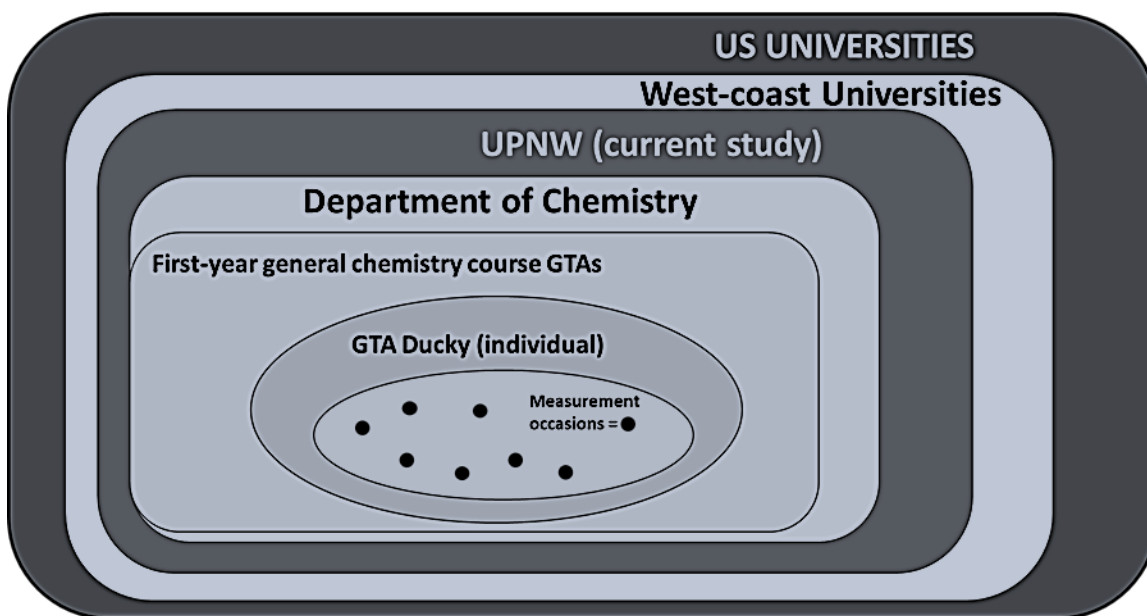


Figure 56: Visualization of an individual GTA as part of a nested population

Piecewise models are another interesting approach in hierarchical linear modeling, where in the effect of a time-lag or a gap can provide direct evidence of the impact of a training program or intervention. With GTAs, the likelihood of positive, rapid growth (i.e., grading trends matching the predictions) is higher during orientation week or the first year of teaching the course. As GTAs gain experience and familiarity with the course goals, logistics and develop a professional routine, their willingness to participate and value for the need for training is likely to decrease or even disappear. This would be reflected in the

“gap” between initial training and re-training when the issue is identified or addressed at a later stage. A piecewise growth model with this hypothesis is an excellent way to provide evidence for training program success and identify a time-based trend to reinforce its necessity. Of course, there is always the equal probability that a piecewise growth model will provide evidence of continued persistence of training and disprove the original hypothesis. Thus, Implementing backreading training at other universities would result in more nested data to explore the effect of this training program at different universities locally (Pacific Northwest), regionally (US West-Coast) and even the US (nationally). Thus, scaling up our data and using nesting variables would not only provide us more robust data about GTA training program success, but also help us identify predictors that can be fine-tuned to make such programs impactful for a diverse population.

5.2.4 The learning curve for grading in online courses and virtual classrooms

In 2016, there were approximately 50 symposia at the Biennial Conference on Chemical Education (BCCE oral presentations) featuring either an online classroom, virtual laboratory, or similar format of virtual learning such as online classes, learning management for online classrooms and student engagement in online courses incorporated into a chemistry curriculum. The extent of such public presentations of on-going research is indicative of a paradigm shift for STEM learning. By the time this dissertation was completed, the COVID-19 pandemic had disrupted every imaginable sphere of our lives. The world as we know is reconstructing itself to around in-person, virtual and hybrid learning spaces and workplaces. This toggling between modes of learning has also caused

a tremendous shift in designing courses, materials, compatible learning strategies, assignments, and assessments.

We have already discussed the importance of laboratory experiences for chemistry students in the introduction chapter. However, the growth of full-blown online courses in chemistry now presents serious challenge to evidence-based approaches like guided inquiry. Hands-on learning experiences in the chemistry laboratory can never be fully replicated in a virtual setting, despite state-of-the-art the graphics, tools, and simulation software. A large portion of these approaches is the ability to perform experiments in groups or active learning settings.^{20, 22, 25, 60, 144, 167}

Student assessment in online learning can be holistic, but difficult to authenticate for actual evidence of learning. Multiple-choice responses to typed responses can always be worked around, or for lack of a better metaphor, the system can be gamed. However, there are some promising studies that are paving the way for navigating novel learning experiences in chemistry for lecture and laboratory courses.^{153, 161, 162, 168}

Although there are recent reports of utilizing digital badges as an assessment tool for successful completion of a laboratory practical exam,²⁷ the virtual-ness of today's world where the "accuracy" of a response might be the most complex variable available, limits our operational room for grading. This makes the online assessment simplistic (yes/no, Likert scale or similar simplified models of assessment), limits merit-based awarding for students' responses and challenging to train evaluator to look beyond the simplified grading scale and provide feedback. Also, back-reading in a virtual setting can become much more isolating, and even cause participants to turn away from an additional online meeting in the day. Ensuring accuracy, reliability and consistency of grading depends

heavily on the content being assessed, the rubric being used and the grader's skill set. All these variables are likely to be watered down in online assessments and will require a major makeover to the back-reading training protocol.

5.2.5 Summary of Research Results

We have reported our findings using a modified format of weekly staff meetings with chemistry GTAs. The focus of these modifications was GTAs' grading of student chemistry laboratory reports. In preliminary studies, we determined there were significant errors in GTAs' ability to grade a laboratory report accurately and consistently. Our modified approach to weekly staff meetings is adapted from existing protocols implemented in College Board Advance Placement (AP) Readings. When implemented over two-years the results indicate that GTAs who participate in training of grading and the back-reading process, are better able to apply rubrics to accurately grade chemistry laboratory reports. Our study provides valuable information about GTAs' requirements for training in grading laboratory reports and the positive impact of such on-going professional development on GTAs' grading accuracy and consistency.

APPENDIX A: COMPONENTS OF LAB REPORT

An outline of the expected components of the post-laboratory report for “Density Exploration” as provided to students.

EXPERIMENT: DENSITY EXPLORATION

BEGINNING QUESTIONS

List of research questions for this experiment.

CLAIMS, EVIDENCE AND ANALYSIS

Claims made based on data, observations, and calculated results

Evidence to support claims

Analysis of results comparing individual data to pooled (class) data

REFLECTION

Error Analysis

Extend the experiment to explore additional research questions

Explain connection to concepts/ideas from the lecture course

Explain how experiment connects to real life scenarios

Report Literature sources that explain, confirm or dispute experimental findings.

APPENDIX B: GENERIC RADING RUBRIC

Grading rubric provided to the GTAs at UO for scoring the post-laboratory report on “Density Exploration.”

<u>Criteria</u>	<u>Points awarded</u>	<u>Points possible</u>
<u>General Details: Name, Partners, Title, Date</u>		1
<u>Beginning Question(s):</u> What question(s) did I have? What question(s) did the class group decide to investigate?		4
<u>Claim, Evidence and Analysis:</u> What results can be reported to answer the beginning questions? (Claim)?		5
What is my interpretation of my data (graphs, class data, trends, or other analysis) to support my claim(s) (Evidence)?		5
<u>Reading and Reflection:</u>		
a. What are the possible sources of error? How would those errors impact my results?		5
b. How have my ideas changed, what new questions do I have, or what new things do I have to think about?		5
c. How does this work tie into concepts about that I have learned in class?		5
d. To what can I refer in my text, my notes, or some real-life application to make a connection with this laboratory work?		5
e. What related reading have I done to explain, confirm, or dispute what I have learned via this laboratory experience, how have I tied what I have read into my work for this activity, and what resources have I cited?		5
TOTAL:		40

APPENDIX C: SAMPLE LABORATORY EXPERIMENT

The Food Dyes Laboratory (reproduced from laboratory manual)⁸

Spectrophotometric Determination of a Food Dye

INTRODUCTION

Food additives, such as artificial colors and flavors, are ubiquitous in the American diet. There is often concern about whether these substances are safe to consume, or more importantly, whether they are safe to consume in the quantities that are present. It is not the purpose of this experiment to determine the safety or efficacy of food dyes but rather, to determine the amount that is present in several popular drinks. This experiment introduces **spectrophotometry** as a method for quantitative analysis.

BACKGROUND

Spectrophotometry

Read Appendix E, *Spectrophotometry*.

The Beer-Lambert law states:

$$A = \epsilon bc$$

where

A is absorbance (unitless)

ϵ is the molar absorptivity coefficient ($\text{M}^{-1} \text{cm}^{-1}$)

b is pathlength (cm)

c is concentration (M)

We can see that using this relationship, the concentration of a solution is directly proportional to its absorbance.

⁸ (105) Exton. *General Chemistry Lab Manual vol.2*.

To find the concentration of an unknown, a standard (or calibration) curve is made by measuring the absorbance, A , of several **standard solutions**, which are solutions of known concentration. Note that when A is plotted versus c , a straight line passing through the origin and with a slope of (ϵb) results. The concentration of the unknown is determined by using this equation and the absorbance of the unknown.

It should be noted that when solution concentrations are too high or too low there are deviations from Beer's law and there is no longer a linear relationship between absorbance and concentration. In this case it is no longer possible to utilize a standard curve to determine concentration and solutions must be diluted until their absorbance is within the linear range.

