August 2013

Studies in Analytical Chemistry and Chemical Education. Part 1: Characterization of Complex Organics By Raman Spectroscopy and Gas Chromatography. Part 2: Differential Item Functioning on Multiple-choice General Chemistry Assessments

Lisa Kay Kendhammer
University of Wisconsin-Milwaukee

STUDIES IN ANALYTICAL CHEMISTRY AND CHEMICAL EDUCATION.

PART 1: CHARACTERIZATION OF COMPLEX ORGANICS BY RAMAN SPECTROSCOPY AND GAS CHROMATOGRAPHY.

PART 2: DIFFERENTIAL ITEM FUNCTIONING ON MULTIPLE-CHOICE GENERAL CHEMISTRY ASSESSMENTS.

by

Lisa K. Kendhammer

A Dissertation Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy in Chemistry

at

The University of Wisconsin-Milwaukee

August 2013

ABSTRACT

STUDIES IN ANALYTICAL CHEMISTRY AND CHEMICAL EDUCATION.

PART 1: CHARACTERIZATION OF COMPLEX ORGANICS BY RAMAN SPECTROSCOPY AND GAS CHROMATOGRAPHY.

by

Lisa Kendhammer

The University of Wisconsin-Milwaukee, 2013

Under the Supervision of Professor Joseph H. Aldstadt, III

The analytical chemistry component of this thesis focused on instrumentation and methods to address challenges in art conservation, particularly the identification, quantitation, and reactivity of a set of representative varnishes and their degradation products. Methods for characterizing varnishes are of great interest to art conservators to restore art work more accurately. A database was created as a means to identify and quantify the composition of aged varnishes. Fourier Transform (FT)-Raman Spectroscopy was used to study common organic acids found in varnishes. The database included nine short-chain carboxylic acids, four di-carboxylic acids, and six medium-to-long-chain fatty acids. Four varnish samples (Linseed Oil, Tung Oil, Dammar, and Mastic) were studied as well. Through visual comparison and fingerprinting analysis comparison, identification of components in the Raman Spectral Database were

recognized as components of the varnish samples. Singular Value Decomposition (SVD) was conducted to determine how well the database represented the unknown varnish samples. SVD was applied to the 19 standards collected in building the database. To reduce the amount of data, seven singular values were chosen. The seven singular values were then used to model several unknowns – Linseed Oil, Tung Oil, Dammar, and Mastic. The root-mean square (RMS) error for the unknowns were 0.08, 0.13, 0.21, and 0.21 Raman Intensity units, for Linseed Oil, Tung Oil, Dammar, and Mastic, respectively. If those values are compared to the largest peak in the unknown spectra, the % relative RMS errors are 1.7%, 1.7%, 4.9%, and 6.4%, respectively.

A method based upon Gas Chromatography (GC) was developed to characterize carboxylic acids formed as a result of varnish degradation. In this method, a headspace solid-phase microextraction (SPME) approach was optimized in which a 75 µm carboxen-polydimethylsiloxane (CAR/PDMS) SPME fiber was used to analyze mono carboxylic acids. For quantitative determinations, the injection port was in the splitless mode and held at 250°C for 1.0 min for the desorption of the analytes from the SPME fiber. After the initial minute, the injector was switched to a 1:100 split ratio. The temperature program consisted of the oven being initially set to a temperature of 30°C and held for 1 min, and then ramped at 25°C/min to 200°C, where the temperature was held for 1 min, thereby resulting in a total run time of 8.80 min. The PFPD was held at 200 °C for the entire run with a 0.5 ms gate delay, and the gate width was set to 20.0 ms. The mono carboxylic acids that were studied were Formic, Acetic, Propanoic, Butyric, Valeric, and Caproic Acid. A linear relationship was observed between the number of carbons in the carboxylic acid and the retention time ($y = 0.75x + 1.55$, R^2=0.95).

Quantitation of Acetic Acid was done by calibration using a first-order regression fit. The model yielded: $y = 0.29x + 0.92$ (R^2=0.95). Using a second-order model, a better fit was found: $y = 0.0025x^2 - 0.0016x + 5.9$ (R^2=0.99).

An ageing chamber was designed, fabricated, and tested as a means for better understanding the decomposition of varnishes over time as a function of temperature, humidity, and ultraviolet light. The goal in the development of the ageing chamber was to demonstrate that it may be possible to create Standard Reference Materials (SRMs) artificially that resemble authentically aged varnishes. This is possible by the use of the ageing chamber that was built because it is directly incorporated into a GC oven where temperature, where UV radiation, humidity levels, and pollutants can be precisely controlled and carefully monitored. The GC method for carboxylic acids described above was developed to aid in the measurement of carboxylic acid fragments that could arise from the ageing process. There are promising results of the Raman Intensity increasing as the sample aged.

ABSTRACT

STUDIES IN ANALYTICAL CHEMISTRY AND CHEMICAL EDUCATION.

PART 2: DIFFERENTIAL ITEM FUNCTIONING ON MULTIPLE-CHOICE GENERAL CHEMISTRY ASSESSMENTS.

by

Lisa Kendhammer

The University of Wisconsin-Milwaukee, 2013

Under the Supervision of Professor Kristen Murphy

Over the past 30 years, there have been a plethora of studies on gender differences. Some of the earlier studies found that male students typically outperform female students in visual-spatial and quantitative abilities, whereas female students outperform male students in verbal abilities. In later studies it was reinforced that female students still tended to outperform male students in verbal abilities while the gap in science and mathematics (the latter as an extension of visual-spatial and quantitative abilities) closed greatly. During this same time, more female students entered the science, technology, engineering, and mathematics (STEM) fields. In 1966, only 25% of all STEM bachelor's degrees were obtained by female students, whereas in 2010 that percentage had grown to 50%. Specifically in chemistry, 49.9% of the bachelor's degrees were earned by women compared to the 18.5% in 1966.[1] With assessments as a

large source of the student's overall course grade, it is imperative that those assessments be valid and unbiased. One way to determine this is to use Differential Item Functioning (DIF). DIF occurs when subgroups of equal abilities perform statistically different on an item on an assessment where typically students that are matched with equivalent ability would have an equivalent possibility of answering the question on the assessment correctly. Because of the difficulty in determining students' ability often times the subgroups are matched on their proficiency or the score they received on an assessment.

This dissertation focused on four main questions. The first question focused on identifying items that exhibited DIF. The second question was to determine if DIF was real, i.e. did it persist no matter the set of students or the matching criteria used? The third question focused on determining the causes of DIF by cloning the items by content and construct (format). Lastly, it was hypothesized that one of the reasons behind why DIF is happening was due to the students' problem-solving process and examining these through the use of incorrect heuristics.

Data for the first part of the study was collected from two American Chemical Society-Examinations Institute (ACS-EI) trial tests (Form A and Form B) that were given to students who had completed one term of general chemistry. This data was analyzed using the Mantel-Haenszel statistic to determine which items exhibited possible DIF. Along with the Mantel-Haenzel statistic a two stage DIF analysis[2] was conducted. Out of the 140 items, 33 exhibited DIF. On Form A there were 14 items which exhibited DIF, seven that favored male students and seven that favored female students. On Form B there were 19 items which exhibited DIF, 11 that favored female students and eight that favored male students. Those items that exhibited the highest probability of DIF were

cloned and included on hourly examinations. These items were examined for DIF persistence against both stages of the two-stage analysis and other relevant measures of proficiency. As more results were collected, patterns emerged for persistent DIF items. On the 24 hourly examinations that were included in this analysis, there were a total of 687 items: 33 (5%) had a significant value using the Mantel-Haenszel statistic, thereby exhibiting persistent DIF. Of those 33 items, 15 were flagged with persistent DIF that favored female students and 18 were flagged with persistent DIF that favored male students. On the three standardized examinations, there were a total of 140 items; 19 (14%) had a significant value using the Mantel-Haenszel statistic, thereby exhibiting persistent DIF. Of those 19 items, two of the items that were flagged with persistent DIF favored female students and 17 of the items that were flagged with persistent DIF favored male students.

Along with these items, certain content areas and formats of the items were found to favor one gender. Over six semesters of testing, the content areas that consistently showed DIF that favored male students were measurement (density), greatest/least number of atoms, limiting reagents, ideal gas equation, and crystal structures; the content areas that favored female students were nomenclature and molecular orbital theory. The formats that tended to favor male students were visual-spatial, reasoning, and computation; the format that favored female students was specific chemical knowledge. By cloning these items, it was found that some of the possible causes of persistent DIF for certain items were the content and/or the format.

Lastly semi-structured interviews were conducted and it was found that for seven items the possible reason why DIF was happening was due to one subgroup using an

incorrect heuristic. These items were in the specific content areas of measurement (density), greatest/least number of atoms, stoichiometry-general, and crystal structures. Additionally, the format inclusions of visual-spatial, reasoning, and computation for these items could also be contributing factors to the observed results.

References

1. S&E Degrees: 1966-2010: National Center for Science and Engineering Statistics. http://www.nsf.gov/statistics/nsf11316/content.cfm?pub_id=4062&id=2 (accessed May 26).

2. Zenisky, A. L.; Hambleton, R. K., Detection of Differential Item Functioning in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach. *Educational and Psychological Measurement* **2003a,** *63* (1), 51-64.

This work is dedicated to my family for all of the love and support

Table of Contents

List of Figures

Analytical Chemistry Section

Chapter 1:

Chapter 2

Chapter 3

Chemical Education Section

Chapter 2

Chapter 3

List of Tables

Analytical Chemistry Section

Chapter 2

Chapter 3

Chemical Education Section

Chapter 3

Acknowledgements

There are so many people I'd like to thank for their help and guidance. First and foremost I'd like to thank my two advisors, Joseph H. Aldstadt, III and Kristen Murphy. Their knowledge, guidance, and patience lead me and my projects to where they are today. Thank you for all your time and dedication.

Secondly, I'd like to thank my committee, Joseph H. Aldstadt, III, Kristen Murphy, Mark Dietz, Andrew Pacheco, and April Zenisky. These group of individuals helped me to explore new avenues and strengthen my projects.

Next I'd like to thank Neil Korfhage, from the UWM glass shop, for his superb work and craftsmanship with the ageing chamber, the glass coupons, and the Raman sample tubes. His work is truly outstanding.

Another huge component of my work is the people who have supported me along the way. Specifically my past and present group member, Dr. Diab Qadah, Dr. Beth Ruddy, Dr. John Frost, Dr. Steven Kopitzke, Dr. Megan McCallum, Scott Schlipp, April Grant, Ryan Schmeling, Veronica Marco Alvarez, Anahit Campbell, Erin O'Connell, Karrie Gerlach, Jaclyn Trate and Shalini Srinivasan. Also I'd like to thank many of my classmates.

Finally, I'd like to thank my family and friends for all their help and support. Thank you for believing in me, and all the reassurances you gave. You are all priceless.

Chapter 1: Introduction

In this chapter, background information on the nature of the problem studied in art conversation is presented, including a discussion of the ageing and characterization of artistic material. The structure of the research study is then described, followed by an examination of the theory and practice of the key analytical techniques that were employed.

1.1 Background

In 2009, a workshop entitled “Chemistry and Materials Research at the Interface Between Science and Art” was co-sponsored by the National Science Foundation and the Andrew W. Mellon Foundation [1]. This workshop brought together scientists from industry, academia, museums, and many other areas. The purpose of the workshop was to identify aspects of Art Conservation in which "knowledge gaps" existed, and to propose scientific solutions to fill these gaps. The report of the workshop outlined the areas of research in which more study is needed.

Three main areas were classified as “Grand Challenges.” The first "Grand

Challenge" focused on the characterization of materials and the specific chemical structures present, the second described the material degradation and ageing process, and the third emphasized the strengthening, stabilization, monitoring, and repair of materials. Of the many areas where "knowledge gaps" were identified, several main points stood out. The first issue was that there are not simple solutions to these problems, because the chemistry of artistic media is both diverse and complex. Furthermore, the specific materials that were used (i.e., the artist's "recipes") are unknown. The second issue was that unlike other areas in chemistry, there are no Standard Reference Materials (SRMs) for artistic media, so researchers are unable to validate their methods. Lastly, in an effort to understand ageing, many people have looked at individual environmental factors that contribute to ageing, but few have looked at all of the environmental factors simultaneously [1].

Based upon the report from this workshop, the Conservation Laboratory at the Milwaukee Art Museum (MAM) approached the Aldstadt research group with an issue of particular concern. They recently obtained some pieces of decorative art by George Mann Niedecken, who worked with Frank Lloyd Wright [2]. Two of the pieces were chairs that had varnish that was cracking and yellowing – and therefore needed to undergo restoration. The conservationists at MAM wanted to know the composition of the varnish that was originally applied to the chairs. One of the common misconceptions with varnish is that it is applied solely for protection. While it does certainly protect the material, the main reason for varnish application to pieces of art (not only furniture but also to painted surfaces) is for aesthetic reasons. The varnish makes the colors of the underlying material more saturated and provides an overall gloss [3]. Thus in restoring a

piece of art, the conservationist ideally would like to know the original varnish material that was used so that the aesthetic intent of the artist can be conserved.

1.1.1 Varnish Ageing

The ageing of artistic material is a major concern in art conservation. With varnishes there is always the chance of the varnish yellowing, cracking, and flaking, which in turn can change the appearance of the object [3]. The mission of the art conservationist is to accurately restore these deteriorating objects, with the application of a new varnish often as the most challenging task. To maintain the integrity of the work of art, ideally the conservationist would strive to apply a new varnish that is identical to the original varnish [3]. Unlike today where there are many different varnishes from which to choose, along with detailed knowledge of the chemical composition of each varnish, art work created before ~1950 is often of mysterious composition. Artists generally did not keep detailed records of their formulations ("recipes") and often – depending on what geographical region or era in which the object was created – the varnishes had different compositions [4,5].

Tasked with the effort of reapplying a varnish that closely matches the original varnish, it is important for art conservators to combine their knowledge with that of the scientist [6]. One way to determine the composition of the original varnish is to study the ageing process of the varnish. There are three main goals when studying the composition of artistic media: (a) determining the mechanism of the degradation process along with identifying causal factors (e.g., heat, light, moisture, etc.), (b) determining the physical

differences between the original and aged media, and (c) determining if previous restoration(s) has been done to the piece [6]. All of these goals revolve around how the varnish has changed and determining the most effectively authentic means with which to restore it. The challenge to the analytical chemist is answering the conservator's question: not "What *is* it?" but rather "What *was* it?"

1.1.2 Varnish Compositions

While there are many types of varnishes applied to art work, triglycerides and triterpenoids are the two most common classes [3,7-9]. As shown in Figure 1, triglycerides have medium- and long-chain fatty acids attached to a glycerol backbone, whereas triterpenoids are made up of units of isoprene in either chains or rings. Upon oxidation of the triglycerides, carboxylic acids are formed as shown in the reaction scheme in Figure 2. Triglycerides, triterpenoids, and many other types of artistic media contain carboxylic acids [7]. Not surprisingly, studies have shown that carboxylic acids are the common products of varnish degradation [10]. Zumbul *et al.*, found that the artificial light-induced ageing of Dammar and Mastic, commonly used triterpenoid varnishes, led to the production of small carboxylic acids, specifically propanoic acid, which would probably polymerize with the parent triterpenoid and thereby change the overall polarity of the varnish [10]. Carboxylic acids are commonly found in artistic material, particularly in varnishes, thereby making them an ideal class of analytes to examine in more detail.

1.1.3 Approach

The goal of this study was to develop the instrumentation and methods needed to study the ageing process in representative varnishes. This entailed the development of analytical methods for characterizing native varnishes (i.e., medium- and long-chain fatty acids) and degradation products (i.e., small carboxylic acids) as well as a testing apparatus to simulate the ageing process. Therefore, there are three main parts to this dissertation: (a) to construct a Raman spectral database of large carboxylic acids to gain more insight into the composition of the complex organic molecules that are found in varnishes; (b) to develop a method using Gas Chromatography (GC) to characterize the small carboxylic acids that are created during the ageing process in varnishes; and (c) to design, fabricate and test an ageing chamber as a means to study the various environmental factors that affect the ageing process of varnishes.

Several databases have been reported, including those for biological materials [11,12], pharmaceutical excipients [13], semiprecious stones [14], porcelain [15], and pigments [16]. Raman Spectroscopy was chosen because of its high sensitivity, selectivity, and high resolution – characteristics that are well-suited for generating high spectral detail at low concentration. Raman spectroscopy has in fact been used to study the ageing of varnishes because the high spectral detail can be used before to identify changes in naturally aged material [17].

To apply the database to varnish unknowns, Singular Value Decomposition (SVD) was applied to mathematically represent the spectral data and compare known standards to unknown varnish samples. SVD is a powerful tool for pattern recognition

studies – it has been used widely as a sensitive measure of often subtle differences in the comparison of complex samples, such as image recognition and palmprint identification [18,19]. However, the application of SVD to the study of artistic materials has been relatively limited [20,21]. This is somewhat surprising, given that SVD would provide an in-depth comparison not only of fresh and aged samples, but also in determining the differences between aged varnishes of similar compositions. In a study conducted by Brissaud *et al.*, SVD was used to determine thicknesses and layer sequences of different types of paints that had been applied to an easel painting [20]. A study by the Rochester Institute of Technology's Munsell Color Science Laboratory used SVD to track color management for a new camera system that was applied to imaging decorative art [21]. While both of these studies illustrate the utility of SVD applied to studies of cultural heritage objects commonly encountered in the art conservation laboratory, one would expect SVD to be even more successful in the comparison of Raman spectra because of the high spectral detail of the spectra and the pattern recognition ability of the SVD technique. Thus in the work reported herein, SVD was explored as a novel means to compare the Raman spectra of standards and unknown varnishes.

The second part of the study consisted of developing a GC method for the characterization of the carboxylic acids generated in the ageing studies. The GC method development consisted of optimizing the key factors for the method and then conducting qualitative and quantitative analysis of carboxylic acid standards. Several studies have been reported for the characterization of artistic material by GC, usually with mass spectrometric (MS) detection [22,23]. Artistic samples that have been analyzed include paint samples from ancient Coptic icons (images painted on wooden supports) [24], oils

and oil binders found in painting samples [25], and varnishes on musical instruments [26] to name just a few.

The last phase of the study was to design, fabricate and test an ageing chamber to create SRMs and to study how various environmental factors (e.g., heat, light, moisture) affect the ageing process of varnishes. The goal of the ageing chamber development was to address some of the issues identified in the 2009 workshop discussed earlier, as well as giving chemists who study the composition and ageing of varnishes an opportunity to validate their methods by the use of SRMs.

1.2 Analytical Techniques Used in the Study

1.2.1 Raman Spectroscopy

When creating a database of carboxylic acids to better understand their composition, there are many choices of analytical techniques. For this project, it was important to use a technique that would give great detail for the comparison of standards and unknowns. Vibrational spectroscopies are obvious choices for providing high-resolution, detailed spectra of molecular structures. Additionally, the non-destructive nature is a key advantage in the measurement of artistic media. The greater sensitivity of Raman spectroscopy, and the facility with which it can handle aqueous phase samples, made it the preferred choice for this work.

1.2.1.1 Background

Raman spectroscopy was first described by C.V. Raman in 1928, for which he was awarded the 1931 Nobel Prize in Physics [27-29]. However, it was not until the 1980s that the technique became widespread because of the improvement of many of the components of the Raman spectrometer [30]. Some of these improvements included a longer wavelength laser to avoid interfering fluorescence and sample decomposition that plague higher energy sources, more efficient filters to remove Rayleigh (elastic) scattering, the use of an interferometer for better frequency precision and a multiplexing effect, and more sensitive detectors [27,28,30,31]. All of this became manifest as the Fourier Transform (FT) Raman spectrometer with neodymium-doped yttrium aluminum garnet (ND:YAG) laser [32], equipped with a holographic filter, and either a germanium (Ge) or a indium gallium arsenide (InGaAs) detector became available [30].

1.2.1.2 Theory

When a photon of light from the incident monochromatic light source interacts with a sample, the sample may emit scattered radiation (Figure 3). When the sample interacts with source light, the sample will absorb that energy until it reaches a virtual state. This is not a quantized energy level; instead it is an electronic state between that of the ground state energy level and that of an excited state energy level [28]. In the virtual state, there will be a transient distortion of the electron cloud, causing the molecule to be temporarily polarized (i.e., an induced dipole moment). Relaxation of the molecule from the virtual state to the ground state causes emission of scattered light [27,28,33].

Rayleigh (elastic) scattering occurs if the energy of the scattered light is equal to that of the incident light. However, the energy of the scattered light may be less than that of the incident light, which is known as Stokes scattering. If the energy of the scattered light is greater than the incident light, anti-Stokes scattering is said to have occurred. Both Stokes and anti-Stokes scattering are forms of inelastic scattering known as Raman scattering [27,28]. Note that anti-Stokes scattering is less likely than Stokes scattering because in the former case, the molecule must be in an excited vibrational level of the ground state when the incident photon is initially absorbed. Figure 4 is an example of a recreated Raman spectrum for chloroform showing the peaks resulting from Rayleigh and Raman scattering processes [27]. The Stokes and anti-Stokes shifts are symmetrical about the Rayleigh shift because the change in energy equal to one vibrational level is measured. This is also what makes Raman spectroscopy a complimentary technique to Infrared Spectroscopy in that Infrared also measures vibrational energy levels, albeit by an absorption process. Therefore the frequency shifts that are observed are identical for both techniques [27].

In Figure 4, the Stokes peaks appear at a lower frequency because they are lower in energy compared to the incident light, whereas the anti-Stokes peaks are found at a higher frequency because they are greater in energy compared to the incident light [27,33]. In Raman spectra, only the Stokes shifts are shown because the Rayleigh peak is so great in intensity compared to Raman scattering (approximately 10^6-fold) [27]. The Stokes shifts are also chosen because they are of greater intensity than the anti-Stokes shifts because most molecules are in the lowest vibrational level of the ground state

electronic energy level at room temperature. Note that the Stokes shifts (Δcm^{-1}) are reported as positive values by convention [27].

For Raman scattering to occur, an induced dipole must occur in the molecule. To derive a mathematical expression for Raman scattering, one must first develop an equation that describes an induced dipole [27-29,34,35].

Consider incident light and its associated electric field, as represented in Equation 1-1.

$$E = E_0 \cos(2\pi \nu_{ex} t) \qquad 1\text{-}1$$

In Equation 1-1, E is the electric field associated with the incident light, E_0 is the amplitude of the electric field, ν_{ex} is the frequency of the incident light, and t is time. When the incident light comes into contact with the molecule, an induced dipole moment (m) can occur as seen in Equation 1-2 where α is the polarizability of the bond.

$$m = \alpha E \qquad 1\text{-}2$$

Combination of Equations 1-1 and 1-2 produces Equation 1-3.

$$m = \alpha E_0 \cos(2\pi \nu_{ex} t) \qquad 1\text{-}3$$

In Equation 1-3, the induced dipole is a function of both the electric field of the incident radiation as well as the polarizability of the molecule. The polarizability of the electron cloud is primarily a function of the distance between the nuclei of the atoms in the bond, which is shown in Equation 1-4

$$\alpha = \alpha_0 + (r - r_{eq})(\delta\alpha/\delta x) \qquad 1\text{-}4$$

In Equation 1-4, α_0 is the polarizability of the bond at its equilibrium distance (r_{eq}) for a nuclear separation (r) at a specific point in time. This distance can change depending on the extent of the molecular vibrations (v_v) that are present, as shown in Equation 1-5:

$$r - r_{eq} = r_m \cos(2\pi\, v_v t) \qquad 1\text{-}5$$

where r_m is the maximum separation of the nuclei compared to their equilibrium positions. Substituting Equation 1-5 into 1-4, then substituting 1-4 into 1-3 produces Equation 1-6.

$$m = \alpha_0 E_0 \cos(2\pi v_{ex} t)$$

$$+ (E_0/2) r_m (\delta\alpha/\delta x) \cos[2\pi(v_{ex} - v_v)t]$$

$$+ (E_0/2) r_m (\delta\alpha/\delta x) \cos[2\pi(v_{ex} + v_v)t] \qquad 1\text{-}6$$

The first term in Equation 1-6 represents Rayleigh scattering, the second term is for Stokes scattering because there is a loss in energy which corresponds to a loss in frequency from the incident frequency ($v_{ex} - v_v$), and the third term is for anti-Stokes scattering because there is a gain in energy which corresponds to a gain in frequency from the incident frequency ($v_{ex} + v_v$) [27,29].

Analytical chemists are interested in obtaining both qualitative (i.e., frequency shift) as well as quantitative information. The intensity of the Raman signal is based upon the probability of inelastic scattering, or the "Raman cross section" [28]. Because knowledge of the integrated cross section over all the three-dimensional modes in which a mixture of molecules would in-elastically scatter light, it is more common to simplify the problem to that of a solid angle (in steradians, sr) in which scattering occurs [28].

This is known as the differential cross section, β (cm^2 $molecule^{-1}$ sr^{-1}). The relationship between the intensity and the differential cross section is shown in Equation 1-7, were L is the specific intensity of scattered radiation (photons cm^{-1} sr^{-1} sec^{-1}), P_D is the power density (photons cm^{-2} sec^{-1}), D is number density of scattered light (molecules cm^{-3}) and dz (cm^2) is the path length of the laser beam in the sample [28].

$$L = P_D \beta D(dz) \qquad 1\text{-}7$$

There are many factors that affect the differential cross section, such as the types of atoms, types of bonds, whether extended or conjugated π systems are present, as well as the physical state of the sample [28]. Understanding the differential cross section will help to provide more information about the nature of the observed Raman intensity to predict the type of samples that will have greater intensity compared to others. When comparing two spectra, information about the differential cross sections of the molecules under study may be useful to explain unexpected and otherwise perplexing features in the spectra.

1.2.1.3 Instrumentation

A Raman spectrometer is shown schematically in Figure 5 [27]. The laser sends a beam of incident light towards the sample. Raman scattering from the sample is then collected by the wavelength selector, which is set to a specific spectral range. The light then travels to a photon transducer to convert the light energy into an electrical signal for computer processing [27]. A schematic of the instrument used in this study, a Thermo Nicolet Nexus 670 FT-IR with Thermo Nicolet Nexus 900 Raman attachment (formally

Thermo Nicolet, now Thermo Fisher Scientific, Madison, WI), is shown in Figure 6. The laser source is an Nd:YAG laser (1064 cm^{-1}). This laser is common in FT-Raman instruments because its low energy (in the near infrared range, NIR) reduces not only photodecomposition but also fluorescence – major problems with higher energy (shorter wavelength) sources [27,28,30,31]. While there is some compromise to the signal strength by using a NIR laser, the benefit particularly in reducing interfering fluorescence is more profound [28]. A holographic filter is used to efficiently block the Rayleigh line because it has a much larger intensity (usually $\geq 10^6$) that would obscure the Raman scattering peaks [30]. The Michelson interferometer is used to convert high frequency radiation to low frequency radiation by the use of a beam splitter and a moving mirror. The transformation of frequency- to time-domain information means that the signal can be resolved by the detector. In this process, the scattered radiation is equally divided by the beam splitter; one beam is directed to a moving mirror while the other is directed to a stationary mirror. When these polychromatic beams are recombined, they interfere both constructively and destructively (i.e., depending on the position of the moving mirror at a given time). By knowing the speed of the moving mirror and the distance that the moving mirror has traveled, a Fourier transformation algorithm is employed to convert the time domain information into the frequency domain (i.e., to produce a spectrum) [36]. Finally, the instrument schematic shows the most recent improvement to the FT-Raman spectrometer — detection. Advances in detector technology have resulted in the ability to measure spectra over wide wavelength ranges (e.g., 0.8 – 1.6 μm) [30]. There are two common detectors used: InGaAs or Ge [33]. These are both photoconductive devices that cover a large spectral range including the longer wavelength regions [30].

1.2.2 Gas Chromatography

1.2.2.1 Solid-Phase Microextraction (SPME)

Solid-phase microextraction (SPME) was first introduced by Pawliszyn and Arthur in 1990 [37]. SPME is based upon the use of a thin polymer coating on a fused silica rod. The basic idea is that the analyte of interest (or class of analytes) will partition and/or absorb into the polymer coating on the fiber. The fiber is then directly inserted into the GC injection port where the analyte will thermally desorb from the polymer coating into the chromatographic column [37]. This process is illustrated schematically in Figure 7. In Figure 7-a, the coated fiber is unexposed inside a hollow stainless steel needle of the SPME holder, which is used to puncture the septum of the sample vial. The plunger on the SPME holder is then depressed and the fiber is exposed to the analytes in the solution for a specific period of time (Figure 7-b). Once the analyte has partitioned into the coating on the fiber, the fiber is retracted into the hollow needle (Figure 7-c). Next in the procedure, the stainless steel needle punctures the septum on the injection port (Figure 7-d). The plunger is depressed again, the fiber is exposed to the hot injector, and the analytes will desorb (Figure 7-e). After the analytes have thermally desorbed, the fiber can be retracted and the stainless steel needle removed from the injector (Figure 7-f) [38]. In this way, only a relatively specific set of analytes are determined, based upon the selectivity of the polymeric coating on the fiber [39]. Thus SPME combines five steps into one procedure: sampling, extraction, pre-concentration, matrix removal, and GC introduction. In addition to being a "solvent-free" method, other advantages that SPME has over traditional extraction techniques is that it is fast, simple, and relatively

inexpensive [38]. The primary disadvantage of SPME is that it is not an exhaustive sampling technique but rather is based on the equilibrium established, i.e., it is matrix-dependent. Thus a SPME method must be calibrated to a given sample matrix, e.g., environmental waters from different origins must have separate quantitation models.

While SPME is a useful technique for many analytical problems, depending on the sample, it is not always the best choice. Some samples could either destroy the fiber (mineral acids or bases) or have such small partitioning coefficients that very long sampling times are required [40]. One way to resolve this issue is to use headspace SPME. Figure 8 shows the headspace SPME procedure. Although headspace SPME appears similar to the liquid-phase process depicted in Figure 7, a key difference is that the fiber is not immersed in the liquid phase but rather is exposed to the volatile components in the headspace above the liquid surface. The primary advantage in using headspace SPME-GC is the shorter extraction time that is typically needed because diffusion times for the gas-phase analyte into the fiber phase are much shorter than those typically found in the liquid phase. Additionally, the degree of selectivity is higher for headspace SPME because non-volatile analytes in complex matrices are not present in the headspace [40]. Compared to the direct syringe injection of a sample of the headspace of a solution, headspace SPME-GC is advantageous because atmospheric interferences (e.g., oxygen, moisture, pollutants) are typically excluded from the SPME fiber [40]. There are many applications for which headspace SPME-GC methods have been reported, including food chemistry, microprocessor fabrication, and artistic media [41-44].

1.2.2.2 Pulsed Flame Photometric Detector (PFPD)

The pulsed flame photometric detector (PFPD) was chosen for this project because it was important to have a detector with high sensitivity and good selectivity. The PFPD was first introduced by Amirav in 1991 [45]. This variation on the design of the conventional flame photometric detector not only improved sensitivity and selectivity, but also reduced the amount of combustion gases that were needed. The key difference in the design of the PFPD compared to the conventional FPD is that the ignitor is not located at the end of the chromatographic column but rather some distance away from it (Figure 9). Pulsing occurs because the flame is not created until the combustion gases fill the chambers and reach the ignitor. This process is shown in more detail in Figure 10 [46]. During the initial "fill" stage, a mixture of combustible gases (usually air and hydrogen) fills the combustion and ignition chambers. When the gas mixture comes into contact with the heated ignitor coil, the flame is created. The flame then propagates through the ignition chamber and into the combustion chamber, where the column effluent is found. Once the combustion gases are consumed, the flame is extinguished. The gases will typically combust twice every second providing the pulsed flame effect. However, the column effluent is also fed into the combustion chamber so that as the flame comes into contact with the analyte, not only will the analyte break down into smaller fragments, but the flame may also create molecular excited states. As these states relax, some will emit light that is detected by a photomultiplier tube (PMT) detector [45-47]. Depending on the analyte being determined, different time-dependent emission profiles will be observed. This time dependence distinguishes the PFPD from the

conventional FPD, where only spectral selectivity is possible from the different energies observed during the emission process [46].

1.2.3 Ageing Chamber

In the Mellon report [1], a lack of SRMs for artistic media was identified as a key short-coming in art conservation research. SRMs are needed to validate methods. While it would be ideal to have a naturally aged varnish for use as a SRM, that is not feasible. First, there would need to be a large quantity of the aged varnish to support an SRM program serving many laboratories. This is impractical because of the relatively small quantities available. An additional complication is that very few records have been kept as to what type of varnish was applied to a given object, as well as any previous restoration that took place [48]. In several reported studies of naturally aged varnishes from paintings, the records to indicate if the piece had been restored were not available [49-51]. Because it is not always possible to obtain naturally aged samples, there are many studies where the researchers artificially age samples as a means to resemble naturally aged samples [17,51-57]. Thus there is a need not only for SRMs in general, but also ways to create them [1].

1.2.3.1 Physical Design and Experimental Approach

The degradation of any artistic material is a combination of four primary factors: heat, light, humidity, and pollutants [1]. In this work, the goal was to create an ageing

chamber where one can control the levels of these factors simultaneously as a means to study the degradation of varnishes [1]. A simple and efficient way to do this was to build a chamber that can be directly incorporated into a GC oven. In this way the effect of heat can be precisely controlled by the oven and the various ports on the GC could be used to introduce humid air and pollutants. A UV lamp can also be readily placed inside the oven so that the effects of radiation on the degradation process can be studied. Also by incorporating the chamber into the GC oven it also makes studies in the absence of light quite simple as well. By using the ageing chamber, the exact levels of these factors can be set, therefore making the ageing experiments relatively simple to conduct. Once a procedure for artificially ageing a varnish is developed, the resultant SRM would be straightforward to produce in repeatable form in large quantities.

Along with using the ageing chamber to create SRMs, it can also be used to account for another important issue identified in the Mellon Report [1]: the fact that few, if any, researchers have studied all the factors that affect ageing simultaneously. Schönemann and Edwards [52] studied the differences in the Raman and FTIR spectra obtained from aged oils compared to fresh oils. In this study, they artificially aged the oils in a UV cabinet at ambient temperature and humidity [52]. Other researchers also studied the degradation of varnishes by exposure of artistic media to UV radiation in the form of a xenon arc fadeometer [53,54] or a carbon arc fadeometer [56]. While few studied more than one factor associated with the ageing of varnishes, some did look at multiple effects. de la Rie and colleagues studied the degradation pathways of Dammar resin using a UV light source and heat from a conventional oven [54]. Recently, a study was conducted using an ageing chamber to study how pollutants and humidity affect

ageing [57]. Bonaduce *et al.*, developed ageing chambers consisting of three separate water-jacketed cylinders at 22 °C and low UV light levels. They were able to either pump humid air or different pollutants such as NO_2, O_3, and acetic acid into the chamber to study how these factors affected ageing [57]. Because few studies been reported for the simultaneous variation of multiple factors in varnish ageing, the ageing chamber developed herein will help to fill the current knowledge gap in this area of research.

Figure 1: Examples of triglyceride and a triterpenoid compounds [7].

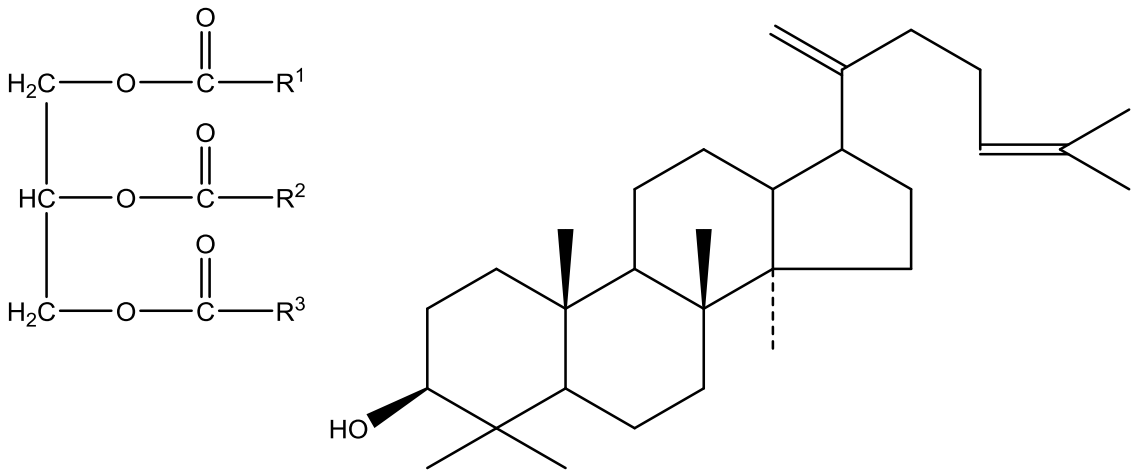


Figure 2: The reaction scheme for the oxidation of drying oils [7].

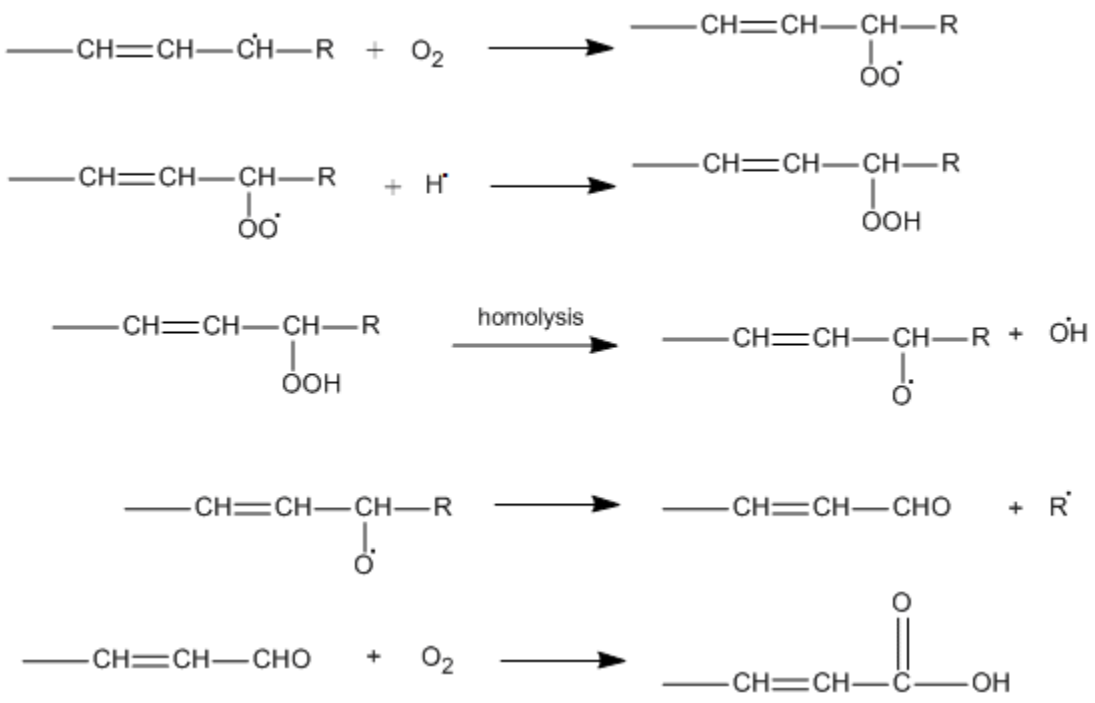


Figure 3: Energy level diagram of Raman scattering [27].