Food Dyes

Coloring agents have been used as food additives for centuries. They help us to identify foods visually. For instance, lime and orange sherbets would be nearly indistinguishable based on appearance if not for the added green and orange colors. They add a festive appearance to foods—M&M's candies would still taste just the same if they were all colored gray, but where's the fun in that? They are also added to foods because we have very strong expectations about what colors should be associated with certain foods. All else being equal, would you be more likely to buy a bright orange-colored orange, or one that is a mottled brown-green?

Coloring agents have been added to foods for less legitimate reasons as well. At the beginning of the 20th century, when there was no regulation of color additives in this country, coloring agents were added to food to mask inferior or spoiled foods, and some coloring agents marketed for inclusion in food were just flat out poisonous. Since passage of the Federal Food, Drug, & Cosmetic (FD&C) Act of 1938, color additives in the US have been the responsibility of the Food and Drug Administration (FDA). A recent controversy in the news concerns the addition of a dye, canthaxanthin, to farm raised salmon. The dye gives the fish the deep red color consumers expect. After a lawsuit was filed in Seattle by a consumer advocate group, local grocery chains were forced to label all fish containing the dye.

The FDA splits coloring agents into two categories: "certifiable" and "exempt from certification." The former are derived primarily from petroleum; the latter includes agents derived largely from mineral, plant, or animal sources. Certified colors are further broken down into water-soluble "dyes" and water-insoluble "lakes," with most colors being available in both forms. At this time (2014) there were nine color additives certifiable for food use. Three of these will be used in this lab.

Red food dyes have a history of controversy. In 1960, additions to the FD&C Act of 1938 included the so-called Delaney amendment. This amendment prohibits the marketing of any coloring agent that has been found to cause cancer in animals or humans, *no matter what the dose*. For many years, FD&C Red No. 3 was the most important red dye used in foods. But, in 1983, a single study found that FD&C Red No. 3 could be associated with thyroid cancer in male rats. On the basis of that

study, the FDA banned all uses of Red Lake No. 3 and several uses of Red Dye No. 3.⁹ FD&C Red No. 2 met a similar end several years earlier,¹⁰ with the curious result that, for a time, there were no red M&M's candies. As of today, Red Dye 3 remains certified for use in foods. However, food manufacturers have almost entirely abandoned this dye in favor of FD&C Red No. 40.

You can read more about color additives in foods, drugs, and cosmetics, at

<https://www.fda.gov/food/food-additives-petitions/color-additives-food>
(last accessed June 2019)

CHEMISTRY IN A SUSTAINABLE WORLD

Food dyes and other colored substances owe their color to the presence of *chromophores*, the part of a molecule that absorbs visible light. For reasons that are beyond the scope of this book, chromophores are frequently found in organic molecules containing alternating single and double bonds, such as exist in Allura Red. Complexes containing metal ions can also be chromophores. Some of these metal ions, particularly those that are referred to as *heavy metals*, are of concern in the environment due to their toxicity to plants, animals, humans, wildlife, and aquatic life. Chemists are engaged in an on-going effort to mitigate the effects of these pollutants in the environment, and more importantly in terms of global sustainability, to minimize their use.

Traditionally, students studying the Beer-Lambert law performed experiments using brightly colored solutions of heavy metal ions such as copper, cobalt, or chromium. This resulted in the generation of large quantities of waste that required hazardous waste disposal. To eliminate the problem of toxic waste and to enact green chemistry principles, this experimental procedure has been modified to use food dyes, relatively harmless substances that can safely be washed down the drain when the experiment is finished.

The drive to replace heavy metal chromophores with less hazardous substances is not limited to academic teaching laboratories. Red, orange, and yellow pigments for the paint industry and others have historically been created using toxic heavy metals such as lead, chromium, and cadmium. In response to concerns about these compounds in the environment, Engelhard Organic Pigments has developed an environmentally friendly line of pigments for use in packaging and entirely phased out its use of heavy metals. Prior to this transition, the company produced 6.5 million pounds of metal containing pigments per year. In addition to eliminating heavy metals, a water-based manufacturing process was incorporated instead of using the organic solvents typically associated with the creation of pigments. For this, Engelhard Organic Pigments (now BASF Corporation) received the 2004 Designing Safer Chemicals Green Chemistry Achievement Award from the Environmental Protection Agency.¹¹

⁹ Dept. of Health and Human Services, FDA, *Federal Register* 1990, 55(22), 3515–3543.

¹⁰ U.S. Food and Drug Administration, *Color Additives Fact Sheet*, <https://www.fda.gov/industry/color-additives-cosmetics/color-additives-and-cosmetics-fact-sheet> (last accessed 5/2020)

¹¹ <http://www2.epa.gov/green-chemistry/presidential-green-chemistry-challenge-winners> (accessed 5/2020)

PROCEDURE

CAUTION!

Although the material used in this experiment are food dyes, treat them with the same respect that you would show for any chemical unknown: do not taste, do not inhale the dust, and minimize skin contact.

There are 3 different food dyes that will be used this week, FD&C Yellow No. 5, FD&C Red No. 40 (Allura Red), and FD&C Blue No. 1. You will work individually to prepare solutions and make a Beer's Law graph for one of the dyes, as assigned by your lab instructor.

Part A: Preparation of Stock Solution

1. Using the information in Table 13-1, calculate the mass of your assigned food dye that is necessary to produce 500 mL of a solution of the target molarity. To check your work, verify that your calculated value is less than the mass shown in the last column of the table.

Table 13-1. Food Dye Information

Dye	Molar Mass (g/mol)	Target Molarity	Mass should be less than:
FD&C Yellow No. 5	534.37	1.75×10^{-4} M	0.06 g
FD&C Red No. 40	496.42	1.90×10^{-4} M	0.06 g
FD&C Blue No. 1	792.86	3.65×10^{-5} M	0.02 g

2. Before carrying out the next steps, make sure your hands are clean and *dry*! Take a 500 mL volumetric flask to the balance with you.
3. Place a square of weighing paper on the balance and tare the balance. Carefully weigh to the nearest 0.0001 g the quantity of food dye calculated in step 1. (Start with a very small scoop of material—the total quantity required is comparable in size to your small fingernail.) Don't walk away from or re-tare the balance yet.
4. Quantitatively transfer the food dye to the 500-mL volumetric flask. Place the weighing paper (now devoid of food dye) back on the same balance and record the mass. Take the weighing paper back to your station with you.
5. Fill the volumetric flask to the line with deionized water and mix well. This constitutes the stock solution from which you will prepare standard solutions. Dribble a few drops of water on the area of the weighing paper where the food dye sat. Record what you see.

Part B: Preparation of Standard Solutions

1. Clean and dry four containers of at least 100 mL capacity and label them A, B, C, and D.
2. Refer to Table 13-2 for the recommended quantity of stock solution to use to prepare your standard solutions. Transfer some of your stock solution to a small beaker and pipet from this beaker during steps 3 and 4. Because of the risk of contamination you should never pipet directly from the stock solution flask.

Table 13-2. Quantity of Stock Solution for Preparing Standard Solutions

Dye	Solution A	Solution B	Solution C	Solution D
FD&C Yellow No. 5	1 mL	5 mL	10 mL	15 mL
FD&C Red No. 40	1 mL	5 mL	10 mL	15 mL
FD&C Blue No. 1	2 mL	5 mL	10 mL	15 mL

3. Prepare standard solution A by pipetting the quantity of stock solution given in Table 13-2 into a 100-mL volumetric flask. Dilute to volume and mix well. Transfer this to your container marked A, where it will be stored until you need it later.
4. Repeat, using the quantities in Table 13-2 for solutions B, C, and D to prepare the remaining three standard solutions. The standard solutions will be used to generate the calibration curve, and to determine λ_{\max} . For convenience, the standard solutions will hereafter be referred to as std. soln. A, std. soln. B, etc.

Part C: Collection of Absorption Spectra and Determination of λ_{\max}

λ_{\max} is the wavelength at which a sample absorbs most strongly, i.e., at which the absorbance is the largest. Whatever their concentration, all samples of the same substance have the same wavelength of maximum absorbance (λ_{\max}). The *amount* of light absorbed may vary, but the *energy*, or *wavelength*, of light absorbed remains the same. To determine λ_{\max} for the food dye you are using, an absorbance spectrum is collected over the visible wavelength range and the wavelength with maximum absorbance is determined to be λ_{\max} .

It is always necessary to use a “blank solution” when calibrating a spectrometer. This is a solution that contains all species that may be present *except* the species of interest. In the case of an aqueous solution, such as the food dye solution, the blank is deionized water. By calibrating with a blank, you are ensured that the measured absorbance is only due to the species of interest.

Standard Solutions

1. Start the Logger *Pro* 3.4.5 software.

2. To calibrate the Spectrometer, select **Experiment > Calibrate > Spectrometer**. The calibration dialog box will display the message: “Waiting ... seconds for lamp to warm up.” Allow the spectrometer to warm up for at least three minutes. Follow the instructions in the dialog box to complete the calibration. The small container used to hold the sample is referred to as a “cuvette.” For this experiment, where water is the solvent, fill the “blank” cuvette about $\frac{3}{4}$ full with de-ionized water. When handling the cuvette, touch the ridged sides only to avoid getting fingerprints on the windows. Insert the cuvette in the spectrometer with the non-ridged sides facing top and bottom (toward the words “Vernier Spectrometer”). Click **Finish Calibration**, and when that is finished Click **OK**.
3. Fill a cuvette about $\frac{3}{4}$ full of your “std. soln. D.” Place the sample in the cuvette holder of the Spectrometer and click **Collect** (green button).
4. Click **Stop** (red button) to end data collection.
5. To optimize the view of the absorbance spectrum that you have just collected, select **Analyze > Autoscale > Autoscale**.
6. Fill a cuvette about $\frac{3}{4}$ full of your “std. sol. C,” place in the cuvette holder and click **Collect**. When the dialog box opens, click the blue button “Store Latest Run”; this will allow you to show (overlay) your “std. sol. C” on the same screen as your “std. sol D.” Click **Stop** (red button) to end data collection.
7. Repeat with “std. sol. B” and “std. sol. A.”
8. Now all 4 absorbance spectra should be on the screen, with “std. sol. D” the “highest” and ‘std. sol. A’ the “lowest” peak.
9. To change the absorbance spectrum colored lines to black, **Double-click** on the box just above the colored row of data you want to change (“Abs” box), click on the **Options** radio button, then on the **Color** tab, then the “scroll arrows,” scroll down to the bottom to **Black** and click. Repeat for remaining colored lines.
10. To find the wavelength of maximum absorbance (λ_{\max}), select **Analyze > Examine**. This will bring up a line on the screen that you can move to the wavelength of maximum absorbance. Record this value. It will also bring up a box with the Wavelength and Absorbance values that you can “grab” and move anywhere on the graph, and by **Double-clicking** in this box you can increase the font size also. Notice that moving the mouse moves the line.
11. **Double-click** anywhere else on the graph to bring up the “Graph options” box to type your name in the title box.
12. Once you have your graph with the line at exactly maximum absorbance (λ_{\max}), without moving the mouse, type Command (⌘) P to print the absorbance spectrum.
13. The data can now be selected in the data table to the left, copied, and pasted into Excel for further analysis.