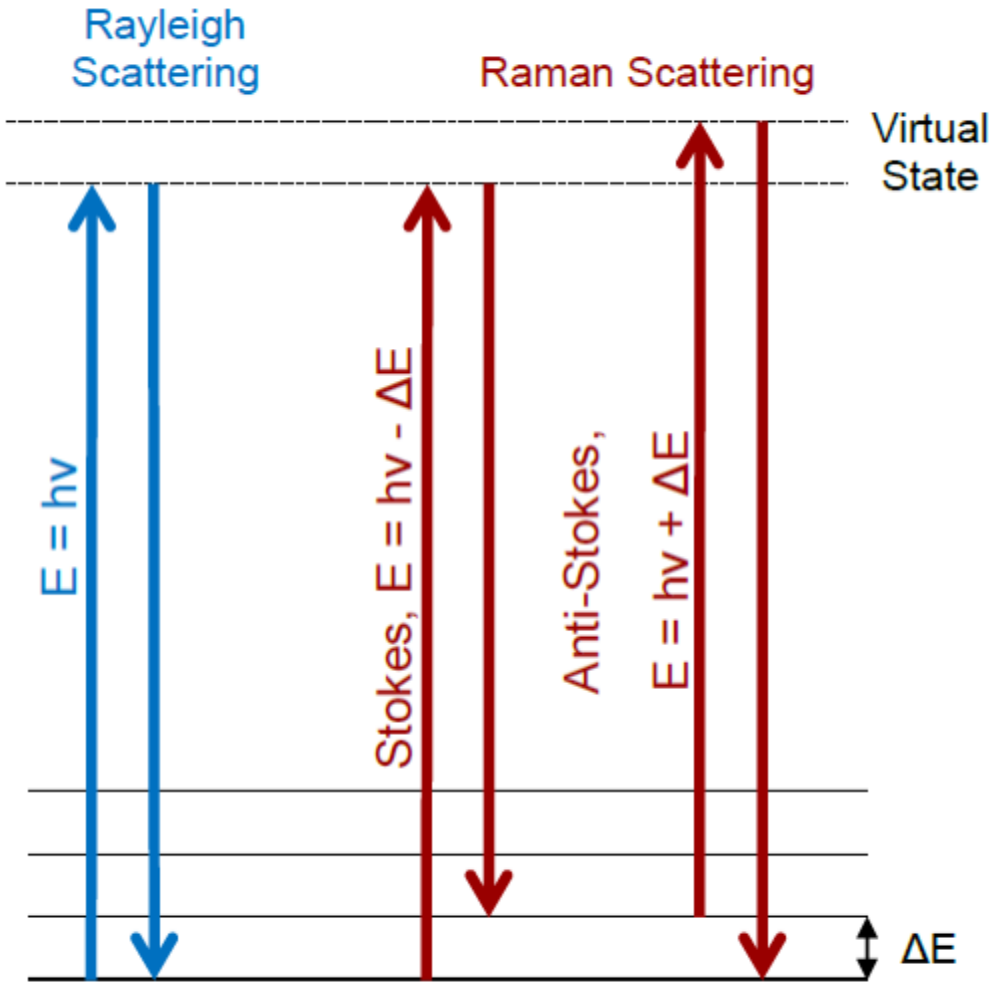


Figure 4: Recreated Raman spectrum of chloroform showing the Rayleigh scattering band, the Stokes scattering shift, and the anti-Stokes scattering shift [27].

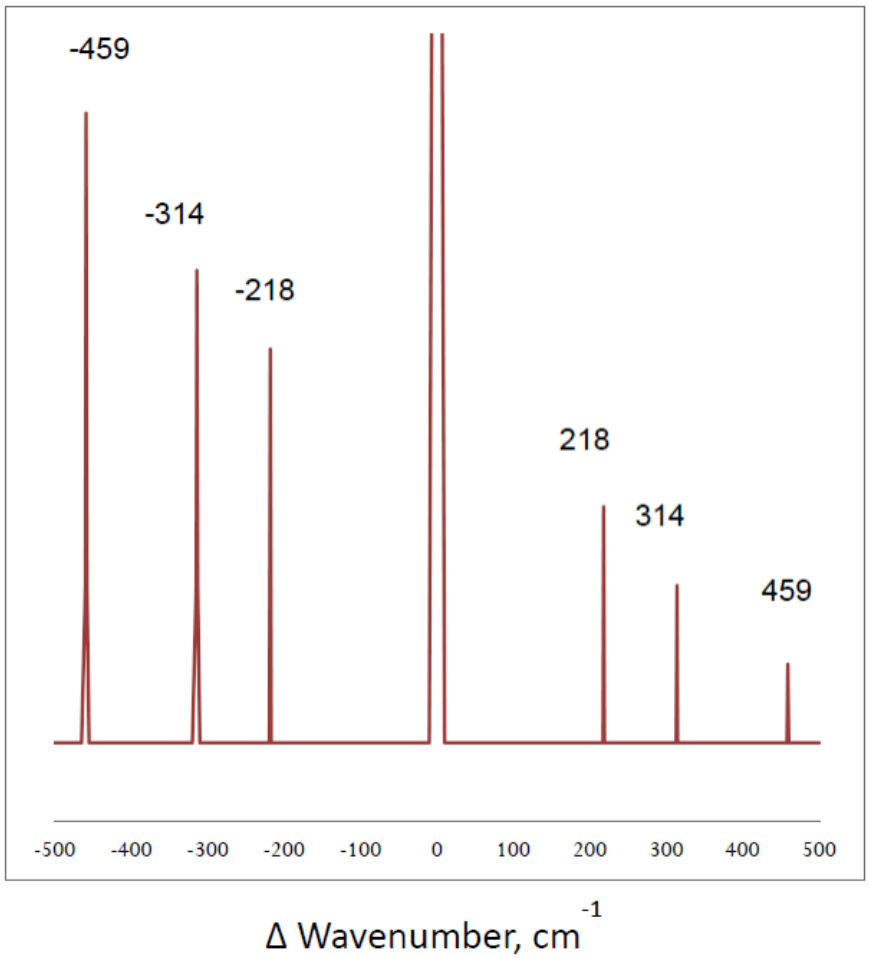


Figure 5: Generic schematic diagram of a Raman Spectrometer [27].

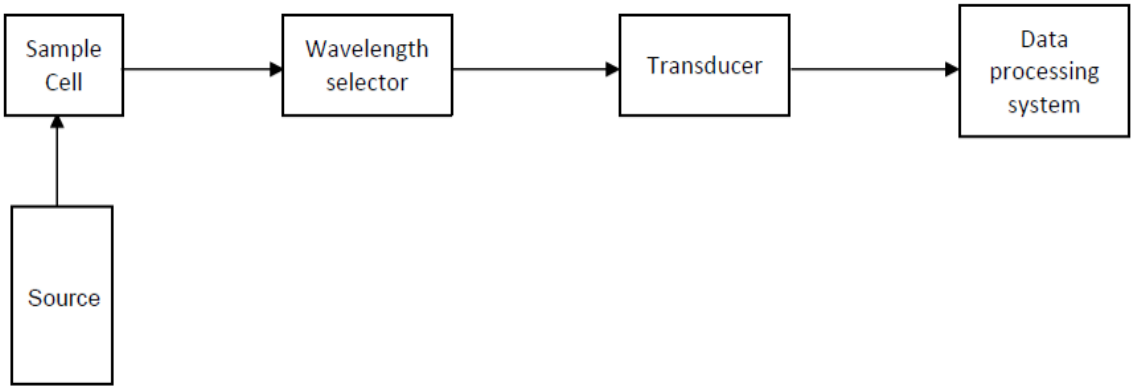


Figure 6: Schematic diagram of the Thermo Nicolet Nexus 670 FT-IR with Nexus 900 Raman attachment including a Michelson Interferometer [60] (M = mirror, MM = moving mirror, and FM = fixed mirror).

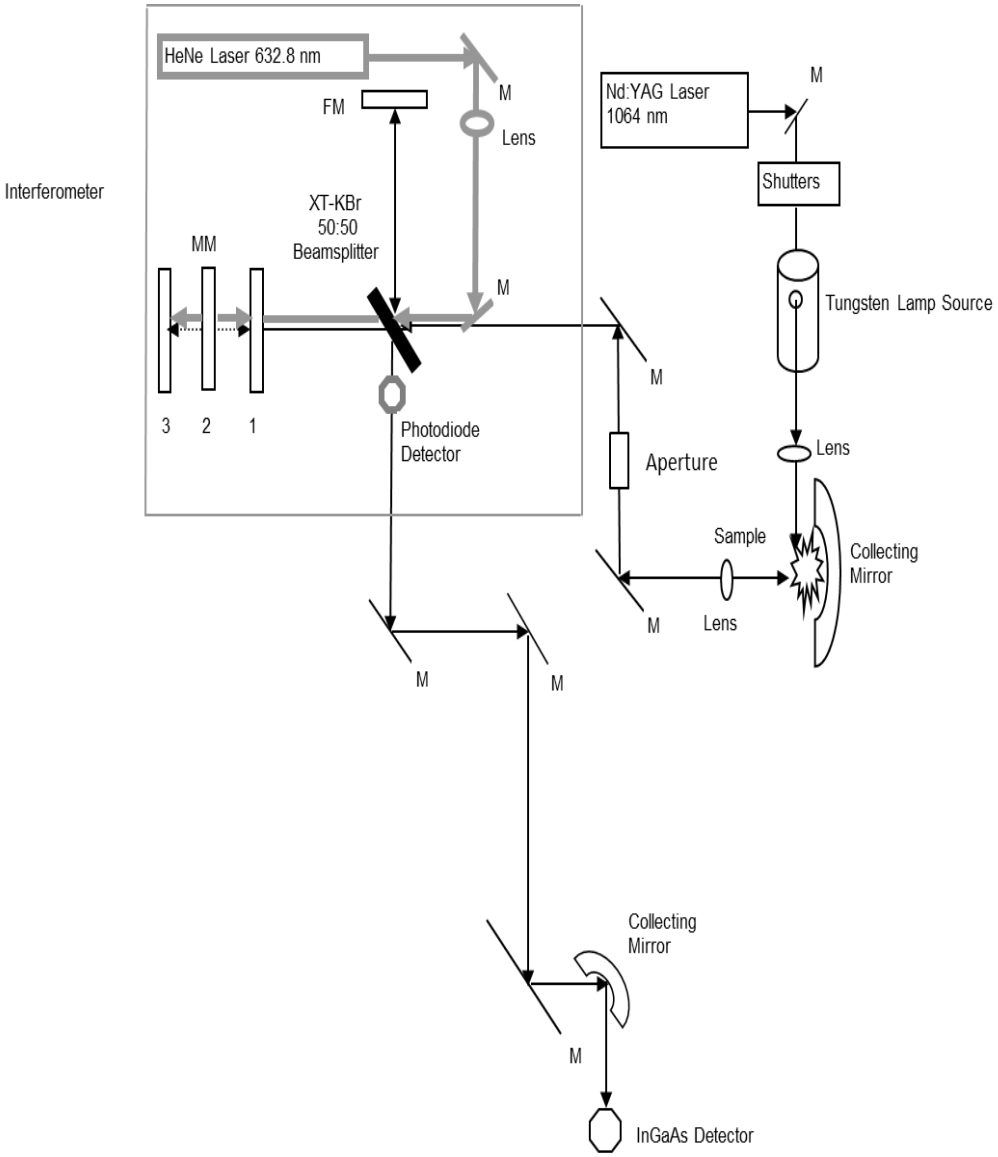


Figure 7: Sample collection and GC introduction process for solid-phase microextration (SPME) [38].

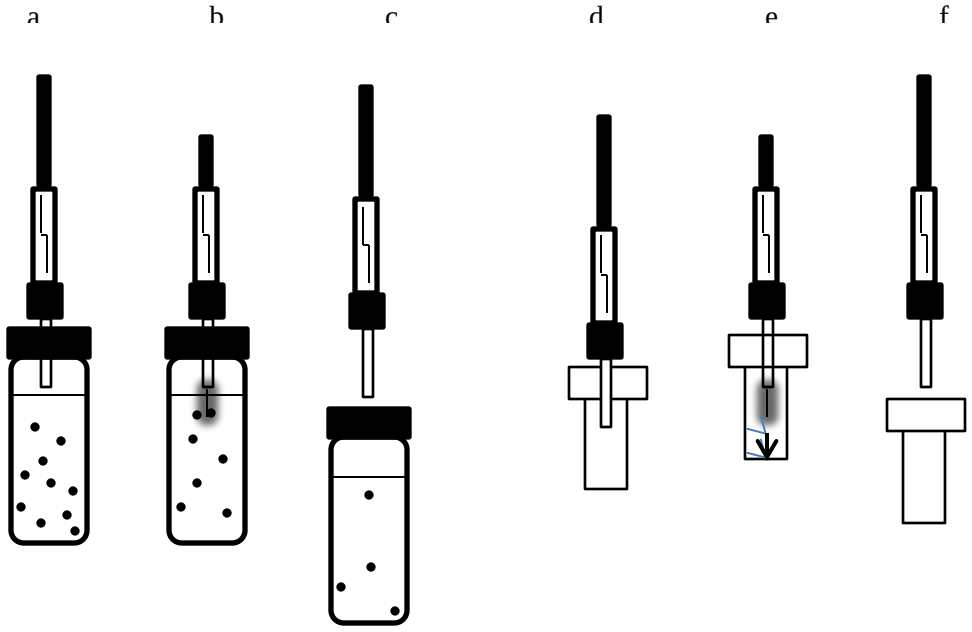


Figure 8: Sample collection and GC introduction process for headspace solid-phase microextraction (SPME) [38].

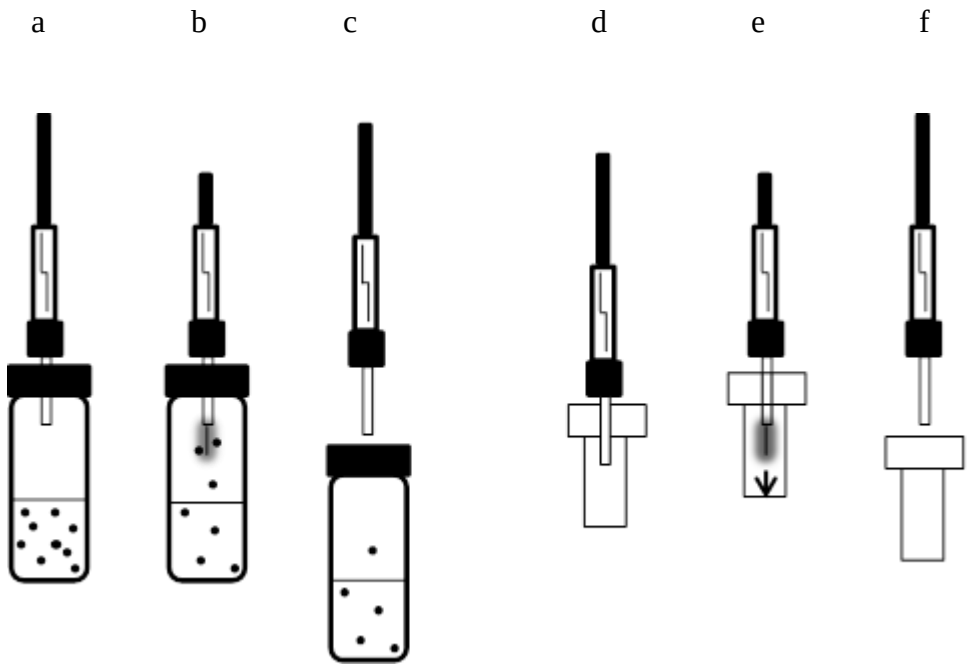


Figure 9: Schematic diagram of the pulsed flame photometric detector (PFPD) [46] (PMT = photomultiplier tube).

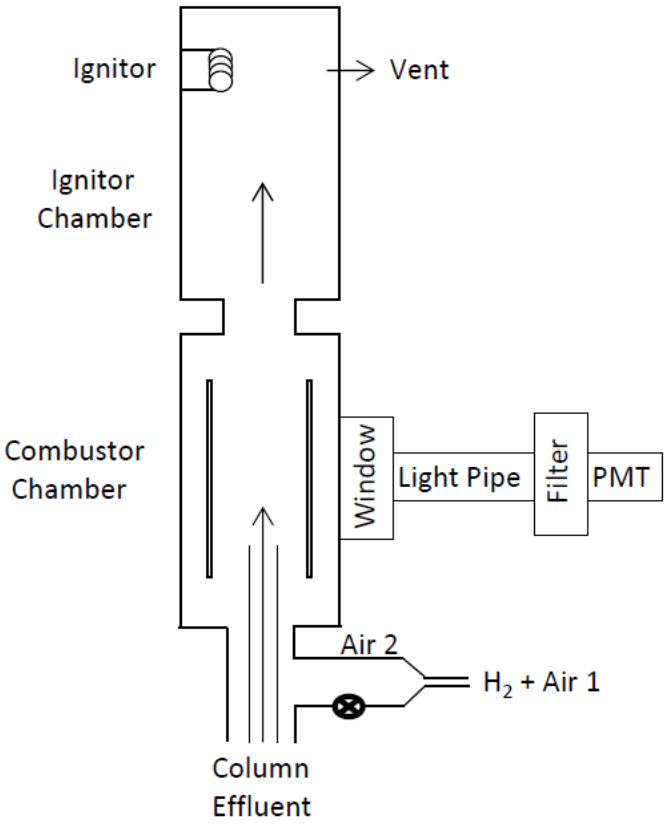


Figure 10: The four main stages of the detection of analytes using a pulsed flame photometric detector [46] (PMT = photomultiplier tube).