Part D: Concentration of Food Dye in Commercial Drinks

There are samples of 6 commercial drinks on the center bench. These drinks contain either one food dye (single component dye) or a combination of two of the food dyes that you and your classmates have made calibration curves for. Working with the other person at your table who used the same food dye as you, use a 50-mL beaker to obtain about 30 mL of the single component drink that contains “your” food dye. You can get more later if needed.

Consider the intensity of the color of this drink and decide whether the absorbance of the drink will fall within the range of your standard solution data. If uncertain, make a measurement of the absorbance using the previous procedure.

If the absorbance is not within the standard solution range, prepare a dilution of the drink solution to obtain one that is within the data range. Be sure to record the dilution factor that you used because you will need to use it in your calculations to determine the concentration of the food dye in the original solutions.

Once you have prepared a solution that is within range, measure and record the absorbance at λ_{\max} of that solution.

Working with all of the others at your lab bench, use a 50-mL beaker to obtain about 30 mL of the drink that contains both of the food dyes used at your lab bench. Repeat the above process to determine the concentrations of both of the food dye components in the drink. It may be necessary to prepare different dilutions to measure the different dyes.

CALCULATIONS

Prepare a data table that, for each standard solution (including std. soln. D), includes concentration and measured absorbance at λ_{\max} . Remember to give a sample calculation for the determination of concentration.

Generate a calibration curve by plotting absorbance vs. concentration. Perform a linear regression analysis and plot the line of best fit.

Use the equation for the line and the absorbance of the drink solution to determine the molarity of the food dyes in the drink. Remembering that the drink mixes were diluted and weaker than the normal drinks would be, calculate the concentration of the food dyes present in the beverage when consumed at their normal concentration.

REPORT

In addition to the normal components of every lab report, your report for this experiment should include the identity of the dyes that you studied, the spectrum (absorbance vs. wavelength) used to determine m_{\max} , the calibration curve, the molar absorptivity coefficient of the food dye that you studied, the molarity of the food dyes in the diluted drink solutions and the molarity of the food dyes in undiluted drinks. Additional items to consider in your discussion:

- Why do we want to know / use λ_{\max} ?
- When you weighed and then transferred the food dye to the volumetric flask, you made the assumption that the quantity you weighed all actually made it into the flask—that’s what “quantitative transfer” implies. The point of re-weighing the empty weighing paper and then dribbling water on it was to help you evaluate the goodness of that assumption. *Did* all of the food dye get into the flask? If not, can you estimate (order of magnitude) how much did not? How does all of this affect the concentration of your stock solution?
- According to the Beer-Lambert law the calibration curve should be perfectly linear. The amount of “scatter” in the points is a reflection of experimental error. Comment on the linearity of your data and suggest a major source of error that would account for the scatter.

QUESTIONS

1. A student finishes with her std. soln. A and properly, thoroughly rinses out the 100-mL volumetric flask using distilled water. Then, however, she immediately proceeds to prepare her second standard solution without making any attempt to dry or shake out the residual water in the flask. What effect will this have on her std. soln. B? Will the concentration be too large? Too small? Explain.

2. Many common materials that we ingest (table salt, aspirin), though quite safe in reasonable quantities, become toxic when taken in very large doses. A measure of toxicity is the LD50 value (Lethal Dose, 50%). It is the quantity of material, expressed in mg of material per kg of subject-body-weight that, if administered to a population of subjects, will cause 50% of the population to perish. The LD50 value for FD&C Yellow No. 5 is 12,750 mg/kg in mice, FD&C Red Dye No. 40 is > 10,000 mg/kg in rats and for FD&C Blue No. 1 it is > 2,000 mg/kg in rats.

Assume that the LD50 value for humans is the same as for mice and rats. Calculate the number of mg of food dye present in an eight-ounce glass of the single component beverage you used in this lab. How many such glasses would a typical adult human have to ingest in order to reach the LD50 concentration? Be sure to state explicitly any major assumptions you make in order to do the calculation. Considering only the potential acute toxicity of the food dye and neglecting anything else that may be present in the drink, do you believe that the beverage you analyzed is a safe product?

3. Calculate the volume of aqueous waste that you produced while completing this experiment. Use this value to calculate the volume of aqueous waste that would be produced by a General Chemistry course with 800 students enrolled.

This lab procedure has frequently been performed using copper sulfate pentahydrate ($\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$) instead of a food dye. Consult the MSDS (Material Safety Data Sheet) for copper sulfate pentahydrate to determine the concerns associated with this compound and appropriate methods of disposal. Would you be able to pour this waste down the drain as you did the food dye waste?

APPENDIX D: STUDENT REPORT GUIDELINES

Student Report Guidelines for the Food Dyes Laboratory

Introduction (10 points)

- Theory: Use your notes, textbook resources and lab lecture material to write briefly about the theory and principles underlying this experiment. Some of the points you may want to include are *Beer's Law*; *Food Dyes*; *Sustainable Chemistry*
- BQs: State your beginning questions, class questions and the purpose of the lab. Remember that beginning questions are usually “researchable” OR explorable through the experiment you performed.
- Procedure: Include a highly concise summary of the experimental procedure which does NOT exceed 3 sentences* Safety: Highlight any safety measures/ points to note about chemicals and equipment.

Experimental (27 points)

- Data tables and observations must be included in an organized fashion (i.e., tabulated where needed) Always list the chemical identity/ names and details (formula, molar mass etc.) of the dyes you examined (knowns, unknowns, blanks etc.)
- Show Part 1 calculations (include all work) for preparation of Stock solutions and Standard solutions.
- Calibration Curve: Include your calibration curve for absorbance versus concentration of standard solutions as a graph with a good line-of-fit. Ensure the graph has all necessary features labeled. Show calculations for the slope and report the values that are needed for further calculations (with correct significant figures and units)
- Details of Unknown sample analyzed
- Include all pertinent details (Identification, source, amount used, dilution –if needed)
- Tabulate the data collected for absorbance of unknown
- Show all work for calculations of dilution and concentration of the unknown(s)

- Graph Outputs

Attach any graphs/spectrum output that are needed for unknown identification and calculations.

Discussion Section: (33 points)

- Write 1-2 claim statements as “answers” to your BQs and provide evidence for your answers using experimental results. Be sure to reference tables/ graphs as needed.
- Reflect on how the purpose of the lab was addressed. What did you measure? What did you calculate? What experimental steps were critical?
- How do the calculated/ output values you got match the literature values?
- If your results were off/ unexpected, what sources of error could have caused this? What assumptions did you make while performing the experiment? What errors could have affected the data you have from the graph/ spectra?
- What aspects of this lab were focused on green chemistry principles? How did this help in making the experiment safer for the environment?
- Answer all questions on page 94 and 95 of your laboratory manuals.

REFERENCES

- (1) Duffin, E. *Doctorate recipients' primary sources of financial support in the United States in 2020 [Graph]*. In *Statista*. NCSES; National Science Foundation; ID 240166; *Statista* October 05, 2021.
- (2) Duffin, E. Undergraduate enrollment in U.S. 4-year postsecondary institutions 1970-2029. In *Statista* NCES, March 31, 2021.
- (3) Gardner, G. E.; Jones, M. G. Pedagogical Preparation of the Science Graduate Teaching Assistant: Challenges and Implications. *Science Educator* **2011**, *20* (2), 31-41. From EBSCOhost eric.
- (4) Dotger, S. Exploring and Developing Graduate Teaching Assistants' Pedagogies via Lesson Study. *Teaching in Higher Education* **2011**, *16* (2), 157-169. From EBSCOhost eric.
- (5) Nicklow, J. W.; Marikunte, S. S.; Chevalier, L. R. Balancing pedagogical and professional practice skills in the training of graduate teaching assistants. *Journal of Professional Issues in Engineering Education and Practice* **2007**, *133* (2), 89-93. DOI: 10.1061/(asce)1052-3928(2007)133:2(89).
- (6) Brendel, W.; Cornett-Murtada, V. Professors Practicing Mindfulness: An Action Research Study on Transformed Teaching, Research, and Service. *Journal of transformative education* **2019**, *17* (1), 4-23. DOI: 10.1177/1541344618762535.
- (7) dik, H.; Laura April, M. A Graduate Teaching Assistant Workshop in a Faculty of Science. *Canadian journal of higher education (1975)* **2009**, *39* (2), 101. Curzan, A. *First day to final grade : a graduate student's guide to teaching*; Ann Arbor : University of Michigan Press, 2000. Nowlis, V. *The graduate student as teacher*; Washington American Council on Education, 1968. Kessel, W. G. Introductory college chemistry; Laboratory manual for introductory college chemistry (Quick, Floyd J.). *Journal of Chemical Education* **1966**, *43* (2), 109. DOI: 10.1021/ed043p109.1.
- (8) Marincovich, M.; Prostko, J.; Stout, F. *The Professional development of graduate teaching assistants*; Bolton, MA : Anker Publishing Company, 1998.
- (9) Nyquist, J. D.; Wulff, D. H. *Working Effectively with Graduate Assistants*; 1996.
- (10) Wisconsin Univ, M. C. o. L. a. S. Manual for Teaching Assistants. 1985.
- (11) Diamond, R. M.; Gray, P. J. A National Study of Teaching Assistants. ASHE 1987 Annual Meeting Paper. 1987.
- (12) Bloom, L. Z. *The Promise and the Performance: What's Really Basic in Teaching TAs*; 1976.
- (13) Prieto, L. R.; Meyers, S. A. *The Teaching Assistant Training Handbook: How To Prepare TAs for Their Responsibilities*; 2001.