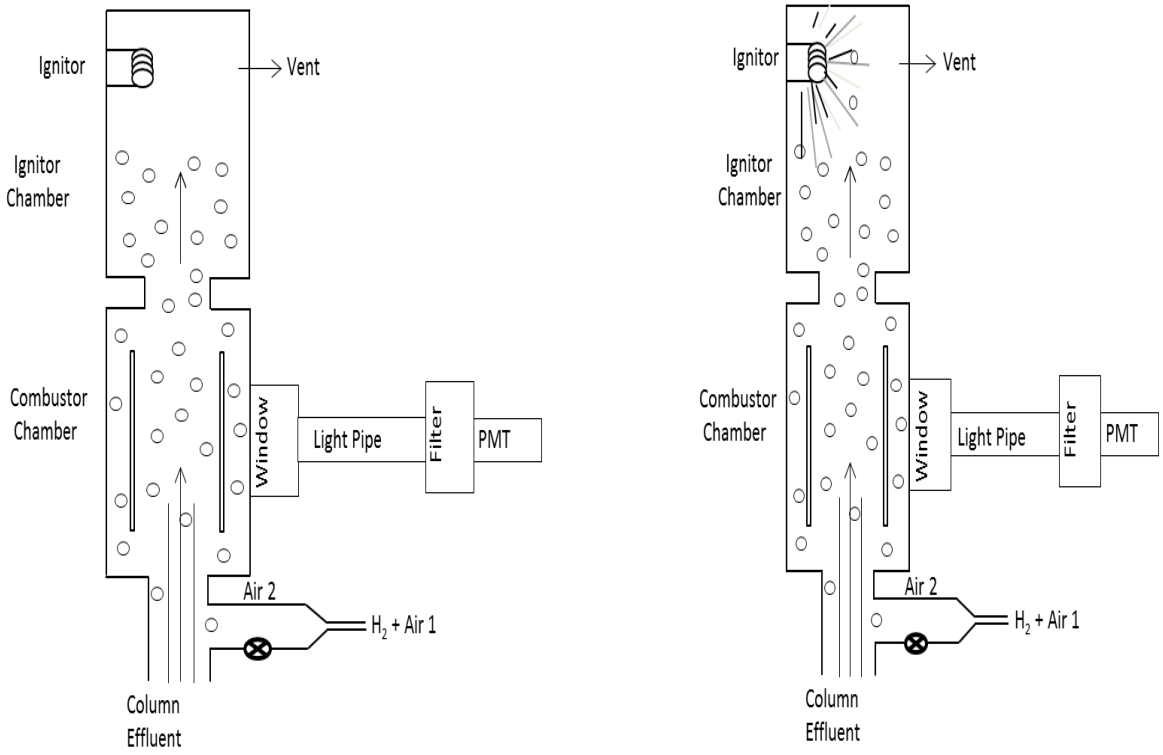
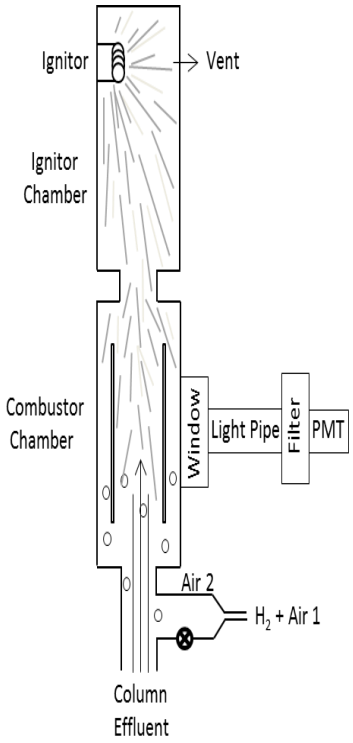
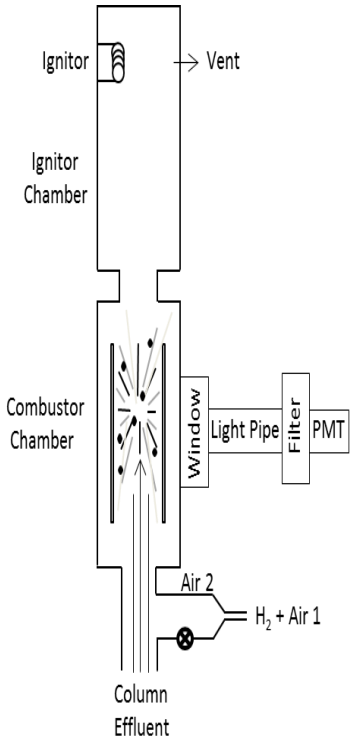


Figure 10: Cont.





References

(1) Leona, M.; Van Duyne, R.; Berrie, B.; Casadio, F.; Ernst, R. R.; Faber, K. T.; Sgamellotti, A.; Trentelman, K.; Whitmore, P., Chemistry and materials research at the interface between science and art; report of a workshop cosponsored by the National Science Foundation and the Andrew W. Mellon Foundation, 2009.

(2) Buchanan, M. *From the collection–George Mann Niedecken*, 2010.

(3) de la Rie, E. R., Old master paintings, *Analytical Chemistry,* 1989, *61*, 1228A-1240A.

(4) Penn, T. Z., Decorative and protective finishes, 1750-1850: Materials, Process, and Craft, *Bulletin of the Association for Preservation Technology,* 1984, *16*, 3-46.

(5) Mussey, R. D., Early varnishes, *Fine Woodworking,* 1982, 54-57.

(6) Lahanier, C.; Preusser, F. D.; Van Zelst, L., Study of conservation of museum objects: Use of classical analytical techniques, *Nuclear Instruments and Methods in Physics Research*, 1986, *B14*, 1-9.

(7) Mills, J. S.; White, R. *The Organic Chemistry of Museum Objects*, 2nd ed.; Butterworth Heinemann: London, 1994, p 206.

(8) van der Doelen, G. A. *Molecular studies of fresh and aged triterpenoid varnishes*. University of Amsterdam, 1999.

(9) de la Rie, E. R., The influence of varnishes on the appearance of paintings, *Studies in Conservation*, 1987, *32*, 1-13.

(10) Zumbul, S.; Knochenmuss, R.; Wulfert, S.; Dubois, F.; Dale, M. J.; Zenobi, R., A Graphite-assisted laser desorption/ionization study of light-induced aging in triterpene dammar and mastic varnishes, *Analytical Chemistry,* 1998, *70*, 707-715.

(11) Movasaghi, Z.; Rehman, S.; Rehman, I. U., Raman Spectroscopy of biological tissues, *Applied Spectroscopy Reviews*, 2007, *42*, 493-541.

(12) De Gelder, J.; De Gussem, K.; Vandenabeele, P.; Moens, L., Reference database of Raman spectra of biological molecules, *Journal of Raman Spectroscopy*, 2007, *38*, 1133-1147.

(13) de Veij, M.; Vandenabeele, P.; De Beer, T.; Remon, J. P.; Moens, L., Reference database of Raman spectra of pharmaceutical excipients, *Journal of Raman Spectroscopy*, 2009, *40*, 297-307.

(14) Lowry, S.; Wieboldt, D.; Dalrymple, D.; Jasinevicius, R.; Downs, R. T., The use of a Raman spectral database of minerals for the rapid verification of semiprecious gemstones, *Spectroscopy*, 2009, *24*.

(15) Ricciardi, P.; Colomban, P.; Fabbri, B.; Milande, V., Towards the establishment of a Raman database of early European porcelain, *e-Preservation Science*, 2009, *6*, 22-26.

(16) Castro, K.; Perez-Alonso, M.; Rodriquez-Laso, M. D.; Fernandez, L. A.; Madariaga, J. M., On-line FT-Raman and dispersive Raman spectra database of artists' materials (e-VISART database), *Analytical Bioanalytical Chemistry*, 2005, *382*, 248-258.

(17) Nevin, A.; Comelli, D.; Osticioli, I.; Toniolo, L.; Valentini, G.; Cubeddu, R., Assessment of the ageing of triterpenoid paint varnishes using fluorescense, Raman, and FTIR spectroscopy, *Analytical Bioanalytical Chemistry*, 2009, *395*, 2139-2149.

(18) Jian, M.; Lam, K.-M.; Dong, J., A novel face-hallucination scheme based on singular value decomposition, *Pattern Recognition*, 2013, *46*, 3091-3102.

(19) Nibouche, O.; Jiang, J., Palmprint matching using feature points and SVD factorisation, *Digital Signal Processing*, 2013, *23*, 1154-1162.

(20) Brissaud, I.; Guilló, A.; Lagarde, G.; Midy, P.; Calligaro, T.; Salomon, J., Determination of the sequence and thicknesses of multilayers in an easel painting, *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 1999, *155*, 447-451.

(21) Berns, R. S.; Taplin, L. A.; Nezamabadi, M.; Zhao, Y.; Okumura, Y., High-Accuracy digital imaging of cultural heritage without visual editing; RIT Munsell Color Science Laboratory: Rochester, New York.

(22) Colombini, M. P.; Andreotti, A.; Bondaduce, I.; Modugno, F.; Ribechini, E., Analytical strategies for characterizing organic paint media using Gas Chromatography/Mass Spectrometry, *Accounts of Chemical Research*, 2010, *43*, 715-727.

(23) van Keulen, H., Gas chromatography/mass spectrometry methods applied for the analysis of a Round Robin sample containing materials present in samples of works of art, *International Journal of Mass Spectrometry*, 2009, *284*, 162-169.

(24) Abdel-Ghani, M.; Edwards, H. G. M.; Janaway, R.; Stern, B., A Raman microscope and gas chromatographic-mass sectrometric study of two 19th century overlapping Coptic icons of Anastasy Al-Romit, *Vibrational Spectroscopy*, 2008, *48*, 69-75.

(25) Blasko, J.; Kubinec, R.; Husova, B.; Prlkrul, P.; Pacakova, V.; Stulik, K.; Hradllova, J., Gas chromatography/mass spectrometry of oils and oil binders in paintings, *Journal of Separation Science*, 2008, *31*, 1067-1073.

(26) Echard, J. P.; Benoit, C.; Peris-Vicente, J.; Malecki, V.; Gimeno-Adelantado, J. V.; Vaiedelich, S., Gas chromatography/mass spectrometry characterization of historical varnishes of ancient Italian lutes and violin, *Analytica Chimica Acta*, 2007, *584*, 172-180.

(27) Skoog, D. A.; Holler, F. J.; Crouch, S. R., Raman Spectroscopy In *Principles of Instrumental Analysis*, Kiselica, S., Ed.; Thomson Brooks/Cole: Canada, 2007, pp 481-497.

(28) McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons, Inc.: Canada, 2000; Vol. 157.

(29) Ferraro, J. R.; Nakamoto, K. *Introductory Raman Spectroscopy*; Academic Press, Inc.: San Diego, CA, 1994.

(30) Chase, B., FT-Raman Spectroscopy: A catalyst for the Raman explosion?, *Journal of Chemical Education,* 2007, *84*, 75-80.

(31) Chase, D. B.; Rabolt, J. F. *Fourier Transform Raman Spectroscopy*; Academic Press, Inc.: San Diego, CA.

(32) Raschotta, R. *YAG Lasers*; RP Photonics Encyclopedia.

(33) Strommen, D. *Raman Spectroscopy*, Settle, F. A., Ed.; Prentice Hall, 1997, pp 285-307.

(34) Szymanski, H. A. *Raman Spectroscopy: Theory and Practice*; Plenum Press: New York, NY, 1970.

(35) Koningstein, J. A. *Introduction to the Theory of the Raman Effect*; D. Reidel: Dordrecht, Holland, 1972.

(36) Skoog, D. A.; Holler, F. J.; Crouch, S. R.,Components of Optical Instruments In *Principles of Instrumental Analysis*, Kiselica, S., Ed.; Thomson Brooks/Cole: Canada, 2007, pp 164-211.

(37) Arthur, C. L.; Pawliszyn, J., Solid phase microextraction with thermal desorption using fused silica optical fibers, *Analytical Chemistry,* 1990, *62*, 2145-2148.

(38) *Solid Phase Microextraction: Theory and Optimization of Conditions*; Supelco, 1998, pp 1-8.

(39) *Solid Phase Microextration Troubleshooting Guide*; Supelco, 2004, p 12.

(40) Zhang, Z.; Pawliszyn, J., Headspace solid-phase microextraction, *Analytical Chemistry*, 1993, *65*, 1843-1852.

(41) Krist, S.; Stuebiger, G.; Unterweger, H.; Bandion, F.; Buchbauer, G., Analysis of volatile compounds and triglycerides of seed oils extracted from different poppy varieties, *Journal of Agricultural and Food Chemistry*, 2005, *53*, 8310-8316.

(42) Hinz, D. C.; Kwarteng-Achaempong, W.; Wenclawiak, B. W., Analysis of volatile varnishes of coated wires by SPME, *Fresenius' Journal of Analytical Chemistry*, **1999**, *364*, 641-642.

(43) Lattuati-Derieux, A.; Thao, S.; Langlois, J.; Regert, M., First results on headspace-solid phase microextration-gas chromatography/mass sectrometry of volatile organic compounds emitted by wax objects in museums, *Journal of Chromatography A*, 2008, *1187*, 239-249.

(44) Godoi, A. F. L.; Van Vaeck, L.; Van Grieken, R., Use of Solid-phase microextraction for the detection of acetic acid by ion-trap gas chromatography-mass spectrometry and application to indoor levels in museums, *Journal of Chromatography A*, 2005, *1067*, 331-336.

(45) Atar, E.; Cheskis, S.; Amirav, A., Pulsed Flame -- A novel concept for molecular detection, *Analytical Chemistry*, 1991, *63*, 2064-2068.

(46) *Pulsed Flame Photometric Detector (PFPD) for 3800*; Varian Associates, Inc., 1998, pp 1-4 - 1-10.

(47) Cheskis, S.; Atar, E.; Amirav, A., Pulsed-flame photometer: A novel gas chromatography detector, *Analytical Chemistry*, 1993, *65*, 539-555.

(48) Caruso, F.; Saverwyns, S.; Van Bos, A.; Chillura Martino, D. F.; Ceulemans, A.-E.; de Valck, J.; Caponetti, E., Micro-X-Ray Fluorescence and the old masters: Non-destructive in situ characterisation of the varnish of historical low countries stringed musical instruments, *Applied Physics A*, 2012, *107*.

(49) van der Doelen, G. A.; Jan van den Berg, K.; Boon, J. J., Comparative Chromatographic and Mass-Spectrometric studies of triperpenoid varnishes: Fresh material and aged samples from paintings, *Studies in Conservation*, 1998, *43*, 249-264.

(50) van der Doelen, G. A.; Jan van den Berg, K.; Boon, J. J.; Shibayama, N.; de la Rie, E. R.; Geniut, W. J. L., Analysis of fresh triterpenoid resins and aged triterpenoid varnishes by high-performance liquid chromatography-atmospheric pressure chemical ionisation (tandem) mass spectrometry, *Journal of Chromatography A*, 1998, *809*, 21-37.

(51) van der Doelen, G. A.; Boon, J. J., Artificial ageing of varnish triterpenoids in solution, *Journal of Photochemistry and Photobiology A: Chemistry*, 2000, *134*, 45-57.

(52) Schonemann, A.; Edwards, H., G. M., Raman and FTIR microspectroscopic study of the alteration of Chinese tung oil and related drying oils during ageing, *Analytical Bioanalytical Chemistry*, 2011, *400*, 1173-1180.

(53) de la Rie, E. R.; McGlinchey, C. W., Stabilized Dammar picture varnish, *Studies in Conservation*, 1989, *34*, 137-146.

(54) de la Rie, E. R., Photochemical and thermal degradation of films of Dammar resin, *Studies in Conservation*, 1988, *33*, 53-70.

(55) Feller, R. L.; Bailie, C. W., Solubility of aged coatings based on Dammar, Mastic, and Resin AW-2, *Bulletin of the American Group. International Institute for Conservation of Historic and Artistic Works*, 1972, *12*, 72-81.

(56) Feller, R. L., A note on the exposure of Dammar and Mastic varnishes to fluorescent lamps, *Bulletin of the American Group. International Institute for Conservation of Historic and Artistic Works*, 1964, *4*, 12-14.

(57) Bonaduce, I.; Odlyha, M.; Di Girolamo, F.; Lopez-Aparicio, S.; Grontoft, T.; Perla Colombini, M., The role of organic and inorganic indoor pollutants in museum environments in the degredation of dammar varnish, *Analyst*, 2013, *138*, 487-500.

Chapter 2: Experimental

2.1 Raman Spectroscopy

2.1.1 Reagents

All reagents used were analytical reagent grade (AR) or better. The organic acids used are listed in Table I. The other chemicals used were Linseed Oil which was purchased from Aldrich (Milwaukee, WI, USA). The Tung Oil was Formby's Tung Oil which was purchased at a local hardware store. Dammar and Mastic were purchased through Kremer Pigmente (New York, NY, USA). Most samples were run neat, i.e., in the sample's original state. Dilute samples were also prepared using 18 MΩ-cm deionized water obtained from a Thermo Scientific Nanopure Ultrapure water system (Waltham, MA, USA). Glassware was washed and then soaked for at least 48 hr in 5% (v/v) HNO_3 (AR grade, Fisher Scientific, Pittsburgh, PA, USA) followed by rinsing with 18 MΩ-cm deionized water.

2.1.2 Raman Instrument

The Raman instrument that was used to characterize complex organic molecules was a Thermo Nicolet Model Nexus 670 Fourier Transform Infrared Spectrometer (FT-IR) with Model Nexus 900 Raman attachment (Madison, WI, USA) (Figure 1). To reduce the chance of fluorescence and sample decay, the longer wavelength, lower

energy Nd:YAG laser was used as the excitation source (1064 nm) [1]. The power of the laser was adjustable and tended to have small fluctuations throughout the experiment. The stability of the laser power was studied over a period of time as shown in Figure 2. It was thus determined that the spectrometer should warm up for approximately 90 min before use for a more stable signal. The laser power for the collected spectra ranged from 2.4 – 2.5 W on average. The sample and sample holder were placed on a platform that was controlled by a Haydon Switch and Instruments, Inc. Stepper Motor (Waterbury, CT, USA) to allow for automatic and/or manual sample alignment. This was controlled through the OMNIC software program (Thermo Scientific, Madison, WI, USA) with alignment by use of a standard as well as manual control to optimize the sample placement to maximize the Raman signal intensity. The Raman scattered radiation caused by the interaction from the laser and the sample was efficiently focused onto the aperture through a series of optical filters (Kaiser Optical Systems, Inc., Ann Arbor, MI, USA), lenses, and mirrors (Thermo Fisher Scientific, Madison, WI, USA). The adjustable aperture in which the Raman scattered radiation travels was set to 100% of the maximum aperture area. With a larger aperture the signal to noise ratio was reduced, thereby allowing more of the scattered radiation to reach the detector. To obtain higher resolution spectra, a smaller aperture is ideal – however, when the Raman signal is weak, a larger aperture can produce a spectrum that has more detail and a stronger signal than would be seen with a smaller aperture [2]. A resolution of 8 cm^{-1} and aperture of 100% provided the optimal tradeoff between a high signal-to-noise ratio and sufficient spectral detail.

After the light traveled through the aperture, it entered the interferometer which makes it possible to measure all frequencies at once using a single channel detector [1,3]. The Raman scattered radiation first is divided by an extended potassium bromide beam splitter (XT-KBr, Thermo Scientific, Madison, WI, USA) that was able to cover frequencies from 11,000 – 375 cm^{-1}. The moving mirror in the interferometer allowed the split beams to recombine and interfere constructively or destructively, depending upon the position of the mirror in space. The speed of the moving mirror provides information on the position of the mirror at a specific time. For the experiments conducted herein, the moving mirror had a velocity equal to 0.3165 cm/s. The constructive or destructive interference then traveled to the indium gallium arsenide (InGaAs) detector, which is a photovoltaic type single channel detector. To obtain a strong signal without compromising the detector, the gain was set to 4 V. The Fourier Transform of the time domain spectrum into the frequency domain spectra was conducted through the OMNIC software installed on a Pentium III personal computer (Optiplex GX150, Dell, Dallas, TX, USA).

2.1.3 Raman Software

Spectra were obtained using OMNIC software (Thermo Scientific, Madison, WI, USA) and saved in the CSV ("comma separated values") format for exportation into Excel (2010, Microsoft, Redmond, WA, USA). For spectra that required digital smoothing, a combination of a moving average (11 point) and a Savitzky−Golay cubic filter (23 point) were used [4]. Singular Value Decomposition (SVD) was conducted by

Professor Sarah Patch of the UWM Physics Department using MatLab Version 2013A (Mathworks, Natick, Massachusetts, USA).

2.1.4 Sample Preparation

To obtain a representative variety of spectra for the database, both mono- and di-carboxylic acids were used as well acids that ranged from short (C1) to long (C20) chain length, as shown in Table I. The samples were stored at room temperature (RT) in opaque polyethylene or glass bottles. Small quantities of the samples (less than 1 mL) were placed into Raman sample tubes (glass, 0.5 mm O.D., 3 in long) (Thermo Fisher Scientific, Madison, WI), wiped clear of impurities on the outside of the tube, and placed into the sample holder. The acquired spectra consisted of an average of 64 scans with a resolution of 8 cm^{-1}. All analyses were performed in triplicate with the resulting spectra averaged. Before use the Raman Spectrometer would be allowed to warm up for approximately 30 min and was then aligned using a standard sulfur (solid) sample.