- (14) *America's lab report : investigations in high school science*; Washington, D.C. : National Academies Press, 2006. Abraham, M.; Cracolice, M.; Aldhamash, A. The nature and state of general chemistry laboratory courses offered by colleges and universities in the United States. *Journal of Chemical Education* **1997**, *74* (5), 591-594. DOI: 10.1021/ed074p591.
- (15) Rudd, J.; Greenbowe, T.; Hand, B. M. Implementing the science writing heuristic in a general chemistry laboratory course. *Abstr. Pap. Am. Chem. Soc.* **2001**, 222, U220-U220.
- (16) Nikolic, S.; Vial, P. J.; Ros, M.; Stirling, D.; Ritz, C. Improving the Laboratory Learning Experience: A Process to Train and Manage Teaching Assistants. *Education, IEEE Transactions on* **2015**, *58* (2), 130-139. DOI: 10.1109/TE.2014.2335712.
- (17) Dragisich, V.; Keller, V.; Zhao, M. S. An Intensive Training Program for Effective Teaching Assistants in Chemistry. *Journal of Chemical Education* **2016**, *93* (7), 1204-1210, Article. DOI: 10.1021/acs.jchemed.5b00577.
- (18) Lawrenz, F.; et al. Training the Teaching Assistant. *Journal of College Science Teaching* **1992**, *22* (2), 106-109. From EBSCOhost eric.
- (19) Burke, K. A.; Hand, B.; Pooch, J.; Greenbowe, T. Using the Science Writing Heuristic: Training Chemistry Teaching Assistants. *Journal of College Science Teaching* **2005**, *35* (1), 36.
- (20) Polacek, K. M.; Keeling, E. L. Easy Ways to Promote Inquiry in a Laboratory Course: The Power of Student Questions. *Journal of College Science Teaching* **2005**, *35* (1), 52. From EBSCOhost eric.
- (21) Miller, K.; Brickman, P.; Oliver, J. S. Enhancing Teaching Assistants' (TAs') Inquiry Teaching by Means of Teaching Observations and Reflective Discourse. *School science and mathematics* **2014**, *114* (4), 178-190. DOI: 10.1111/ssm.12065. Burke, K.; Pooch, J.; Cantonwine, D.; Greenbowe, T.; Hand, B. M. Evaluating the effectiveness of implementing inquiry and the science writing heuristic in the general chemistry laboratory: Teaching assistants and students. *Abstr. Pap. Am. Chem. Soc.* **2003**, 225, U545-U545. Gallet, C. Problem-Solving Teaching in the Chemistry Laboratory: Leaving the Cooks. *Journal of chemical education* **1998**, *75* (1), 72-77. DOI: 10.1021/ed075p72.
- (22) Wheeler, L. B.; Maeng, J. L.; Whitworth, B. A. Teaching assistants' perceptions of a training to support an inquiry-based general chemistry laboratory course. *Chemistry Education Research and Practice* **2015**, *16* (4), 824-842, Article. DOI: 10.1039/c5rp00104h.
- (23) Wheeler, L. B.; Clark, C. P.; Grisham, C. M. Transforming a Traditional Laboratory to an Inquiry-Based Course: Importance of Training TAs When Redesigning a Curriculum. *Journal of Chemical Education* **2017**, *94* (8), 1019-1026. DOI: 10.1021/acs.jchemed.6b00831.

- (24) Chen, H. J.; She, J. L.; Chou, C. C.; Tsai, Y. M.; Chiu, M. H. Development and Application of a Scoring Rubric for Evaluating Students' Experimental Skills in Organic Chemistry: An Instructional Guide for Teaching Assistants. *Journal of Chemical Education* **2013**, *90* (10), 1296-1302, Article. DOI: 10.1021/ed101111g. Kelly, O. C.; Finlayson, O. E. Providing solutions through problem-based learning for the undergraduate 1(st) year chemistry laboratory. *Chemistry Education Research and Practice* **2007**, *8* (3), 347-361, Article.
- (25) Tatli, Z.; Ayas, A. Effect of a Virtual Chemistry Laboratory on Students' Achievement. *Educational technology & society* **2013**, *16* (1), 159-170.
- (26) Wheeler, L.; Sturtevant, H.; Mumba, F. Exploratory Study of the Impact of a Teaching Methods Course for International Teaching Assistants in an Inquiry-Based General Chemistry Laboratory. *Journal of Chemical Education* **2019**, *96* (11), 2393-2402, Article. DOI: 10.1021/acs.jchemed.9b00239.
- (27) Govindarajoo, G.; Lee, J. Y.; Emenike, M. E. Proof of Concept for a Thin-Layer Chromatography Digital Badge Assignment Within a Laboratory Practical Exam for a Nonchemistry Majors' Organic Chemistry Lab. *Journal of chemical education* **2021**, *98* (9), 2775-2785. DOI: 10.1021/acs.jchemed.1c00598.
- (28) Gardner, G. E.; Parrish, J. Biology graduate teaching assistants as novice educators: Are there similarities in teaching ability and practice beliefs between teaching assistants and K-12 teachers? *Biochemistry and molecular biology education* **2019**, *47* (1), 51-57. DOI: 10.1002/bmb.21196.
- (29) Young, S. L.; Bippus, A. M. Assessment of Graduate Teaching Assistant (GTA) Training: A Case Study of a Training Program and Its Impact on GTAs. *Communication Teacher* **2008**, *22* (4), 116-129. From EBSCOhost eric.
- (30) Swan, L. M.; Kramer, S.; Gopal, A.; Shi, L.; Roth, S. M. Beyond Proficiency: An Asset-Based Approach to International Teaching Assistant Training. *Journal of Faculty Development* **2017**, *31* (2), 21-27.
- (31) Randi, J.; Corno, L. Chapter 20 - Teacher Innovations in Self-Regulated Learning. In *Handbook of Self-Regulation*, Zeidner, M. B. R. P. Ed.; Academic Press, 2000; pp 651-685. Randi, J. Teachers as Self-Regulated Learners. *Teachers College record* / **2004**, *106* (9), 1825-1853. DOI: 10.1111/j.1467-9620.2004.00407.x info:doi/10.1111/j.1467-9620.2004.00407.x.
- (32) Sandi-Urena, S.; Gatlin, T. Factors Contributing to the Development of Graduate Teaching Assistant Self-Image. *Journal of Chemical Education* **2013**, *90* (10), 1303-1309. DOI: 10.1021/ed200859e.
- (33) Park, C. The graduate teaching assistant (GTA): lessons from North American experience. *Teaching in Higher Education* **2004**, *9* (3), 349-361, Article. DOI: 10.1080/1356251042000216660.
- (34) Hurst, C.; Sparrow, L. Professional Learning for Teaching Assistants and Its Effect on Classroom Roles. *Mathematics Education Research Group of Australasia* **2012**, er.

- (35) Kadi, A.; Beytekin, O. F.; Arslan, H. A Research on the Burnout and the Teaching Profession Attitudes of Teacher Candidates. *Journal of Education and Training Studies* **2015**, 3 (2), 107-113. From EBSCOhost eric.
- (36) Bengu, E. Adapting to a New Role as an International Teaching Assistant: Influence of Communicative Competence in This Adaptation Process. 2009.
- (37) Hebbani, A.; Hendrix, K. G. Capturing the Experiences of International Teaching Assistants in the US American Classroom. *New Directions for Teaching and Learning* **2014**, 2014 (138), 61-72. DOI: 10.1002/tl.20097.
- (38) Kim, M. A Comparison of Pedagogical Practices and Beliefs in International and Domestic Mathematics Teaching Assistants. *Journal of International Students* **2014**, 4 (1), 74-88. Young, R. *Curriculum Renewal in Training Programs for International Teaching Assistants*; 1990.. Ken, N. M.; Karyn, C. O.; Nanda, D.; Debra, L. D. Evaluating the Differential Impact of Teaching Assistant Training Programs on International Graduate Student Teaching. *Canadian journal of higher education (1975)* **2015**, 45 (3), 34.
- (39) Rosslyn M. Smith, P. B., Gayle I. Nelson, Ralph Pat Barrett and Janet C. Constantinides. Crossing Pedagogical Oceans: International Teaching Assistants in U.S. Undergraduate Education, ASHE-ERIC Higher Education Report No. 8, The George Washington University, School of Education and Human Development, Washington, D.C. (1992), p. xix + 122. *Economics of education review* **1994**, 13 (4), 369-370. DOI: 10.1016/S0272-7757(05)80063-9.
- (40) Okoth, E. A.; Mupinga, D. M. An Evaluation of the International Graduate Teaching Assistants Training Program. 2007; p ment.
- (41) Gorsuch, G. J. Improving Speaking Fluency for International Teaching Assistants by Increasing Input. *TESL-EJ* **2011**, 14 (4), EJ, 2011, Vol.2014(2014).
- (42) Olaniran, B. A. International Graduate Teaching Assistants Workshop: Implications for Training. *The College student affairs journal* **1999**, 18 (2), 56.
- (43) Youssef, S. International Teaching Assistant (ITA) Training Program at Bowling Green State University: Putting the Needs of ITAs and the Expectations of Undergraduate Native English-Speaking Students NESSs in Conversation. ProQuest LLC: 2018.
- (44) Ashavskaya, E. International Teaching Assistants' Experiences in the U.S. Classrooms: Implications for Practice. *Journal of the Scholarship of Teaching and Learning* **2015**, 15 (2), 56-69.
- (45) Zhou, J. Managing Anxiety: A Case Study of an International Teaching Assistant's Interaction with American Students. *Journal of International Students* **2014**, 4 (2), 177-190.
- (46) Chalupa; Lair. Meeting the Needs of International TAs in the Foreign Language Classroom: A Model for Extended Training. **2000**.