2.2 Gas Chromatography (GC)

2.2.1 Reagents

All reagents used were AR grade or better. The organic acids studied are listed in Table II. Dilute samples were also prepared using 18 MΩ-cm deionized water. Glassware was washed and then soaked for at least 48 hr in 5% (v/v) HNO_3 (AR grade, Fisher Scientific) followed by rinsing with 18 MΩ-cm deionized water.

2.2.2 GC Instrument

The GC instrument used for the determination of the organic acids was a Varian Model 3800 (Walnut Creek, CA, USA) gas-liquid chromatograph (Figure 3). The system consisted of the following components: solid-phase microextraction (SPME) apparatus (Supelco, Bellefonte, PA, USA), 75 µm carboxen-polydimethylsiloxane (CAR/PDMS) SPME fiber (Supelco), Varian Model 1079 split/splitless injector, Varian wall-coated open tubular (WCOT) fused silica (CP-SIL 24 CB) capillary column (30 m x 0.25 mm x 0.25 µm), Varian pulsed flame photometric detector (PFPD) (Figure 4), and Model R5070 photomultiplier tube (PMT) detector (Hamamatsu, Bridgewater, NJ, USA) set to 600 V with a 200 mV trigger level. The helium, hydrogen and air used for the GC and PFPD were ultra-high purity (99.99999% v/v) gasses (Praxair, Milwaukee, WI and Airgas, West Allis, WI). The helium was used at a constant flow of 2.0 mL/min, Air 1 was set to 17.0 mL/min, Air 2 was set to 10.0 mL/min, and hydrogen was set to 13.0 mL/min (all flow rates were maintained by using electronic flow control). For quantitative determinations, the injection port was in the splitless mode and held at 250°C for 1.0 min for the desorption of the analytes from the SPME fiber. After the initial minute, the injector was switched to a 1:100 split ratio. The temperature program consisted of the oven being initially set to a temperature of 30°C and held for 1 min, and then ramped at 25°C/min to 200°C, where the temperature was held for 1 min, thereby resulting in a total run time of 8.80 min. The PFPD was held at 200 °C for the entire run with a 0.5 ms gate delay, and the gate width was set to 20.0 ms.

2.2.3 GC Software

GC instrument control and data acquisition were performed on a Pentium II personal computer (Optiplex GX1, Dell, Dallas, TX, USA) using Varian's Saturn software version 5.21 and Varian PFPD analysis software version 1.0.

2.2.4 GC Sample Preparation

Dilute acids were prepared in a 25% (w/v) NaCl (ACS grade, Fisher Scientific) by first weighing NaCl into a 2 mL clear glass vial (National Scientific Company) and then adding 18 MΩ-cm deionized water (Thermo Scientific Nanopure system) and Glacial Acetic Acid (ACS grade, Acros, Bridgewater, New Jersey, USA) (Table III).

A sample of the acid (0.50 mL) was transferred into a 2 mL clear glass vial (National Scientific Company, Rockwood, TN, USA) and closed with a Teflon/rubber seal crimp top. The vial was placed on a hot plate set to 30°C as a way to ensure constant temperature conditions between day-to-day analyses. The fiber was exposed to the headspace of the vial for 20 min before being transferred to the injection port.

The SPME fiber used was a medium polarity fiber that has been reported to work well for the determination of acetic acid in the gas phase (i.e., headspace) [5]. The procedure for headspace SPME was as follows: first, the hollow stainless steel needle of the SPME holder was used to puncture the spectrum of the sample vial. Next, the CAR/PDMS-coated SPME fiber was exposed to the volatile components in the headspace region above the liquid phase in the sample vial. After a 20 min sampling

period, the fiber was retracted into the needle. The stainless steel needle was then used to puncture the septum of the GC injection port and the fiber was then re-exposed so that the volatile components could desorb from the fiber into the injection port [6]. The process is depicted schematically in Figure 5.

2.3 Ageing Chamber

2.2.1 Chamber Design

A schematic drawing of the ageing chamber system is shown in Figure 4. This chamber was an adaptation of a "kinetics chamber" that was designed and built by a previous student (Aaron Roerdink) as part of his dissertation research [7]. The original chamber was made of quartz with deactivated fused silica capillary tubing attached to stainless steel fittings to connect it to the Varian 3800 capillary GC. For this work, a door was added to the chamber to introduce a sample into the chamber (Figure 6).

2.2.2 Ageing Experiments

To test the use of the ageing chamber FT-Raman spectra were collected of "pristine" samples, as well as those that were authentically aged and artificially aged using the chamber. The "pristine" samples include those of pure Linseed Oil that had been applied to a glass coupon (approximately 1 inch squared) and dried in the fume hood overnight. The dried samples were removed from the glass coupon and placed in a

Raman sample tube and analyzed using the same method as described above. The authentically aged sample includes a dried, crust-like piece of Linseed Oil that was removed from the cap of a canister of Linseed Oil (purchased at a local hardware store) that had been in a garage for a couple of years. The artificially aged samples consisted of pure Linseed Oil that had been dried onto a glass piece and then placed in the ageing chamber and artificially aged. Two different experiments were conducted. One consisted of ageing the sample for 24 hours at 100°C and 24 hours under UV light (UVGL-58 Handheld UV Lamp, UVP, Upland, Ca, USA) at 254 nm and 6 W. The other experiment consisted of ageing the sample at varying times (2 hr, 4 hr, and 6 hr) at 100°C without UV light.

Table I: The standard carboxylic acid samples that were studied by Raman spectroscopy. The standards were mono-carboxylic acids, di-carboxylic acids, and medium- and long-chain fatty acids.

Acids	***IUPAC Name***	***Formula***
Formic Acid	Methanoic Acid	HCOOH
Acetic Acid	Ethanoic Acid	CH_3COOH
Propanoic Acid	Propanoic Acid	CH_3CH_2COOH
n-Butyric Acid	Butanoic Acid	$CH_3(CH_2)_2COOH$
Valeric Acid	Pentanoic Acid	$CH_3(CH_2)_3COOH$
Caprioc Acid	Hexanoic Acid	$CH_3(CH_2)_4COOH$
Enanthic Acid	Heptanoic Acid	$CH_3(CH_2)_5COOH$
Caprylic Acid	Octanoic Acid	$CH_3(CH_2)_5COOH$
Pelargonic Acid	n-Nonanoic Acid	$CH_3(CH_2)_7COOH$
Lauric Acid	Dodecanoic acid	$CH_3(CH_2)_{10}COOH$
Palmitic Acid	Hexadecanoic Acid	$CH_3(CH_2)_{14}COOH$
Stearic Acid	Octadecanoic Acid	$CH_3(CH_2)_{16}COOH$
Oleic Acid	(9Z)-Octadec-9- enoic acid	$C_{18}H_{34}O_2$ C18:1 cis-9
Linoleic Acid	cis, cis-9,12-Octadecadienoic acid	$C_{18}H_{32}O_2$ C18:2 cis-9,12
Linolenic Acid	cis,cis,cis-9,12,15-Octadecatrienoic acid	$C_{18}H_{30}O_2$ C18:3 cis 9,12,15
Oxalic Acid	Ethanedioic Acid	HOOC-COOH
Malonic Acid	Propanedioic Acid	HOOC-(CH_2)-COOH
Succinic Acid	Butanedioic Acid	HOOC-$(CH_2)_2$-COOH
Glutaric Acid	Pentanedioic Acid	HOOC-$(CH_2)_3$-COOH

Table II: Organic acids studied by the SPME GC method. These included mono-carboxylic acids ranging from one carbon atom to nine carbon atoms.

Acids	***IUPAC Name***	***Formula***
Formic Acid	Methanoic Acid	HCOOH
Acetic Acid	Ethanoic Acid	CH_3COOH
Propanoic Acid	Propanoic Acid	CH_3CH_2COOH
n-Butyric Acid	Butanoic Acid	$CH_3(CH_2)_2COOH$
Valeric Acid	Pentanoic Acid	$CH_3(CH_2)_3COOH$
Caprioc Acid	Hexanoic Acid	$CH_3(CH_2)_4COOH$
Enanthic Acid	Heptanoic Acid	$CH_3(CH_2)_5COOH$
Caprylic Acid	Octanoic Acid	$CH_3(CH_2)_5COOH$
Pelargonic Acid	n-Nonanoic Acid	$CH_3(CH_2)_7COOH$

Table III: Design of the experiments for constructing a calibration model for the SPME GC method.

Concentration of Acetic Acid (v/v %)	***Mass of NaCl (mg)***	***Volume of H_2O (mL)***	***Volume of Acetic Acid (mL)***
0	0.13	0.50	0.00
20	0.10	0.40	0.10
40	0.08	0.30	0.20
60	0.05	0.20	0.30
80	0.03	0.10	0.40
100	0.00	0.00	0.50

Figure 1: Schematic diagram of the Thermo Nicolet Nexus 670 FT-IR with Nexus 900 Raman attachment including a Michelson Interferometer [60] (M = mirror, MM = moving mirror, FM = fixed mirror) [8].

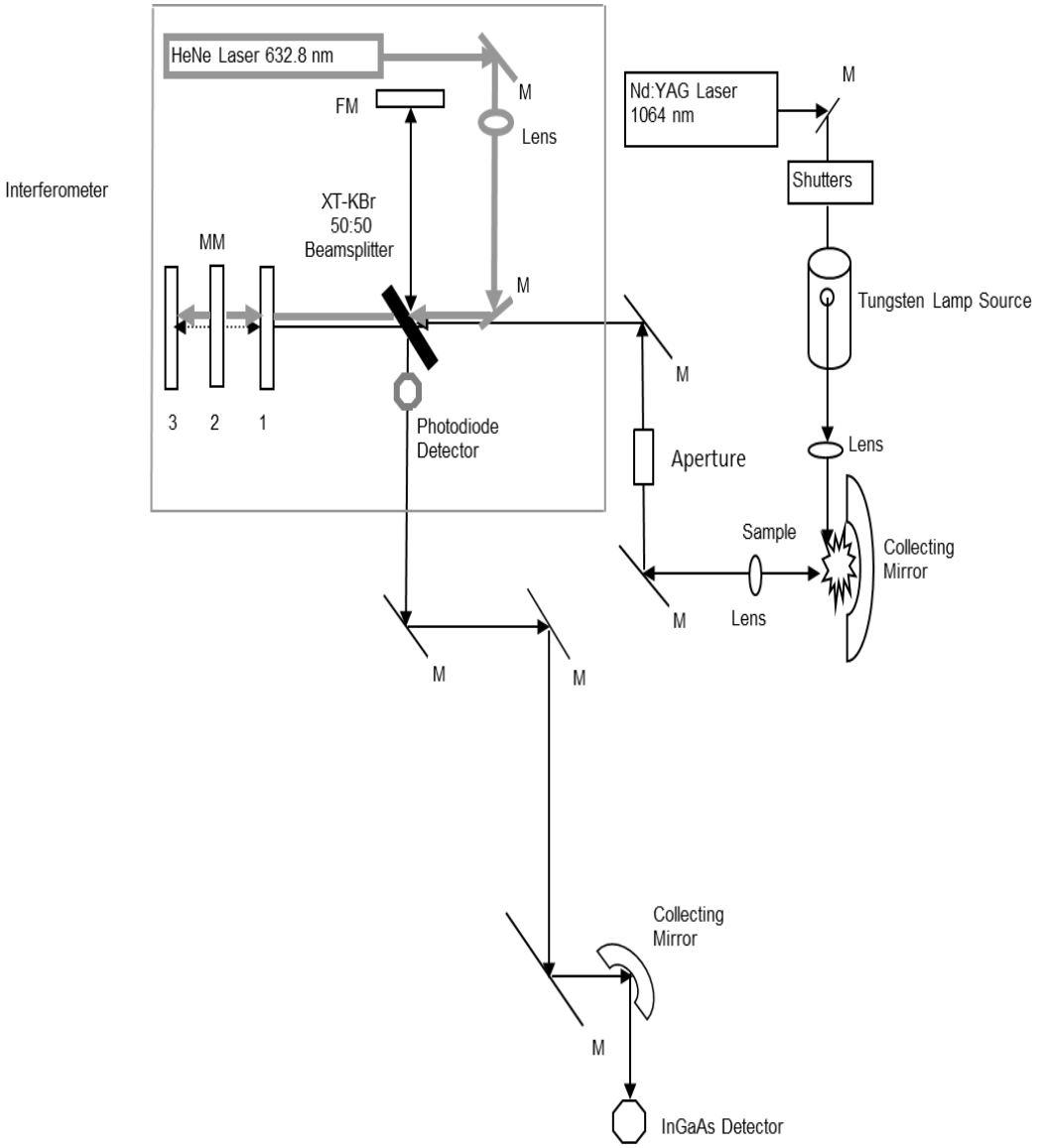


Figure 2: The changes in Raman intensity of three shifts in the sulfur standard were monitored to study the stabilization time of the instrument.

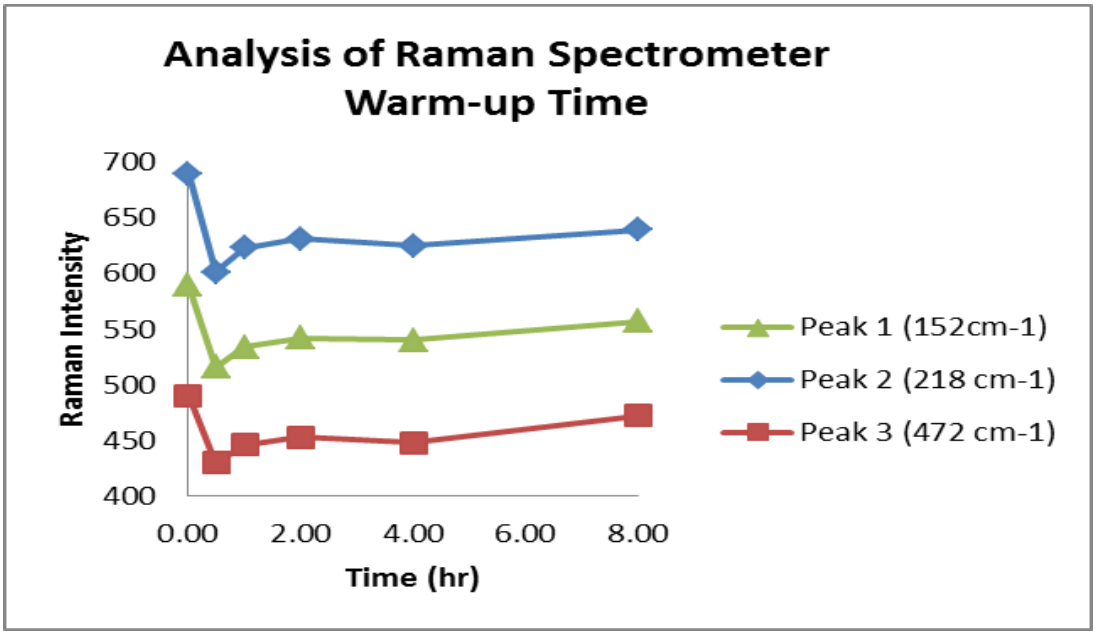


Figure 3: Schematic diagram of the Varian 3800 capillary GC equipped with a pulsed flame photometric detector [7].

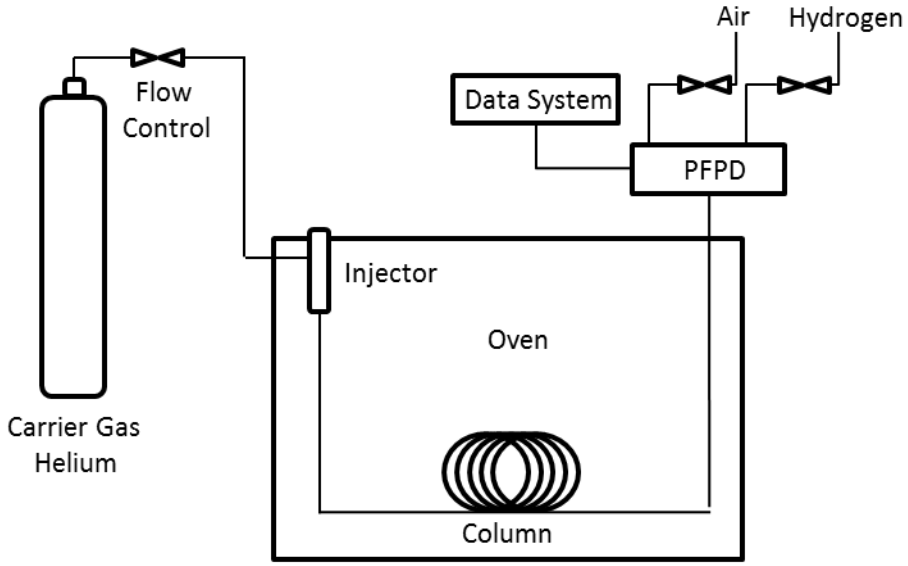


Figure 4: Schematic diagram of a pulsed flame photometric detector [60] (PMT = photomultiplier tube).

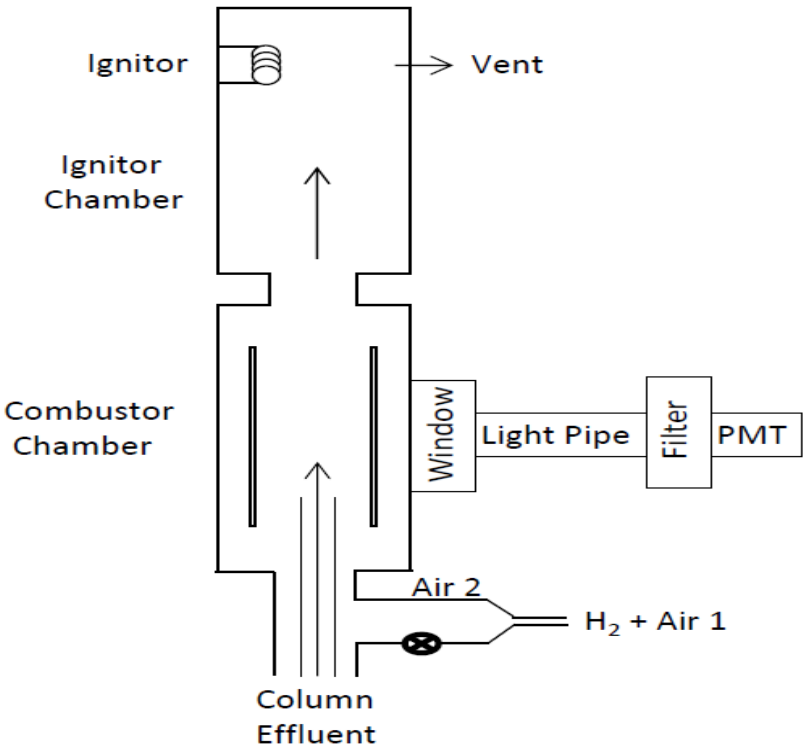


Figure 5: The headspace solid-phase microextraction procedure. For details, see text [6].