- (47) Kang, O.; Rubin, D.; Lindemann, S. Mitigating U.S. Undergraduates' Attitudes toward International Teaching Assistants. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* **2015**, 49 (4), 681-706. From EBSCOhost eric.
- (48) Hofer, S. I. Studying Gender Bias in Physics Grading: The role of teaching experience and country. *International Journal of Science Education* **2015**, 37 (17), 2879-2905, Article. DOI: 10.1080/09500693.2015.1114190.
- (49) Jia, C. J.; Bergerson, A. A. Understanding the International Teaching Assistant Training Program: A Case Study at a Northwestern Research University. *International Education* **2008**, 37 (2), 77-98.
- (50) Zhou, J. What Is Missing in the International Teaching Assistants Training Curriculum? *Journal of Faculty Development* **2009**, 23 (2), 19-24.
- (51) Crooks, T. J. An Evaluation of a Program for Developing Teaching Skills: The Campus Teaching Program. *Journal of Educational Technology Systems* **1980**, 9 (2), 95-122. DOI: 10.2190/9BGG-1BM0-AQ45-C8RR.
- (52) Kanaga, K. R. The Evaluation of a Training Program for Undergraduate Teaching Assistants. 1979.
- (53) Bray, J. H.; Howard, G. S. Methodological considerations in the evaluation of a teacher-training program. *Journal of Educational Psychology* **1980**, 72 (1), 62-70. DOI: 10.1037/0022-0663.72.1.62. Stains, M.; Pilarz, M.; Chakraverty, D. Short and Long-Term Impacts of the Cottrell Scholars Collaborative New Faculty Workshop. *Journal of Chemical Education* **2015**, 92 (9), 1466-1476, Article. DOI: 10.1021/acs.jchemed.5b00324. Linenberger, K.; Slade, M. C.; Addis, E. A.; Elliott, E. R.; Mynhardt, G.; Raker, J. R. Training the Foot Soldiers of Inquiry: Development and Evaluation of a Graduate Teaching Assistant Learning Community. *Journal of college science teaching* **2014**, 44 (1), 97-107.
- (54) Beu, F. A. - a Study of Grading by College Teachers. **1955**, (- 12), - 18. Singleton, R.; Smith, E. R. - Does Grade Inflation Decrease the Reliability of Grades? **1978**, - 15 (- 1), - 41. Hills, J. R.; Gladney, M. B. - Factors Influencing College Grading Standards. **1968**, - 5 (- 1), - 39. Lewis, W. A.; Dexter, H. G.; Smith, W. C. - Grading Procedures and Test Validation: A Proposed New Approach. **1978**, - 15 (- 3), - 227. Warren, J. R.; George Washington Univ, W. D. C. E. C. o. H. E. College Grading Practices: An Overview. 1971. Henderson, C.; Yerushalmi, E.; Kuo, V. H.; Heller, P.; Heller, K. Grading student problem solutions: The challenge of sending a consistent message. *American journal of physics* **2004**, 72 (2), 164-169. DOI: 10.1119/1.1634963. DeBoer, B. V.; Anderson, D. M.; Elfessi, A. M. Grading Styles and Instructor Attitudes. *College teaching* **2007**, 55 (2), 57-64. DOI: 10.3200/CTCH.55.2.57-64. Isenhour, M.; Kramlich, G. Holistic Grading: Are all Mistakes Created Equal? *PRIMUS : problems, resources, and issues in mathematics undergraduate studies* **2008**, 18 (5), 441-448. DOI: 10.1080/10511970701624483. Campbell, C. Learning-Centered Grading Practices. *Leadership* **2012**, 41 (5), 30-33. From EBSCOhost ERIC. National Education Association, W. D. C. *What Research Says to the Teacher: Evaluation and Reporting of Student Achievement*, 1974.

- (55) Mutambuki, J.; Fynewever, H. Comparing Chemistry Faculty Beliefs about Grading with Grading Practices. *Journal of chemical education* **2012**, *89* (3), 326-334. DOI: 10.1021/ed1000284. Barnes, L. L. B.; Bull, K. S.; Campbell, N. J.; Perry, K. M. Effects of Academic Discipline and Teaching Goals in Predicting Grading Beliefs among Undergraduate Teaching Faculty. *Research in higher education* **2001**, *42* (4), 455-467. DOI: 10.1023/A:1011006909774. Caldwell, E.; Hartnett, R. - Sex Bias in College Grading? **1967**, - 4 (- 3), - 132. Brookhart, S. M. - Teachers' Grading Practices: Meaning and Values. **1993**, - 30 (- 2), - 142. Ouazad, A. ASSESSED BY A TEACHER LIKE ME: RACE AND TEACHER ASSESSMENTS. *Education Finance and Policy* **2014**, *9* (3), 334-372, Article. Herridge, M.; Talanquer, V. Dimensions of Variation in Chemistry Instructors' Approaches to the Evaluation and Grading of Student Responses. *Journal of Chemical Education* **2021**, *98* (2), 270-280, Article. DOI: 10.1021/acs.jchemed.0c00944. Petcovic, H. L.; Fynewever, H.; Henderson, C.; Mutambuki, J. M.; Barney, J. A. Faculty Grading of Quantitative Problems: A Mismatch between Values and Practice. *Research in Science Education* **2013**, *43* (2), 437-455, Article. DOI: 10.1007/s11165-011-9268-8. Schmelkin, L. P.; Spencer, K. J.; Gellman, E. S. Faculty Perspectives on Course and Teacher Evaluations. *Research in Higher Education* **1997**, *38* (5), 575-592. From EBSCOhost eric. Goubeaud, K.; Yan, W. Teacher educators' teaching methods, assessments, and grading: A comparison of higher education faculty's instructional practices. *The Teacher educator* **2004**, *40* (1), 1-16. DOI: 10.1080/08878730409555348.
- (56) Nyquist, J. D. E.; et al. *Preparing the Professoriate of Tomorrow to Teach. Selected Readings in TA Training*; 1991.
- (57) Chan, S. Applications of Andragogy in Multi-Disciplined Teaching and Learning. *Journal of Adult Education* **2010**, *39* (2), 25-35.
- (58) Knowles, M. S. *The modern practice of adult education; andragogy versus pedagogy*; New York, Association Press, 1970.
- (59) Fogarty, R.; Pete, B. Professional Learning 101: A Syllabus of Seven Protocols. *Phi Delta Kappan* **2009**, *91* (4), 32-34. DOI: 10.1177/003172171009100407.
- (60) DiBiase, W. J.; Wagner, E. P. Aligning General Chemistry Laboratory With Lecture at a Large University. *School science and mathematics* **2002**, *102* (4), 158-171. DOI: 10.1111/j.1949-8594.2002.tb18198.x.
- (61) Johnstone, A. H. Why is science difficult to learn? Things are seldom what they seem. *Journal of computer assisted learning* **1991**, *7* (2), 75-83. DOI: 10.1111/j.1365-2729.1991.tb00230.x. Taber, K. S. Revisiting the chemistry triplet: drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education. *CHEMISTRY EDUCATION RESEARCH AND PRACTICE* **2013**, *14* (2), 156-168. DOI: 10.1039/c3rp00012e.
- (62) Williams, G. M. Examining Classroom Negotiation Strategies of International Teaching Assistants. *International Journal for the Scholarship of Teaching and Learning* **2011**, *5* (1). DOI: 10.20429/ijstl.2011.050121.
- (63) Ma, R. English Communication for International Teaching Assistants. *Journal of International Students (JIS)*: Jonesboro, 2014; Vol. 4, pp 199-201.

- (64) Teixeira-Dias, J. J. C.; de Jesus, H. P.; de Souza, F. N.; Watts, M. Teaching for quality learning in chemistry. *International Journal of Science Education* **2005**, *27* (9), 1123-1137, Article. DOI: 10.1080/09500690500102813. Hatch, D. H.; Farris, C. R. Helping TAs Use Active Learning Strategies. *New Directions for Teaching and Learning* **1989**. From EBSCOhost eric.
- (65) Clarke, E.; Visser, J. Teaching Assistants managing behaviour - who knows how they do it? A review of literature. *Support for Learning* **2016**, *31* (4), 266-280, Review. DOI: 10.1111/1467-9604.12137.
- (66) López Cupita, L. A. Just in Time Teaching: A Strategy to Encourage Students' Engagement. *HOW* **2016**, *23* (2), 89-105. DOI: 10.19183/how.23.2.163. Muzyka, J. L. ConfChem Conference on Flipped Classroom: Just-in-Time Teaching in Chemistry Courses with Moodle. *Journal of Chemical Education* **2015**, *92* (9), 1580-1581. DOI: 10.1021/ed500904y.
- (67) Ruder, S.; Stanford, C.; Gandhi, A. Scaffolding STEM Classrooms to Integrate Key Workplace Skills: Development of Resources for Active Learning Environments. *Journal of College Science Teaching* **2018**, *47* (5), 29-35.
- (68) Witteck, T.; Most, B.; Kienast, S.; Eilks, I. A lesson plan on 'methods of separating matter' based on the Learning Company Approach - A motivating frame for self-regulated and open lab-work in introductory secondary chemistry lessons. *Chemistry Education Research and Practice* **2007**, *8* (2), 108-119, Article.
- (69) Luft, J. A.; Kurdziel, J. P.; Roehrig, G. H.; Turner, J. Growing a Garden without Water: Graduate Teaching Assistants in Introductory Science Laboratories at a Doctoral/Research University. *Journal of Research in Science Teaching* **2004**, *41* (3), 211-233. From EBSCOhost eric.
- (70) Roehrig, G. H.; Luft, J. A.; Kurdziel, J. P.; Turner, J. A. Graduate teaching assistants and inquiry-based instruction: Implications for graduate teaching assistant training. *Journal of Chemical Education* **2003**, *80* (10), 1206-1210.
- (71) Robinson, J. B. New course promoting the facilitation of active learning by new GTAs in introductory chemistry laboratories: Learning through formative assessment and peer review. *Abstracts of Papers of the American Chemical Society* **2000**, *219*, U430-U430, Meeting Abstract.
- (72) Sharpe, R. A Framework for Training Graduate Teaching Assistants. *Teacher Development* **2000**, *4* (1), 131-143. From EBSCOhost eric.
- (73) Lave, J. *Situated learning : legitimate peripheral participation*; Cambridge England ; New York : Cambridge University Press, 1991.
- (74) Herrington, D. G.; Nakhleh, M. B. What defines effective chemistry laboratory instruction? Teaching assistant and student perspectives. *Journal of Chemical Education* **2003**, *80* (10), 1197-1205, Article.

(75) Armenti, A.; Wheeler, G. F. HAWTHORNE EFFECT AND QUALITY TEACHING - TRAINING GRADUATE TEACHING ASSISTANTS TO TEACH. *American Journal of Physics* **1978**, *46* (2), 121-124, Article. DOI: 10.1119/1.11368.

(76) Berg, C. A. R.; Bergendahl, V. C. B.; Lundberg, B. K. S.; Tibell, L. A. E. Benefiting from an Open-Ended Experiment? A Comparison of Attitudes to, and Outcomes of, an Expository versus an Open-Inquiry Version of the Same Experiment. *International Journal of Science Education* **2003**, *25* (3), 351-372. From EBSCOhost eric.

(77) Sandi-Urena, S.; Cooper, M. M.; Gatlin, T. A. Graduate teaching assistants' epistemological and metacognitive development. *Chemistry Education Research and Practice* **2011**, *12* (1), 92-100. DOI: 10.1039/c1rp90012a.

(78) Gatlin, T.; Sandi-Urena, S. Graduate teaching assistants' potential benefits from teaching general chemistry laboratories. *Abstracts of Papers of the American Chemical Society* **2013**, 245. Sandi-Urena, S.; Gatlin, T. A. Effect of facilitating general chemistry laboratories on graduate teaching assistants' development as scientists. *Abstracts of Papers of the American Chemical Society* **2010**, 239.

(79) Park *, C. The graduate teaching assistant (GTA): lessons from North American experience. *Teaching in Higher Education* **2004**, *9* (3), 349-361. DOI: 10.1080/1356251042000216660.

(80) Sorcinelli, M. D.; Elbow, P. *Writing to learn : strategies for assigning and responding to writing across the disciplines*; San Francisco, Calif. : Jossey-Bass, 1997. Wittek, A. L.; Askeland, N.; Aamotsbakken, B. Learning from and about writing: A case study of the learning trajectories of student teachers. *Learning Culture and Social Interaction* **2015**, *6*, 16-28, Article. DOI: 10.1016/j.lcsi.2015.02.001.

(81) Simmons, A. D.; Larios-Sanz, M.; Amin, S.; Rosell, R. C. Using Mini-reports to Teach Scientific Writing to Biology Students. *American Biology Teacher* **2014**, *76* (8), 551-555, Article. DOI: 10.1525/abt.2014.76.8.9. Brillhart, L. V.; Debs, M. B. Teaching Writing--A Scientist's Responsibility. *Journal of College Science Teaching* **1981**, *10* (5), 303-304. From EBSCOhost eric. Farris, C.; Washington Univ, S. C. f. I. D. a. R. Mentor. A Handbook for New Teaching Assistants. Second Edition. 1985. Graham, S.; Harris, K. R.; Macarthur, C. A. IMPROVING THE WRITING OF STUDENTS WITH LEARNING-PROBLEMS - SELF-REGULATED STRATEGY-DEVELOPMENT. *School Psychology Review* **1993**, *22* (4), 656-670. Breland, H. M.; Gaynor, J. L. - A Comparison of Direct and Indirect Assessments of Writing Skill. **1979**, - 16 (- 2), - 128.

(82) Gupta, T. G.; Burke, K. A.; Fetterly, B. F.; Del Carlo, D. D.; Greenbowe, T. J. Student-centered learning in the laboratory: The Science Writing Heuristic (SWH) approach and its impact on students. *Abstracts of Papers of the American Chemical Society* **2011**, 241.

(83) Wackerly, J. W. Stepwise Approach to Writing Journal-Style Lab Reports in the Organic Chemistry Course Sequence. *Journal of Chemical Education* **2018**, *95* (1), 76-83. DOI: 10.1021/acs.jchemed.6b00630.

- (84) Van Bramer, S. E.; Bastin, L. D. Using a Progressive Paper to Develop Students' Writing Skills. *Journal of Chemical Education* **2013**, *90* (6), 745-750. DOI: 10.1021/ed300312q. Carr, J. M. Using a Collaborative Critiquing Technique to Develop Chemistry Students' Technical Writing Skills. *Journal of Chemical Education* **2013**, *90* (6), 751-754. DOI: 10.1021/ed2007982.
- (85) Cho Y., S. S., French D. P. Need assessment for graduate teaching assistant training: identifying important but under-prepared roles,. University, O. S., Ed.; Proceedings of the 2010 Midwest Section Conference of the American Chemical Society for Engineering Education: 2010.
- (86) Puccio, P. M. *TAs Help TAs: Peer Counseling and Mentoring*; 1986.
- (87) Renfrew, M. M.; Moeller, T. Training of teaching assistants in chemistry. A survey. *Journal of Chemical Education* **1978**, *55* (6), 386. DOI: 10.1021/ed055p386. Shulz, R. A. TA Training, Supervision, and Evaluation: Report of a Survey. *ADFL Bulletin* **1980**, *12* (1), 1-8. DOI: 10.1632/adfl.12.1.1. Simpson, R. D.; Smith, K. S. *Validating Teaching Competencies for Graduate Teaching Assistants: A National Study Using the Delphi Method*; Innovative Higher Education, 1993.
- (88) Bomotti, S. S. *Teaching Assistant Attitudes toward College Teaching*; Review of Higher Education, 1994.
- (89) Pelton, J. A. Assessing Graduate Teacher Training Programs: Can a Teaching Seminar Reduce Anxiety and Increase Confidence? *Teaching Sociology* **2014**, *42* (1), 40-49. DOI: 10.1177/0092055X13500029.
- (90) Rodriguez, R. N. Teaching Teaching to Teaching Assistants. *College Teaching* **1985**, *33* (4), 173-176. DOI: 10.1080/87567555.1985.10532315.
- (91) Seals, M.; Hammons, J. O.; Mamiseishvili, K. Teaching Assistants' Preparation for, Attitudes towards, and Experiences with Academic Dishonesty: Lessons Learned. *International Journal of Teaching and Learning in Higher Education* **2014**, *26* (1), 26-36. From EBSCOhost eric.
- (92) Ozuah, P. O. First, There Was Pedagogy And Then Came Andragogy. *Einstein Journal of Biology & Medicine* **2005**, *21* (2), 83-87, Article. From EBSCOhost aph.
- (93) Ingalls, J. D. *A trainers guide to andragogy : its concepts, experience and application*; Washington, D.C. : U.S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, 1973.
- (94) Knowles, M. S. Andragogy: Adult Learning Theory in Perspective. *Community College Review* **1978**.
- (95) Loeng, S. Alexander Kapp - the first known user of the andragogy concept. *International Journal of Lifelong Education* **2017**, *36* (6), 629-643. DOI: 10.1080/02601370.2017.1363826.

- (96) Sato, T.; Haegele, J. A.; Foot, R. Developing Online Graduate Coursework in Adapted Physical Education Utilizing Andragogy Theory. *Quest* **2017**, 69 (4), 453-466. DOI: 10.1080/00336297.2017.1284679.
- (97) Callary, B.; Rathwell, S.; Young, B. W. Alignment of Masters Swim Coaches' Approaches With the Andragogy in Practice Model. *International Sport Coaching Journal* **2017**, 4 (2), 177-190. DOI: 10.1123/iscj.2016-0102.
- (98) Gina, D. Andragogy: A Fundamental Principle of Online Education for Nursing. *Journal of best practices in health professions diversity* **2016**, 9 (2), 1263-1273.
- (99) Murray, A. Andragogy vs. pedagogy. *Teachers matter* **2018**, (38), 32-33. Soares, A. C.; Brauna, R. D. D.; Saraiva, A. ANDRAGOGY: CONTRIBUTIONS FOR ADULT TEACHER LEARNING. *Comunicacoes* **2019**, 26 (3), 23-38. DOI: 10.15600/2238-121X/comunicacoes.v26n3p23-38.
- (100) Birzer, M. L. The theory of andragogy applied to police training. *Policing-an International Journal of Police Strategies & Management* **2003**, 26 (1), 29-42. DOI: 10.1108/13639510310460288.
- (101) Jayakumar, C. MENTORING AND ADULT LEARNING: ANDRAGOGY IN ACTION. *International journal of management research and reviews* **2013**, 3 (5), 2835.
- (102) Bass, C. Learning Theories & Their Application to Science Instruction for Adults. *American Biology Teacher* **2012**, 74 (6), 387-390. DOI: 10.1525/abt.2012.74.6.6.
- (103) Manning, M. Constructivism Does Not Only Happen in the Individual: Sociocultural Theory and Early Childhood Education.(Brief article). *Childhood education* **2006**, 82 (5), 310.
- (104) Price, P. D.; Kugel, R. W. The New AP Chemistry Exam: Its Rationale, Content, and Scoring. *Journal of Chemical Education* **2014**, 91 (9), 1340-1346. DOI: 10.1021/ed500034t.
- (105) Exton. *General Chemistry Lab Manual vol.2*.
- (106) Exton. *Experiments in General Chemistry Lab Manual, 1st ed*.
- (107) Rudd, J. A., II; Greenbowe, T. J.; Hand, B. M. Using the Science Writing Heuristic to Improve Students' Understanding of General Equilibrium. *Journal of Chemical Education* **2007**, 84 (12), 2007-2011. DOI: 10.1021/ed084p2007.
- (108) Poock, J. R.; Burke, K. A.; Greenbowe, T. J.; Hand, B. M. Using the Science Writing Heuristic in the General Chemistry Laboratory to Improve Students' Academic Performance. *Journal of Chemical Education* **2007**, 84 (8), 1371. DOI: 10.1021/ed084p1371.
- (109) Carroll, J. G. Assessing the Effectiveness of a Training Program for the University Teaching Assistants. *Teaching of psychology* **1977**, 4 (3), 135-138. DOI: 10.1207/s15328023top0403_8.