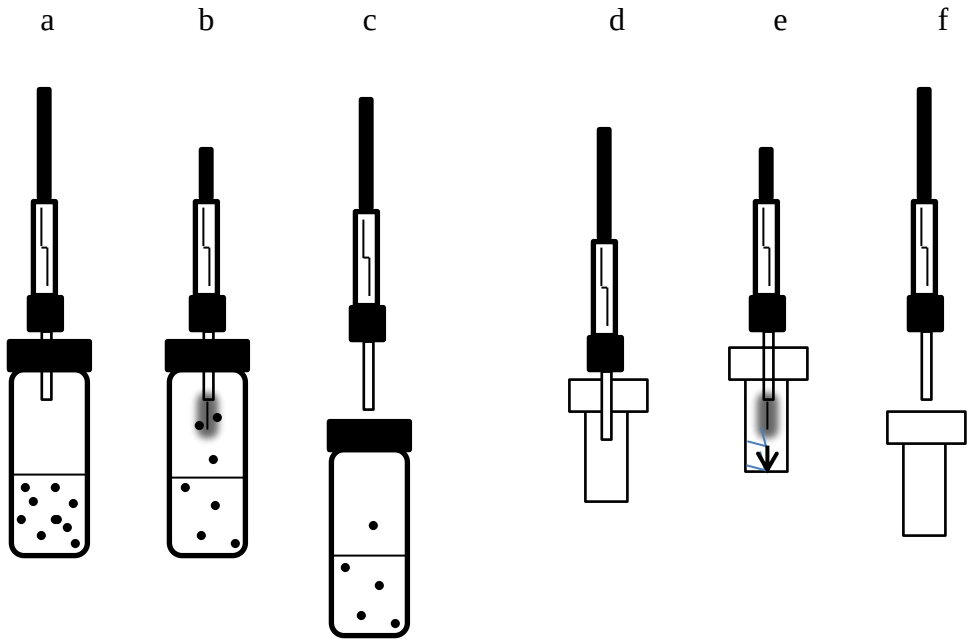
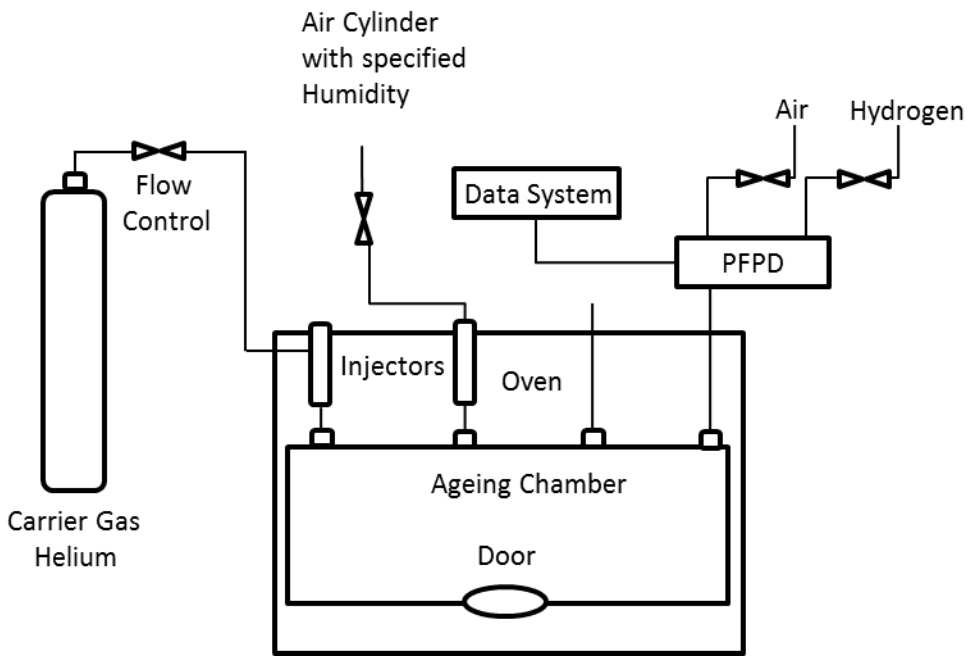


Figure 6: Schematic diagram of the ageing chamber (installed in the GC).



References

(1) Skoog, D. A.; Holler, F. J.; Crouch, S. R.,Raman Spectroscopy In *Principles of Instrumental Analysis*, Kiselica, S., Ed.; Thomson Brooks/Cole: Canada, 2007, pp 481-497.

(2) Wiebolt, D. Understanding Raman Spectrometer Parameters, *Spectroscopy,* 2010, 1-12.

(3) Chase, B., FT-Raman Spectroscopy: A catalyst for the Raman explosion?, *Journal of Chemical Education,* 2007, *84*, 75-80.

(4) Savitzky, A.; Golay, M. J. E., Smoothing and differentiation of data by simplified least squares procedure, *Analytical Chemistry,* 1964, *36*, 1627-1639.

(5) Godoi, A. F. L.; Van Vaeck, L.; Van Grieken, R., Use of solid-phase microextraction for the detection of acetic acid by ion-trap gas chromatography-mass spectrometry and application to indoor levels in museums, *Journal of Chromatography A,* 2005, *1067*, 331-336.

(6) *Solid Phase Microextraction: Theory and Optimization of Conditions*; Supelco, 1998, pp 1-8.

(7) Roerdink, A. R. *Studies of Organoarsenical Speciation and Reactivity in the Environment: Development of Novel Optical and Mass Spectrometric Methods*. University of Wisconsin-Milwaukee, Milwaukee, WI, 2004.

(8) *FT-Raman Module User's guide*; Therno Nicolet Corporation: Madison, WI, 1999, pp 1-86.

Chapter 3: Results and Discussion

3.1 Raman Spectral Database

3.1.1 Introduction

One of the main questions addressed in this dissertation was to determine if Fourier Transform Raman Spectroscopy (FT-Raman) could be used to develop a select database of carboxylic acids as a means to better understand the types of organic structures that are present in artistic media. To accomplish this, a variety of carboxylic acid standards and unknowns were studied by FT-Raman Spectroscopy. To compare spectra obtained for unknown samples to those found in the database, Singular Value Decomposition (SVD) was used as a "fingerprinting technique" as a means to determine which standard or combination of standards were components of the unknown samples.

3.1.2 Method Optimization

Optimization of the method conditions for creation of the FT-Raman spectral database was as follows: sample size was less than 1 mL, laser power was approximately 2.5 W, number of scans averaged was 64, resolution was 8 cm^{-1}, PMT gain was set at level 4, and the aperture was set at 100% of the maximum aperture area (Table I). The liquid samples were approximately 0.4 mL and the solid samples were approximately 0.1 g. The laser power tended to fluctuate but averaged to approximately 2.5 ± 0.1 W

throughout each day. An approximately 60-90 min "warm-up" time was required to properly align the interferometer and thereby stabilize the signal. Resolution was also optimized to create a sufficient amount of spectral detail without introducing too much noise. A resolution of 8 cm^{-1} was determined to be the best level because some of the FT-Raman intensities were relatively low for certain samples. Therefore a relatively lower resolution was used to increase light throughput and consequently allow for more intense Raman scattering to occur. The PMT gain setting of 4 was selected to produce the greatest intensity at a potential difference across the dynodes which would not hasten the demise of the detector. Finally, the aperture of 100% was set to correspond to a resolution of 8 cm^{-1} (set automatically). By choosing a specific resolution, the aperture automatically adjusted to transmit the maximum amount of light.

3.1.3 Collection of Standard Spectra

FT-Raman spectra were collected for 19 standard samples which included: mono-carboxylic acids, di-carboxylic acids, and medium- to long-chain fatty acids (Table II). "Medium -chain" fatty acids were defined as carboxylic acids with an 8-14 carbon atom chain length, whereas "long-chain" fatty acids were defined as carboxylic acids with 16 or more carbon atoms in chain length [1]. The mono-carboxylic acids (C1-C9) and di-carboxylic acids (C2-C5) that were studied are listed in Table II; their structures are shown in Figures 1 and 2, respectively. The structures of the medium- and long-chain fatty acids are shown in Figure 3. All of the samples were measured in the pure (neat) physical state at room temperature. Formic, Acetic, Propanoic, Butyric, Valeric, Caproic,

Enanthic, Caprylic, Pelargoni, Oleic, Linoleic, and Linolenic Acids were examined in the liquid state. Oxalic, Malonic, Succinic, Glutaric, Lauric, Palmitic, and Stearic Acids were studied in the solid state. Dilute samples of each of the standards were also studied, but the spectra obtained were often of only the solvent. When a subtraction of the solvent spectrum (both through the OMNIC program and manually through Excel) was performed from the spectrum of a dilute solution of carboxylic acid, Raman peaks were not discernible in the spectrum. These experiments were tried with both organic and aqueous solvents. For the organic solvents, the spectrum was just that of the solvent. With the aqueous solvents, a 1:1 mixture of acetic acid yielded interpretable spectra. However, for concentrations lower than 50% (v/v), peaks were not discernible. This was both unexpected and disappointing given that one of the (presumed) advantages of Raman compared to IR is that water is a very weak Raman scatterer. Thus it was generally difficult to obtain good FT-Raman spectra of the standard in a dilute solution in either organic or aqueous solvents.

As a way to compare the spectra of the unknown samples to spectra of standard samples, it was imperative that all samples were studied using the same method factors. Table III lists all of the observed frequency shifts for the standard spectra. For example, Formic Acid (Figure 1 in Appendix A) had five major frequency shifts (Table IV). These frequency shifts are similar to those reported in the literature [2-4]. Note that the variety of instrumental and method factors used in different studies account for the small variations in frequency observed in this study. For example, Glutaric Acid (Figure 13 in Appendix A) is a di-carboxylic acid with major frequency shifts that are close to those observed for Formic Acid, but there are some differences. There are eight major

frequency shifts for Glutaric Acid as shown in Table IV. The O-H out of plane bend arises from dimers of the carboxylic acid. Carboxylic acids that are in condensed states at high concentration have been shown to exist in a dimeric state because of the strong hydrogen bonding which can occur between the carbonyl and hydroxyl groups in adjacent carboxylic acid molecules [4,5]. The shifts observed at 445 cm^{-1} and 322 cm^{-1} are difficult to assign vibrational modes because there is a lack of understanding about bands found at low wavenumber values [6]. It should also be noted that while Glutaric Acid and Formic Acid have the same functional groups, there are some frequency shifts observed for Glutaric Acid but not for Formic Acid. There are two possible reasons for this phenomenon: (a) molecules have many bond stretching modes (for example, Glutaric Acid has twice as many C=O bond stretches than Formic Acid), and (b) Raman scattering probabilities ("cross-sections") are a function of molecular structure and phase [7]. Glutaric Acid was studied as a solid, whereas Formic Acid was studied as a liquid. For Linolenic Acid (Figure 16 in Appendix A) six major frequency shifts were identified, as listed in Table IV. The frequency shift at 1657 cm^{-1} could either be a C=O stretch or a C=C stretch, but because this frequency shift is much larger for the unsaturated long-chain fatty acids than observed for the saturated fatty acids like Palmitic Acid (Figure 18 in Appendix A), a C=C stretch probably obscured the C=O stretch. Thus a variety of frequency shifts were observed for each class of carboxylic acid standards – similar trends in the observed frequency shifts are summarized in Table III.

3.1.3 Collection of Unknown Spectra

Along with collecting spectra for the standard carboxylic acids, a select group of varnishes were studied as well. Two groups of varnishes were studied: drying oils and resins. The drying oils were chosen because they have been common varnishes in art since the thirteenth century [8,9]. Another reason for their choice is that it has been shown that the oxidation of the drying oils can lead to small molecular weight products, such as mono- and di-carboxylic acids [9]. Resins are thick, sticky, water-insoluble substances that are excreted by certain trees and plants. Natural resins are most often composed of mixtures of di- or tri-terpenoids; they are commonly used in art work as a coating or adhesive [9].

The two drying oils that were studied were Linseed Oil and Tung Oil (the latter is also known as China Wood Oil). Drying oils are composed of triglycerides, which are individual long-chain fatty acids attached to a glycerol backbone (Figure 4). There has been much research and many theories as to which fatty acids and in what order they attach to the glycerol, but there is no definitive answer yet [9]. With the different components in each drying oil attaching to the glyceride backbone at random, there are many possible combinations, thereby creating a far greater number of triglycerides. While the percentage of each component in a mixture of long-chain fatty acids can vary for a given chemical formula, the individual components are in fact known. An example of the composition of Linseed Oil [9,10] and Tung Oil [9,11] are shown in Figures 5 and 6, respectively. Linseed Oil is composed of Linolenic, Linoleic, Oleic, Palmitic, Stearic, Arachidic Acids, along with a small component comprised of other various long-chain fatty acids. Tung Oil is composed of Eleostearic, Linoleic, Palmitic, and Oleic Acids.

The FT-Raman spectra of Linseed Oil and Tung Oil are shown in Figures 7 and 8, respectively, and the frequency shifts are shown in Table V. For Linseed Oil, the major frequency shifts include: 2906 and 2856 cm^{-1} which is a C-H stretches, 1657 cm^{-1} which is a C=C stretch, 1441 cm^{-1} which is a C-H bend, and 1263 cm^{-1} which is a C=O stretch. Like Linolenic Acid, the stretch at 1657 cm^{-1} could either be a C=C stretch or a C=O stretch, but it is assumed it is a C=C stretch because of its more prominent presence in the unsaturated long-chain fatty acids. For Tung Oil, the frequency shifts include: 2933 and 2871 cm^{-1} which are C-H stretches, 1456 cm^{-1} which is a C-H bend, and 1306 cm^{-1} which is a C-O stretch. Upon a visual comparison of standard spectra from the database to that of Linseed Oil, Linolenic Acid is the closest match, which makes sense because it is the most abundant component of Linseed Oil. However, it is not a perfect match, as can be seen by off-setting these two spectra (Figure 9). A slightly different shape to the C-H stretch frequency shifts as well as for the peak at 1745 cm^{-1} is seen in the Linseed Oil spectrum; there are also several peaks near 1000 cm^{-1} that can be seen in the Linolenic Acid spectrum. Linseed Oil is a very complex mixture, so the similarity of the unknown and standard spectra is high but not perfect, nevertheless allowing one to achieve a better understanding of the possible components that are present. This will also be useful when comparing aged samples to those found in the spectral database. Not only can one examine how the spectrum of fresh Linseed Oil differs from that of aged Linseed Oil, but one can use the database to identify degradation products of Linseed Oil.

Resins comprised a second class of varnishes that were studied. The two resins studied, Dammar and Mastic, are the two most common resins used in art conservation [9,12,13]. Both of these resins are tri-terpenoids (composed of 30 carbon atoms) and are

derived from isoprene [9,14]. Figure 4 shows an example of one of the components found in Dammar [9]. Dammar was first used as a varnish in the 19th Century and while the use of Mastic as an early form a chewing gum is dated to the Ottoman Period (14th Century), it is unknown when this resin was first used as a varnish for artistic purposes [9]. Much like the drying oils, the composition of these varnishes is complex and the specific composition is largely unknown [9]. Because these materials are traditionally in the solid state under ambient conditions, two different types of samples were prepared and then studied by FT-Raman spectroscopy. The first sample type involved grinding the crystals into a powder while the second type of sample preparation involved heating a sample of the powder in a glass sample tube and then allowing it to solidify into a clear solid. Because there was little difference in the spectra obtained from the two methods, the solidified sample was used for comparison. It was thought that the solidified sample would most likely resemble how the varnish would be applied to a piece of art work. The spectra of Dammar and Mastic are shown in Figures 10 and 11, respectively, and the frequency shifts are listed in Table V. The major frequency shifts for Dammar include: 2933 cm^{-1} which is a C-H stretch, 1657 cm^{-1} which is either a C=O stretch or a C=C stretch, and 1452 cm^{-1} which is a C-H bend. For the band observed at 1657 cm^{-1}, given the complex composition of Dammar it is most likely the result of a combination of C=O and C=C stretches. The same effect is a likely explanation for the frequency shift observed at 1660 cm^{-1} in the Mastic spectrum. Other major frequency shifts observed in the Mastic spectrum include: 2929 cm^{-1} which is a C-H stretch, and 1444 cm^{-1} which is a C-H bend. Although the FT-Raman database of carboxylic acid standards does not feature known components found in Dammar and Mastic, it can still be used for

comparison. For instance, when comparing the Dammar spectrum to that observed for the Enanthic Acid standard, one can see many similarities (Figure 12). When the spectra are superimposed, it is easy to see that while many of the frequency shifts are identical, there are still some shifts in the "fingerprint region" that do not align very well. Thus while these spectra are very similar, in Dammar there may be a small amount of Enanthic Acid, but interpretation of these spectra is difficult given the complex composition of Dammar.

3.1.4 Singular Value Decomposition

Singular values are representations of reduced dimensionality of the original data [15]. The optimal number of singular values were selected from the "scree plot" shown in Figure 13. There are multiple ways to determine the optimal number of singular values. One way is to pick the point where the remaining values are small and approximately the same size. These remaining singular values account for a small amount of variation in the data, so they can be excluded. The larger the singular value, the more variance that value represents in the data. As one can see in Figure 13, the downward slope ends approximately between singular values of 6 and 7. These values were then compared to the 19 standards in the Raman Spectral Database (Figure 14). Clearly, using a singular value of 6 does not fit the “typical” standard as well as using 7 singular values, as is obvious from the small negative peak that was observed at approximately 2800 cm^{1}. Therefore a singular value equal to 7 was chosen.

By using SVD, the many thousands of data points in the Raman Spectral Database are reduced to 7 singular values. These values were compared to the four unknowns (Linseed Oil, Tung Oil, Dammar and Mastic), and only small errors were found (Figure 15). Consider the spectral range between 2600cm^{-1} to 3200cm^{-1} as shown in Figure 16. The root-mean square (RMS) error was used as a measure of how well the model based upon 7 singular values compared to the unknown data. The RMS errors were 0.08, 0.13, 0.21, and 0.21 Raman Intensity units for Linseed Oil, Tung Oil, Dammar and Mastic, respectively. If those errors are compared to the largest peak found in the unknown spectra, the percent relative RMSs were 1.7%, 1.7%, 4.9%, and 6.4%, respectively. These are fairly small errors, thus indicating that a singular value of 7 was optimal in accounting for much of the variability in the database while at the same time reducing the dimensionality of the data. This indicates that the Raman Spectral Database, as represented by the 7 singular values, predicts the "unknown" spectra effectively. In further application of the database to other unknown varnishes, a re-consideration of the optimal choice of the singular value data would be necessary to best fit the database to the unknown spectra.

One of the drawbacks of using singular values is that this approach assumes that all of the components of the data are orthogonal (i.e., independent of one another). When the samples are too similar, such as observed in some of the Raman spectra of standards, negative fits of the standards to the unknown(s) can be observed. As an example, Figure 17 shows the SVD fits of standards from the database to a mixture of four carboxylic acids: Formic Acid, Acetic Acid, Heptanoic Acid, and Nonanoic Acid. When separate standards for these four acids are compared to the four-component mixture, a negative

amount of Nonanoic Acid is calculated. However, this is still useful information as one could be able to determine that unknown "Q" is x% A, y% B, and z% (C and/or D). While, this isn't perfect, it still gives a relative idea of the unknown's makeup. This is just one example of what can occur when standards are too similar in structure. Thus the SVD approach is effective because instead of using every data point, the optimal singular value creates a vector that accurately represents the data while reducing its dimensionality. Consequently, the RMS error will also decrease in the process, as can be seen in a comparison of Figure 15 to Figure 17.

3.2 Gas Chromatographic Characterization

3.2.1 Introduction

To incorporate the ageing chamber directly into the Gas Chromatograph (GC) seamlessly, it was necessary to create a method to make sure that the by-products of the aged varnish could be measured using GC with a pulsed flame-photometric detector (PFPD). This section describes the optimization of this method, along with presenting results for both qualitative and quantitative analyses. A method for determining acetic acid using headspace SPME with GC-MS in museums was reported recently [16]. The goal of the research presented herein was to extend the method of Godoi *et al.* to larger carboxylic acids.