- (110) Nenty, H. J.; Moyo, S.; Phuti, F. Perception of Teaching as a Profession and UB Teacher Trainees' Attitude towards Training Programme and Teaching. *Educational Research and Reviews* **2015**, *10* (21), 2797-2805. From EBSCOhost eric.
- (111) Bernice, A. P.; Melissa, A. M. The Status of Teacher Training in U.S. and Canadian Sociology Departments. *Teaching sociology* **1995**, *23* (4), 341-352. DOI: 10.2307/1319163.
- (112) Blouin, D. D.; Moss, A. R. Graduate Student Teacher Training: Still Relevant (And Missing?) 20 Years Later. *Teaching Sociology* **2015**, *43* (2), 126-136. DOI: 10.1177/0092055X14565516.
- (113) Thornburg, N. A.; Wood, F. E.; Davis, W. E. Keeping Established Teaching Assistant Training Programs Vital: What Does It Take? *Journal of Graduate Teaching Assistant Development* **2000**, *7* (1), 77-83.
- (114) Winternitz, T.; Davis, W. E. Lessons Learned during Five Years of the UC Davis Program in College Teaching. *Journal of Graduate Teaching Assistant Development* **2000**, *7* (1), 69-75.
- (115) Enerson, D. M.; et al. Creating a Community of Teachers: The Penn State Course in College Teaching. 1996.
- (116) Buerkel-Rothfuss, N. L.; Gray, P. L. Graduate Teaching Assistant (GTA) Training: The View from the Top. **1989**.
- (117) Enkin, E. Supporting the Professional Development of Foreign Language Graduate Students: A Focus on Course Development and Program Direction. *Foreign Language Annals* **2015**, *48* (2), 304-320. DOI: 10.1111/flan.12131.
- (118) Meadows, K.; Olsen, K.; Dimitrov, N.; Dawson, D. Evaluating the Differential Impact of Teaching Assistant Training Programs on International Graduate Student Teaching. *The Canadian Journal of Higher Education* **2015**, *45* (3), 34-55.
- (119) Meyer, K.; Mao, Y. Comparing Student Perceptions of the Classroom Climate Created by U.S. American and International Teaching Assistants. *Higher Learning Research Communications* **2014**, *4* (3), 12-22. DOI: 10.18870/hlrc.v4i3.206.
- (120) Ervin, G.; Muyskens, J. A. On Training TAs: Do We Know What They Want and Need? *Foreign Language Annals* **1982**, *15* (5), 335-344. From EBSCOhost eric.
- (121) Abraham, R. G.; Plakans, B. S.; Enright, D. S. Evaluating a Screening/Training Program for NNS Teaching Assistants. *TESOL quarterly* **1988**, *22* (3), 505-508. DOI: 10.2307/3587294. Constantinides, J. C. ITA training programs. *New Directions for Teaching and Learning* **1989**, *1989* (39), 71-77. DOI: 10.1002/tl.37219893908.
- (122) Parsons, J. S.; Texas Univ, A. R. a. D. C. f. T. E. Anxiety Self Report (ASR (1,2,3,4,). X. 1973.

- (123) Abraham, R. G.; Plakans, B. S. Evaluating a Screening/Training Program for NNS Teaching Assistants. *TESOL Quarterly* **1988**, 22 (3), 505-508. DOI: 10.2307/3587294.
- (124) Young, R. Curriculum Renewal in Training Programs for International Teaching Assistants. *Journal of Intensive English Studies* **1990**, 4 (spring-fall), 59-77.
- (125) Bateman, W. L. *Open to question : the art of teaching and learning by inquiry*, San Francisco : Jossey-Bass, 1990. Volkert, D. Inquiry Based Learning. *Nevada RNformation* **2012**, 21 (3), 15.
- (126) Randall, D. A. C.; Moore, C.; Carvalho, I. S. An international collaboration to promote inquiry-based learning in undergraduate engineering classrooms. *Campus-wide information systems* **2012**, 29 (4), 259-271. DOI: 10.1108/10650741211253859.
- (127) Hughes, P. W.; Ellefson, M. R. Inquiry-based training improves teaching effectiveness of biology teaching assistants. *PloS one* **2013**, 8 (10), e78540-e78540. DOI: 10.1371/journal.pone.0078540. Wheeler, L. B.; Maeng, J. L.; Whitworth, B. A. Characterizing Teaching Assistants' Knowledge and Beliefs Following Professional Development Activities within an Inquiry Based General Chemistry Context. *Journal of Chemical Education* **2017**, 94 (1), 19-28, Article. DOI: 10.1021/acs.jchemed.6b00373.
- (128) Koenig, K.; Schen, M.; Edwards, M.; Bao, L. Addressing STEM Retention through a Scientific Thought and Methods Course. *Journal of College Science Teaching* **2012**, 41 (4), 23-29. From EBSCOhost eric.
- (129) Nesselroade, J. R., and Baltes, Paul B. *Longitudinal research in the study of behavior and development*; New York : Academic Press, 1979.
- (130) Curran, P. J.; Obeidat, K.; Losardo, D. Twelve Frequently Asked Questions About Growth Curve Modeling. *Journal of cognition and development* **2010**, 11 (2), 121-136. DOI: 10.1080/15248371003699969.
- (131) McCoach, D. B.; O'Connell, A. A.; Reis, S. M.; Levitt, H. A. Growing Readers: A Hierarchical Linear Model of Children's Reading Growth During the First 2 Years of School. *Journal of Educational Psychology* **2006**, 98 (1), 14-28. From EBSCOhost ERIC.
- (132) King, R. B.; McInerney, D. M. Mapping changes in students' English and math self-concepts: a latent growth model study. *Educational psychology (Dorchester-on-Thames)* **2014**, 34 (5), 581-597. DOI: 10.1080/01443410.2014.909009.
- (133) Tutwiler, M. S.; McCoach, D. B.; Hamilton, R.; Siegle, D. Trends in Reading Growth between Gifted and Nongifted Students: An Individual Growth Model Analysis. AERA Online Paper Repository: 2017.

(134) Seltzer, M.; Choi, K.; Thum, Y. M. Examining Relationships between Where Students Start and How Rapidly They Progress: Using New Developments in Growth Modeling to Gain Insight into the Distribution of Achievement within Schools. *Educational Evaluation and Policy Analysis* **2003**, *25* (3), 263-286. From EBSCOhost ERIC. Owens, E. B.; Shaw, D. S. Predicting Growth Curves of Externalizing Behavior Across the Preschool Years. *Journal of abnormal child psychology* **2003**, *31* (6), 575-590. DOI: 10.1023/A:1026254005632.

(135) Huttenlocher, J.; Haight, W.; Bryk, A.; Seltzer, M.; Lyons, T. Early Vocabulary Growth: Relation to Language Input and Gender. *Developmental psychology* **1991**, *27* (2), 236-248. DOI: 10.1037/0012-1649.27.2.236.

(136) Schumann, A.; Stein, J. A.; Ullman, J. B.; John, U.; Rumpf, H.-J.; Meyer, C. Patterns and Predictors of Change in a Smoking Intervention Study: Latent Growth Analysis of a Multivariate Outcome Model. *Health psychology* **2008**, *27* (3S), S233-S242. DOI: 10.1037/0278-6133.27.3(Suppl.).S233.

(137) Ferreira, P. C.; Simao, A. M. V.; da Silva, A. L. Does training in how to regulate one's learning affect how students report self-regulated learning in diary tasks? *Metacognition and Learning* **2015**, *10* (2), 199-230, Article. DOI: 10.1007/s11409-014-9121-3.

(138) Mehtatalo, L.; Peltola, H.; Kilpelainen, A.; Ikonen, V.-P. The Response of Basal Area Growth of Scots Pine to Thinning: A Longitudinal Analysis of Tree-Specific Series Using a Nonlinear Mixed-Effects Model. *Forest science* **2014**, *60* (4), 636-644. DOI: 10.5849/forsci.13-059.

(139) Rossi, A.; Marconi, M.; Di Lucca, G.; Morena, R.; Pogliani, C.; Parati, M. C.; Rossini, C.; Verusio, C. Exploring the form of change: A latent growth curve model of distress in oncologic patients undergoing specific cancer-related psychological intervention—A preliminary study. *Journal of clinical oncology* **2017**, *35* (15_suppl), e21556-e21556. DOI: 10.1200/JCO.2017.35.15_suppl.e21556. Yao, Z.; Enright, R. A longitudinal analysis of social class and adolescent prosocial behaviour: a latent growth model approach (Un análisis longitudinal de clase social y conducta prosocial adolescente: un enfoque de modelo de crecimiento latente). *Revista de psicología social ahead-of-print* (ahead-of-print), 1-29. DOI: 10.1080/02134748.2022.2034292.

(140) Wareham, J.; Dembo, R. A Longitudinal Study of Psychological Functioning Among Juvenile Offenders: A Latent Growth Model Analysis. *Criminal justice and behavior* **2007**, *34* (2), 259-273. DOI: 10.1177/0093854806289828.

(141) van der Veen, I.; Peetsma, T. The development in self-regulated learning behaviour of first-year students in the lowest level of secondary school in the Netherlands. *Learning and Individual Differences* **2009**, *19* (1), 34-46, Article. DOI: 10.1016/j.lindif.2008.03.001. Francis, D. J.; Shaywitz, S. E.; Stuebing, K. K.; Shaywitz, B. A.; Fletcher, J. M. Developmental Lag Versus Deficit Models of Reading Disability: A Longitudinal, Individual Growth Curves Analysis. *Journal of educational psychology* **1996**, *88* (1), 3-17. DOI: 10.1037/0022-0663.88.1.3.

- (142) Peng, Y.; Lord, D. Application of Latent Class Growth Model to Longitudinal Analysis of Traffic Crashes. *Transportation research record* **2011**, 2236 (2236), 102-109. DOI: 10.3141/2236-12.
- (143) Kwon, J.; Lee, H. Why travel prolongs happiness: Longitudinal analysis using a latent growth model. *Tourism management (1982)* **2020**, 76, 103944. DOI: 10.1016/j.tourman.2019.06.019.
- (144) Galloway, K. R.; Bretz, S. L. Measuring Meaningful Learning in the Undergraduate General Chemistry and Organic Chemistry Laboratories: A Longitudinal Study. *Journal of chemical education* **2015**, 92 (12), 2019-2030. DOI: 10.1021/acs.jchemed.5b00754.
- (145) Lee, V. E. Using Hierarchical Linear Modeling to Study Social Contexts: The Case of School Effects. *Educational psychologist* **2000**, 35 (2), 125-141. DOI: 10.1207/S15326985EP3502_6.
- (146) Duffin, E. Digest of Education Statistics 2020, table 318.10. NCES: July 2020.
- (147) Dewey, J. *Experience and education*; New York : Macmillan, 1938.
- (148) Price, J.; Cotten, S. R. Teaching, Research, and Service: Expectations of Assistant Professors. *The American sociologist* **2006**, 37 (1), 5-21. DOI: 10.1007/s12108-006-1011-y.
- (149) Wigfield, A.; Eccles, J. S. Expectancy–Value Theory of Achievement Motivation. *Contemporary educational psychology* **2000**, 25 (1), 68-81. DOI: 10.1006/ceps.1999.1015.
- (150) Leaper, C. More similarities than differences in contemporary theories of social development?: a plea for theory bridging. *Advances in child development and behavior* **2011**, 40, 337-378.
- (151) Canales-Negron, I. Using digital badges design principles in professional continuing education programs: a scoping review. *Revista Digital De Investigacion En Docencia Universitaria-Ridu* **2020**, 14 (2). DOI: 10.19083/ridu.2020.1170.
- (152) Abramovich, S.; Schunn, C.; Higashi, R. M. Are badges useful in education?: it depends upon the type of badge and expertise of learner. *Educational technology research and development* **2013**, 61 (2), 217-232. DOI: 10.1007/s11423-013-9289-2. McDaniel, R.; Lindgren, R.; Friskics, J. Using badges for shaping interactions in online learning environments. 2012, IEEE: pp 1-4. DOI: 10.1109/IPCC.2012.6408619.
- (153) Kopcha, T. J.; Ding, L.; Neumann, K. L.; Choi, I. Teaching Technology Integration to K-12 Educators: A 'Gamified' Approach. *TechTrends* **2016**, 60 (1), 62-69. DOI: 10.1007/s11528-015-0018-z.
- (154) Foster, J. C. The promise of digital badges.(The 21st-Century Classroom)(displaying educational achievements to potential employers). *Techniques - Association for Career and Technical Education* **2013**, 88 (8), 30.

(155) Hakulinen, L.; Auvinen, T.; Korhonen, A. Empirical Study on the Effect of Achievement Badges in TRAKLA2 Online Learning Environment. 2013, IEEE: pp 47-54. DOI: 10.1109/LaTiCE.2013.34.

(156) Askeroth, J. H.; Newby, T. J. Digital Badge Use in Specific Learner Groups. *International journal of innovative teaching and learning in higher education* **2020**, *1* (1), 1-15. DOI: 10.4018/IJITLHE.2020010101. Hensiek, S.; DeKorver, B. K.; Harwood, C. J.; Fish, J.; O'Shea, K.; Towns, M. Digital Badges in Science: A Novel Approach to the Assessment of Student Learning. *Journal of college science teaching* **2017**, *46* (3), 28-33. DOI: 10.2505/4/jcst17_046_03_28. Carey, K. L.; Stefaniak, J. E. An exploration of the utility of digital badging in higher education settings. *Educational technology research and development* **2018**, *66* (5), 1211-1229. DOI: 10.1007/s11423-018-9602-1. Besser, E. D.; Newby, T. J. Feedback in a Digital Badge Learning Experience: Considering the Instructor's Perspective. *TechTrends* **2020**, *64* (3), 484-497. DOI: 10.1007/s11528-020-00485-5. Jones, W. M.; Hope, S.; Adams, B. Teachers' perceptions of digital badges as recognition of professional development: Teachers' perceptions of digital badges. *British journal of educational technology* **2018**, *49* (3), 427-438. DOI: 10.1111/bjet.12557.

(157) Facey-Shaw, L.; Specht, M.; Van Rosmalen, P.; Borner, D.; Bartley-Bryan, J. Educational Functions and Design of Badge Systems: A Conceptual Literature Review. *Ieee Transactions on Learning Technologies* **2018**, *11* (4), 536-544. DOI: 10.1109/tlt.2017.2773508.

(158) Immorlica, N.; Stoddard, G.; Syrgkanis, V. Social Status and Badge Design. In *International World Wide Web Conference*, 2015, ACM: pp 473-483. DOI: 10.1145/2736277.2741664.

(159) Easley, D.; Ghosh, A. Incentives, gamification, and game theory: an economic approach to badge design. In *Electronic Commerce*, 2013, ACM: pp 359-376. DOI: 10.1145/2482540.2482571.

(160) Denny, P. The effect of virtual achievements on student engagement. In *Conference on Human Factors in Computing Systems*, 2013, ACM: pp 763-772. DOI: 10.1145/2470654.2470763.

(161) Seery, M. K.; Agustian, H. Y.; Doidge, E. D.; Kucharski, M. M.; O'Connor, H. M.; Price, A. Developing laboratory skills by incorporating peer-review and digital badges. *CHEMISTRY EDUCATION RESEARCH AND PRACTICE* **2017**, *18* (3), 43-419. DOI: 10.1039/c7rp00003k.

(162) Hennah, N.; Seery, M. K. Using Digital Badges for Developing High School Chemistry Laboratory Skills. *Journal of chemical education* **2017**, *94* (7), 844-848. DOI: 10.1021/acs.jchemed.7b00175.

(163) Towns, M.; Harwood, C. J.; Robertshaw, M. B.; Fish, J.; O'Shea, K. The Digital Pipetting Badge: A Method To Improve Student Hands-On Laboratory Skills. *Journal of chemical education* **2015**, *92* (12), 2038-2044. DOI: 10.1021/acs.jchemed.5b00464.

- (164) Hensiek, S.; DeKorver, B. K.; Harwood, C. J.; Fish, J.; O'Shea, K.; Towns, M. Improving and Assessing Student Hands-On Laboratory Skills through Digital Badging. *Journal of chemical education* **2016**, *93* (11), 1847-1854. DOI: 10.1021/acs.jchemed.6b00234.
- (165) Roberts, A.; O'Neil, A. L.; Barlow, R.; Keeler, C.; Nelligan, W. S.; Westmoreland, T. D. Chemical Hygiene and Safety Badge for an Upper-Level Laboratory Course. *Journal of chemical education* **2021**, *98* (1), 143-149. DOI: 10.1021/acs.jchemed.0c00118.
- (166) Hill, M. A.; Overton, T.; Kitson, R. R. A.; Thompson, C. D.; Brookes, R. H.; Coppo, P.; Bayley, L. 'They help us realise what we're actually gaining': The impact on undergraduates and teaching staff of displaying transferable skills badges. *Active learning in higher education* **2022**, *23* (1), 17-34. DOI: 10.1177/1469787419898023.
- (167) Matz, R. L.; Rothman, E. D.; Krajcik, J. S.; Banaszak Holl, M. M. Concurrent enrollment in lecture and laboratory enhances student performance and retention. *Journal of research in science teaching* **2012**, *49* (5), 659-682. DOI: 10.1002/tea.21016. Westbrook, S. L.; Rogers, L. N. Doing Is Believing: Do Laboratory Experiences Promote Conceptual Change? *School science and mathematics* **1996**, *96* (5), 263-271. DOI: 10.1111/j.1949-8594.1996.tb10239.x. George-Williams, S. R.; Ziebell, A. L.; Kitson, R. R. A.; Coppo, P.; Thompson, C. D.; Overton, T. L. 'What do you think the aims of doing a practical chemistry course are?' A comparison of the views of students and teaching staff across three universities. *CHEMISTRY EDUCATION RESEARCH AND PRACTICE* **2018**, *19* (2), 463-473. DOI: 10.1039/c7rp00177k.
- (168) Box, M. C.; Dunnagan, C. L.; Hirsh, L. A. S.; Cherry, C. R.; Christianson, K. A.; Gibson, R. J.; Wolfe, M. I.; Gallardo-Williams, M. T. Qualitative and Quantitative Evaluation of Three Types of Student-Generated Videos as Instructional Support in Organic Chemistry Laboratories. *Journal of chemical education* **2017**, *94* (2), 164-170. DOI: 10.1021/acs.jchemed.6b00451. Erdmann, M. A.; March, J. L. Video reports as a novel alternate assessment in the undergraduate chemistry laboratory. *CHEMISTRY EDUCATION RESEARCH AND PRACTICE* **2014**, *15* (4), 650-657. DOI: 10.1039/c4rp00107a.