3.2.2 Method Optimization

Table VI shows the experimental factors that were identified for the method optimization. Absorption time, desorption time, split ratio, and the oven temperature program were optimized. Other factors, such as fiber type, film thickness, and injector temperature were not studied because these factors had been extensively studied by Godoi *et al.* and others [16,17]. In head-space SPME, the absorption (or partitioning) time is the amount of time that the SPME fiber is exposed to the volatile components in the headspace of the liquid [17]. One of the important issues with absorption time is that it needs to be long enough to for equilibrium to be reached between the components absorbed into the fiber and those in the headspace, as well as the equilibrium between the analyte in the aqueous phase and the analyte in the volatile phase [17]. These relationships are described by Equation 4-1:

$$n = (C_0V_1V_2K)/(KV_1 + K_2V_3 + V_2) \qquad \text{Equation 4 – 1}$$

where n is the mass of the analyte that is absorbed by the fiber coating; C_0 is the initial concentration of the analyte in the aqueous phase; V_1, V_2, and V_3 are the volumes of the stationary phase coating on the fiber, the aqueous solution, and the headspace, respectively; K is the partitioning coefficient of the analyte between the fiber coating and the aqueous solution; and K_2 is the partitioning coefficient between the volatile headspace phase and the aqueous phase. K is actually the product of K_1 and K_2, where K_1 is the partitioning coefficient between the volatile headspace phase and the stationary phase coating on the fiber [17,18].

Different absorption times for neat Glacial Acetic Acid were tested: 0, 10, 15, 20, and 30 min (Figure 18). Each absorption time was run in triplicate and an average was plotted along with error bars determined by correcting the standard deviation by using Student's t-value (two-tailed, 95% confidence). For the 0 min absorption time analyses, there were not any definite peaks in the resulting chromatogram. The retention time for the peak associated with the 10 min absorption was 3.57 ± 0.08 min and the rest of the retention times for the other absorption time experiments were 3.7 ± (0.03, 0.03, 0.09) min. To include the signal for 0 min desorption time, all of the retention times between 3.45-3.54 min were averaged and the composite value was used. After examining the trends in the data, an optimal absorption time of 20 min was chosen. While Figure 18 shows that a 15 min absorption time seems to be sufficient for all of the volatile compounds to partition into the fiber, 20 min was chosen to ensure that equilibrium had been reached. Note that samples were thermostatted at 30°C to make sure that the equilibrium was established under constant conditions

The next factors optimized were the desorption time and gradient elution (temperature) program conditions. The desorption time is the time for the analyte to desorb from the fiber while in the GC injection port. This was controlled by the split ratio of the mobile phase flow. For instance, for a 1 min desorption time, the flow was set at "splitless" until 1 min into the run and then it was changed to a 1:100 split (i.e., to vent any remaining sample). The desorption conditions were reflected in the temperature program for the oven, i.e., it was held at 30°C for 1 min before the temperature gradient was initiated (25°C/min to 200°C). Desorption times of 0, 1, 2, 4, and 8 min were tested using neat Glacial Acetic Acid. Each desorption time was studied in triplicate and an

average was plotted along with error bars as before (Figure 19). There was overlap of the error bars in the 1-2 min time frame, but because a desorption time of 1 min gave the largest PMT signal, that level was chosen as the optimal desorption time.

The last factor that was optimized was the split ratio. Because capillary columns have low capacity, only a small amount of sample can be injected. Three different split ratios were tested: 1:100, 1:50, and 1:10. Chromatograms for these split ratios are shown in Figure 20. The intensity signals for both 1:50 and 1:10 split ratios were very similar, however when increasing the split ratio not surprisingly the signal decreases. For a 1:100 split ratio and a 1.0 min desorption time, the signal was nevertheless satisfactory, as seen in Figure 21.

3.2.4 Calibration Models for Acetic Acid

Quantitation of Acetic Acid was done by calibration using a first-order regression fit. The model yielded: $y = 0.29x + 0.92$ (R^2=0.95). Using a second-order fit (Figure 22), a relationship of: $y = 0.0025x^2 - 0.0016x + 5.9$ (R^2=0.99). This determination was conducted for Acetic Acid to compare results to the Godoi *et al.* study [16]. To quantify other carboxylic acids, separate calibration models would have to be created.

Increasing the ionic strength (μ) will often create a condition in solution that will improve the extraction of organic analytes ("salting out") [19]. Therefore, a calibration model was constructed for Acetic Acid in a 25% (w/v) NaCl solution. The calibration model obtained is shown in Figure 23. Curiously, the sensitivity did not increase and the linear fit of the model (as embodied in R^2) was worse – a 9.6% decrease compared to low

μ conditions. Thus increasing μ can help improve the extraction when the fiber is immersed into solution, but it does not improve the method when the headspace is sampled because increasing the ionic strength of a solution reduces the solubility of the analyte. Therefore by adding the salt, the fraction of the analyte in the headspace was reduced.

3.2.4 GC Structure-Retention Relationship

Godoi and co-workers [16] specifically studied the levels of Acetic Acid in the ambient air at The Rubens House, an art museum in Antwerp, Belgium. Because the oxidation products of drying oils are low molecular weight carboxylic acids, it is imperative to have a method that determines not only levels of Acetic Acid but also other low molecular weight carboxylic acids. This was accomplished by using the method described in the previous section for Acetic Acid in the headspace. The carboxylic acids studied were: Formic, Acetic, Propanoic, Butyric, Valeric, and Caproic Acid. After collecting chromatograms of these carboxylic acids, it became evident that there was a linear relationship (R^2= 0.95) between the number of carbons in the mono-carboxylic acid and the retention time (Figure 24). The relationship among Formic, Acetic, Propanoic, and Butyric Acid is relatively linear. However, after Butyric Acid, the retention times continue to increase but with a more second-order dependence. Mono-carboxylic acids with greater than six carbon atoms were also studied; they did not follow a linear trend which was not surprising given their low volatility. The relationship between the retention time of the mono-carboxylic acids and their boiling points are shown in Figure

25 [20]. Figure 24 will be useful to determine the type of mono-carboxylic acid that is present in an aged (authentic) sample. By using the retention time of the sample and the linear relationship that was observed ($y = 0.75x + 1.55$, R^2=0.95), it will be possible to identify the mono-carboxylic acid that is present. Then by using the calibration model as shown in Figure 23, it will be possible to quantify the specific mono-carboxylic acid as well.

3.3 Ageing Chamber

3.3.1 Design of the Ageing Chamber

As a way to address some of the concerns that arose in a workshop sponsored by the National Science Foundation and the Andrew W. Mellon Foundation [21], an ageing chamber was designed and fabricated. In this report it was stated that: "We must prepare standard test materials with known properties, measure them, and encourage communication between labs to establish these standards." This idea of a standard reference material (SRM) relates to all areas for cultural heritage research. Without SRMs one can neither validate a method nor compare a method to others in the field. It was also stated that: "We must have a robust baseline framework from which to design analytical tests, interpret findings, and attempt to predict the effects of age and treatment on these objects" [21]. One of the ways to achieve this is through the creation and use of SRMs. One of the reasons that the conservationist community does not have SRMs is that they would need a significant mass of an authentically aged sample for an SRM program to be practical. This is inherently a problem because minimally invasive

analytical measurements are imperative for monetary reasons [21]. Another reason that the conservationist community does not have SRMs is that most pieces of art contain material of unknown identity. Not only is the unknown identity challenging from an analytical chemistry standpoint, but also many pieces of art have been restored multiple times, thereby changing the original composition of the material [22-24].

There are a broad range of restorative materials for the various types of objects, with varnishes of primary interest to conservators. To create a SRM for varnishes, ideally one would use an authentically aged sample, but as mentioned before there is the problem of limited sample size and the identity of the original varnish. The design and fabrication of the ageing chamber described in this study was intended to address the need for SRMs. The ageing chamber was built to be directly incorporated into the GC oven, as shown in Figure 26. Incorporation of the ageing chamber into the oven allows simultaneous study of four factors that have primary effects on ageing: heat, UV radiation, atmosphere, and time. As a "proof of concept", an experiment was designed and tested to show that by introducing the sample into the ageing and exposing it to two different environmental factors (heat and light), ageing in a repeatable fashion will occur.

3.3.2 Testing of the Ageing Chamber

To test the ageing chamber, FT-Raman spectra were collected for "pristine" (i.e., unaged) samples, as well as for samples that were "authentically aged" (i.e., under ambient conditions) and "artificially aged" (i.e., by using the chamber). The "pristine" samples were pure Linseed Oil that had been applied to a glass coupon (approximately 1

inch square) and dried in the fume hood for five days. The dried sample was then removed from the glass coupon and placed in a sample tube for Raman analysis as described above. The "authentically aged" sample was a dried, crust-like piece of Linseed Oil that had been removed from the cap of an old canister of Linseed Oil (purchased at a local hardware store ~15 years ago). The "artificially aged" sample consisted of pure Linseed Oil that had been dried onto a glass slide and then placed in the ageing chamber. The sample was subjected to experiments under the following conditions: (a) 24 hours at 100°C in the dark, and (b) 24 hours at 100˚C under intense UV light at 254 nm and 6 W.

3.3.2 Characterization of Linseed Oil by FT-Raman Spectroscopy

To test the use of the ageing chamber FT-Raman spectra were collected of "pristine" samples, as well as those that were authentically aged and artificially aged using the chamber. The "pristine" samples, include those of pure Linseed Oil that had been dried and the Raman spectrum is shown in Figure 27. The intensity is lower than that of the liquid Linseed Oil (Figure 7), but the frequency shifts are the same. To compare that of "pristine" linseed oil with a sample of Linseed oil that had been authentically aged a sample of a dried, crust-like piece of Linseed Oil that was removed from the cap of a canister of Linseed Oil (purchased at a local hardware store) that had been in a garage for a couple of years was studied. The Raman spectrum of the authentically aged sample is shown in Figure 28. To better compare the authentically aged sample with that of the "pristine" sample the two spectra were graphed together in

Figure 29 and the difference was plotted in Figure 30. There are very few differences, except a slight change in Raman Intensity and seen in Figure 29. In Figure 30 one can see the small change in the spectra. The peaks around 3000 cm^{-1} most likely arise from the small difference in the width of the frequency shift, however, the positive peaks around 1300 cm^{-1}, 1450 cm^{-1}, and 1650 cm^{-1} shows that the frequency shifts for the "pristine" samples have greater intensity than those of the authentically aged sample.

3.3.3 Linseed Oil Artificially 'Aged' using the Ageing Chamber

To test that the ageing chamber works, two experiments were performed. The first experiment was to age a sample of Linseed Oil for 24 hours at 100°C and 24 hours under UV light at 254 nm. The Raman spectrum that was collected on the aged sample is shown in Figure 31. To again compare the aged sample to that of "pristine" Linseed Oil Figure 32 shows them graphed together and Figure 33 is the difference spectra. In this case the artificially aged sample has a greater Raman Intensity than the "pristine" sample. In fact, when looking at the difference spectra it is almost a complete mirror image of the Linseed Oil spectra, with only a small positive peak at around 2800 cm^{-1} which could be a result of a small narrowing on the frequency shift. This same trend of the Raman Intensity of artificially aged samples being larger than that of the "pristine" Linseed Oil is evident on the second experiment as well. This experiment consisted of ageing the sample at varying times (2 hr, 4 hr, and 6 hr) at 100°C and the Raman spectra can be seen in Figure 34. As the amount of the ageing time increases so does the Raman intensity. Figure 35 shows the difference spectra of each of the Raman spectra of the aged sample

for the varying time experiment compared to that of the "pristine" Linseed Oil. Again, all of the frequency shifts are negative, indicating that as a sample is aged the Raman Intensity increases. This was not the case with the authentically aged sample, which suggests that either the authentically aged sample was not in fact aged very long, that the authentically aged sample is not a good representation of an actual aged varnish, or that the ageing chamber is not producing samples that are representative of an aged sample. Therefore, to prove that the ageing chamber can be used to create standard reference materials a new authentically aged sample, such as that from a piece of artwork, needs to be studied.

Table I: Instrumental factors for the FT-Raman spectrometer.

Sample Size	< 1 mL
Laser	1064 nm Nd:YAG
Detector	InGaAs
Laser Power	~2.5 W
Number of Scans	64
Resolution	8 cm^{-1}
Gain	4
Aperture	100%

Table II: List of carboxylic acids studied using FT-Raman spectroscopy.

Acids	***IUPAC Name***	***Formula***
Formic Acid	Methanoic Acid	HCOOH
Acetic Acid	Ethanoic Acid	CH_3COOH
Propanoic Acid	Propanoic Acid	CH_3CH_2COOH
n-Butryic Acid	Butanoic Acid	$CH_3(CH_2)_2COOH$
Valeric Acid	Pentanoic Acid	$CH_3(CH_2)_3COOH$
Caprioc Acid	Hexanoic Acid	$CH_3(CH_2)_4COOH$
Enanthic Acid	Heptanoic Acid	$CH_3(CH_2)_5COOH$
Caprylic Acid	Octanoic Acid	$CH_3(CH_2)_5COOH$
Pelargonic Acid	n-Nonanoic Acid	$CH_3(CH_2)_7COOH$
Lauric Acid	Dodecanoic acid	$CH_3(CH_2)_{10}COOH$
Palmitic Acid	Hexadecanoic Acid	$CH_3(CH_2)_{14}COOH$
Stearic Acid	Octadecanoic Acid	$CH_3(CH_2)_{16}COOH$
Oleic Acid	(9Z)-Octadec-9- enoic acid	$C_{18}H_{34}O_2$ C18:1 cis-9
Linoleic Acid	cis, cis-9,12- Octadecadienoic acid	$C_{18}H_{32}O_2$ C18:2 cis-9,12
Linolenic Acid	cis,cis,cis-9,12,15- Octadecatrienoic acid	$C_{18}H_{30}O_2$ C18:3 cis 9,12,15
Oxalic Acid	Ethanedioic Acid	HOOC-COOH
Malonic Acid	Propanedioic Acid	$HOOC-(CH_2)-COOH$
Succinic Acid	Butanedioic Acid	$HOOC-(CH_2)_2-COOH$
Glutaric Acid	Pentanedioic Acid	$HOOC-(CH_2)_3-COOH$

Table III: Frequency shifts for all of the carboxylic acid standards.

Carboxylic Acids	Frequency Shifts								
Formic Acid	2956	1676	1398	1209	681				
Acetic Acid	1672	1429	893	623	445				
Propanoic Acid	2949	1660	1421	1078	847				
Butyric Acid	2941	2879	2744	1660	1452	1043	866	777	357
Valeric Acid	2941	2875	2740	1660	1448	1306	1109	1059	916
	827								
Caprioic Acid	2937	2875	2733	1660	1444	1306	1113	1066	916
	854								
Enanthic Acid	2941	2733	1660	1441	1306	893			
Caprylic Acid	2933	2733	1664	1441	1306				
Pelargonic Acid	2933	2856	2733	1664	1441	1302			
Oxalic Acid	1780	1726	1479	1174	827	542	465		
Malonic Acid	2952	1687	1433	1182	920	766	642	407	
Succinic Acid	2968	2929	2571	1657	1421	1294	1228	1086	935
	685	580	388						
Glutaric Acid	2922	1653	1417	1298	1066	939	667	445	322
Oleic Acid	3006	2898	2856	1653	1441	1302			
Linoleic Acid	3010	2902	1657	1441	1263				
Linolenic Acid	3014	2933	1657	1441	1267	866			
Lauric Acid	2883	2848	2725	1437	1298	1063	893		
Palmitic Acid	2883	2848	2725	1437	1298	1128	1063	893	
Stearic Acid	2883	2848	2721	1437	1298	1128	1063	893	

Table IV: Frequency shifts for Formic, Glutaric, and Linolenic Acid.

Functional Group	***Expected Frequency Shift (Δcm^{-1})***	***Observed Frequency Shift (Δcm^{-1})***
Formic Acid		
C=O stretch	1750-1600 [3]	1676
C-H bend	1398 [3]	1398
C-O stretch	1208 [3]	1209
O-C-O	725-650 [3]	681
Glutaric Acid		
C-H stretch	2943 [2]	2922
C=O stretch	1750-1600 [3,5]	1653
C-H bend	1398 [3]	1417
C-O stretch	1208 [3]	1298
O-C-O stretch	1046 [4]	1066
O-H out of plane bend	920 [4,5]	939
Linolenic Acid		
C-H stretch	3013 [25]	3014
C-H stretch	2967 [4]	2933
C=C stretch	1657 [25]	1657
C-H bend	1439 [25]	1441
C-O stretch	1250 [5]	1267
C-C stretch	1200-800 [4]	866

Table V: Frequency shifts for the unknowns.

Unknown	Frequency Shifts					
Linseed Oil	3014	2906	2856	1657	1441	1263
Tung Oil	3072	2933	2871	2729	1456	1306
Dammar	2933	1657	1452			
Mastic	2929	1660	1444			

Table VI: Experimental factors for the SPME GC method.

Fiber Coating	75 µm Carboxen-PDMS
Sample Size	0.5 mL
Sample Temperature	30°C
Absorption Time	20 min
Desorption Time	1 min
Split ratio	1:100
Stationary Phase	CP-SIL 24 CB, 0.25 µm film thickness
Column dimensions	30 m x 0.25 mm
Mobile Phase	Helium, 99.999% (v/v)
Mobile Phase Flow Rate	2 mL/min
Injector Temperature	300°C
Oven Program	30°C for 1 min, 25°C/min to 200°C, 200˚C for 1 min

Figure 1: Structures of mono-carboxylic acids studied by FT-Raman spectroscopy. The structures are: a) Formic Acid, b) Acetic Acid, c) Propanoic Acid, d) Butyric Acid, e) Valeric Acid, f) Caproic Acid, g) Enanthic Acid, h) Caprylic Acid, and i) Pelargonic Acid.

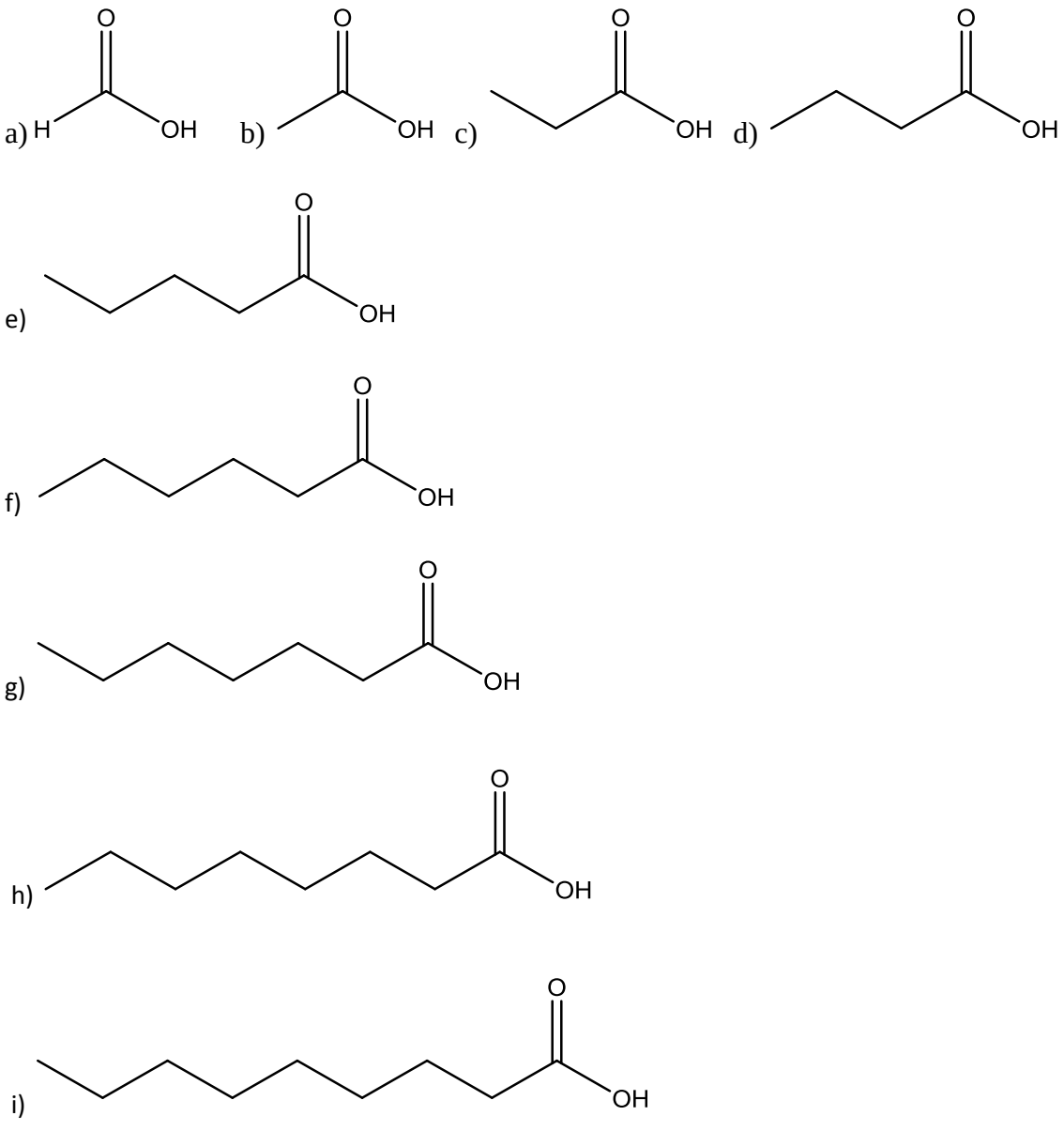
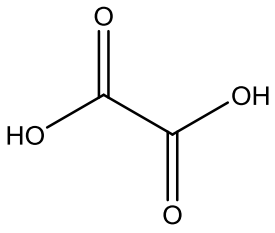
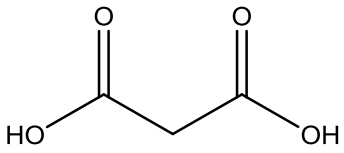


Figure 2: Structures of di-carboxylic acids studied by FT-Raman spectroscopy. The structures are: a) Oxalic Acid, b) Malonic Acid, c) Succinic Acid, and d) Glutaric Acid.

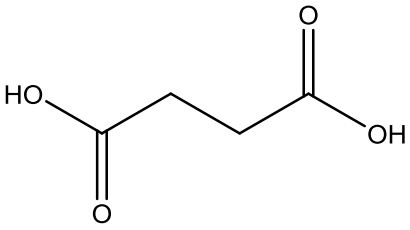
a)



b)



c)



d)

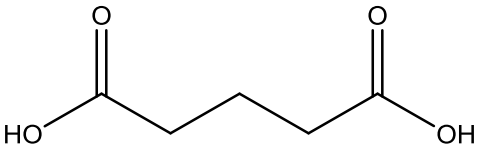


Figure 3: Structures of medium and long-chain fatty acids studied by FT-Raman spectroscopy. The structures are: a) Lauric Acid, b) Palmitic Acid, c) Stearic Acid, d) Oleic Acid, e) Linoleic Acid, and f) Linolenic Acid.

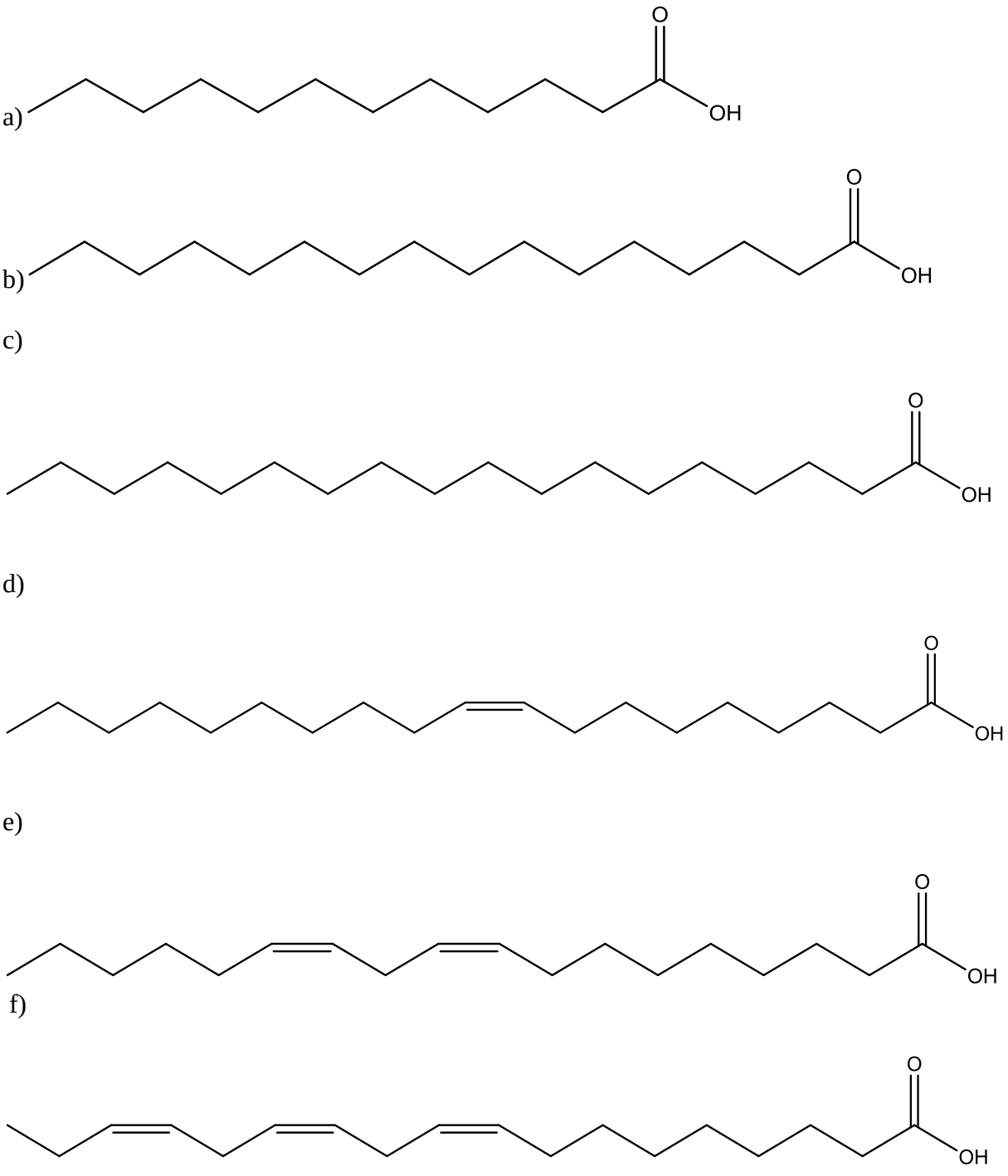


Figure 4: Examples of a triglyceride compound and a triterpenoid compound [9].

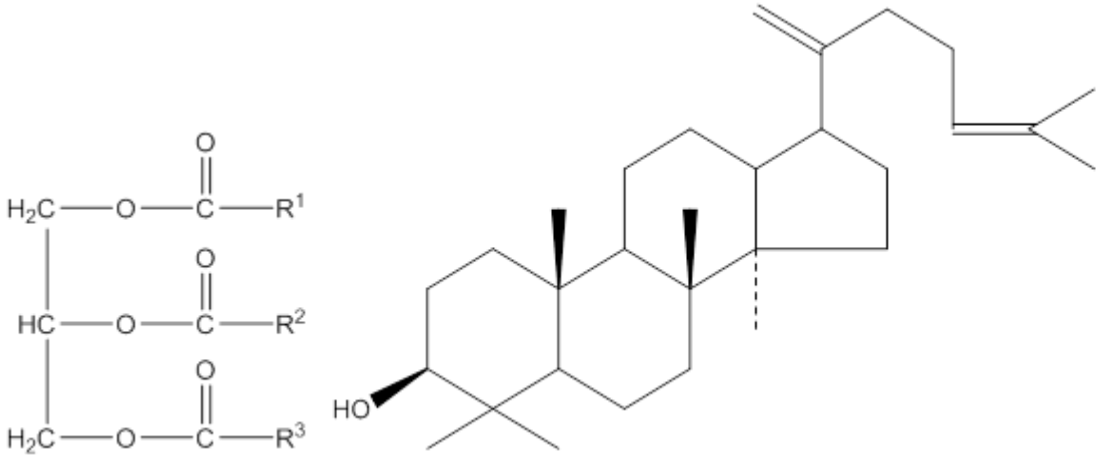


Figure 5: The six major components of Linseed Oil [9,10].

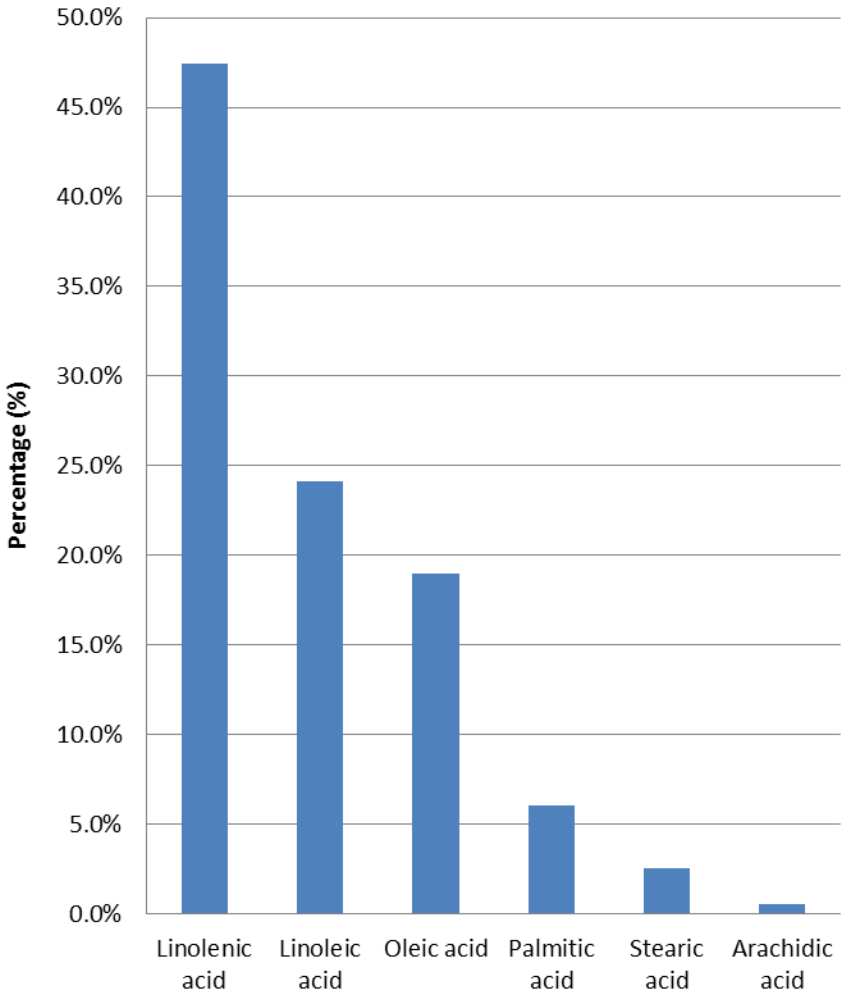


Figure 6: The four major components of Tung Oil [9,11].

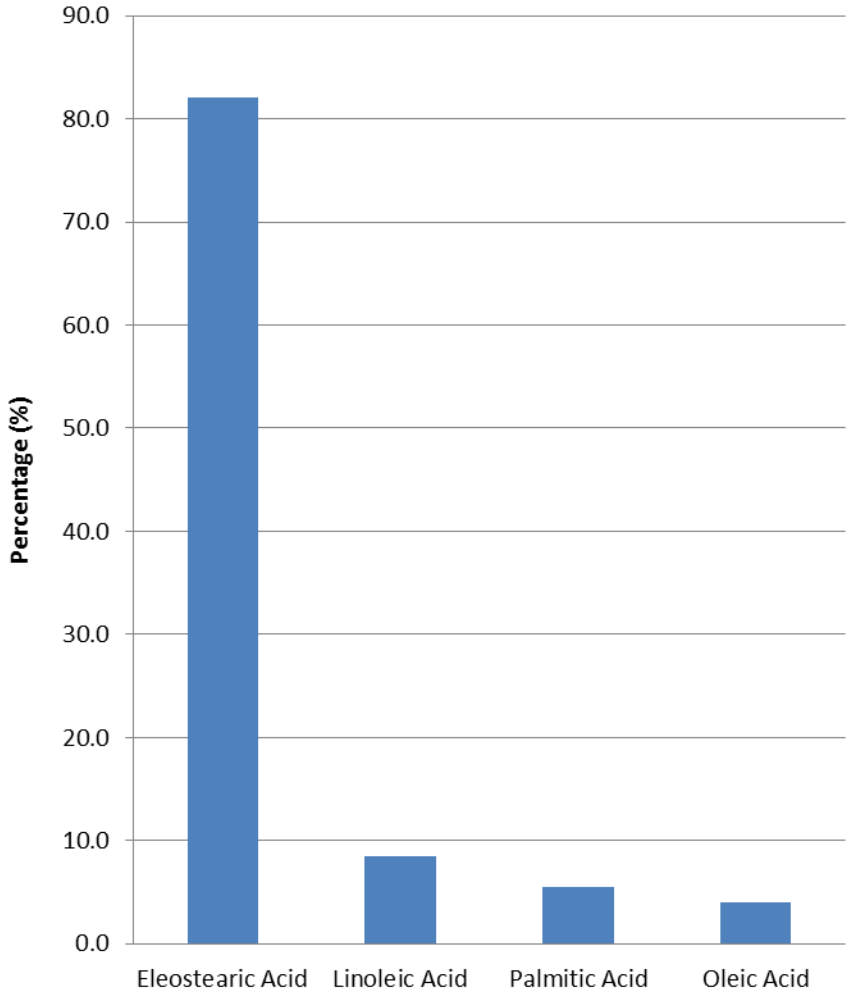


Figure 7: FT-Raman spectrum of Linseed Oil (n = 3).

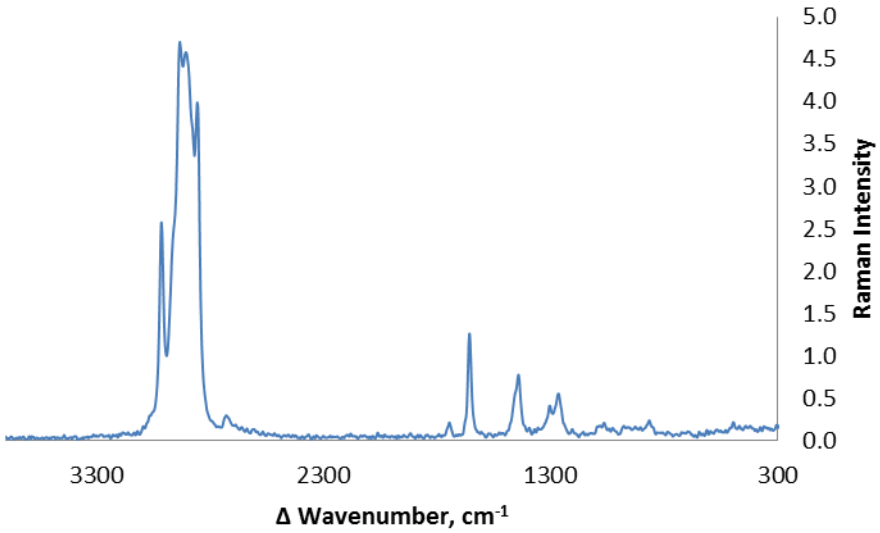


Figure 8: FT-Raman spectrum of Tung Oil (n = 3).

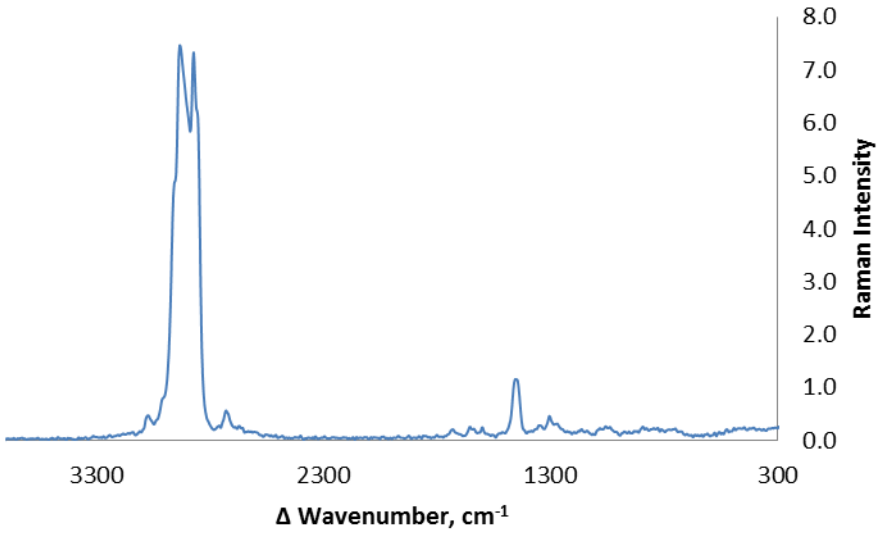


Figure 9: Comparison of FT-Raman Spectra of Linseed Oil to Linolenic Acid. The Linolenic Acid spectrum is off-set by one intensity unit for clarity.

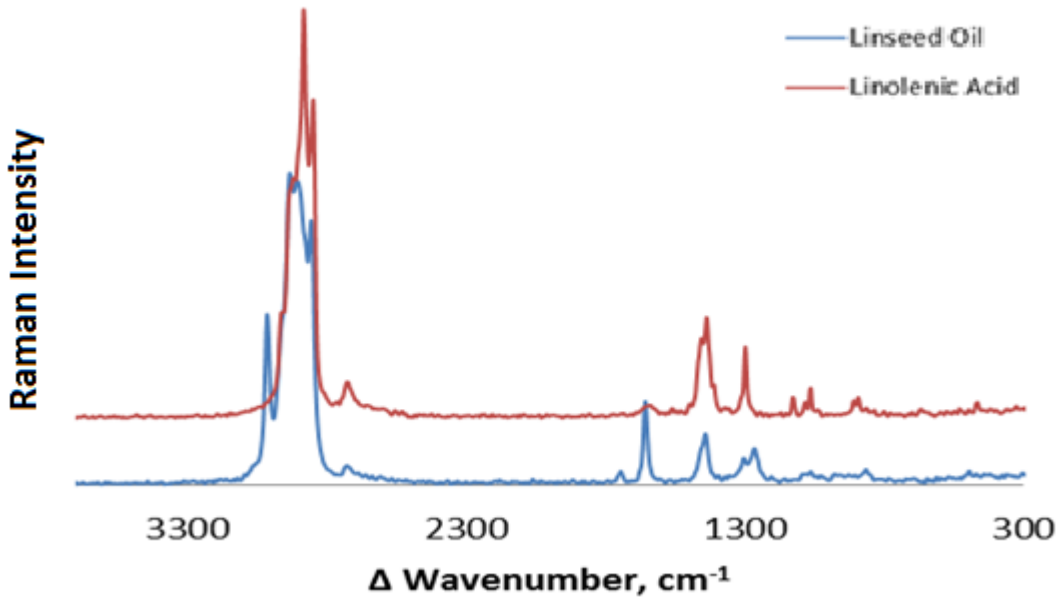


Figure 10: FT-Raman spectrum of Dammar (n = 3).

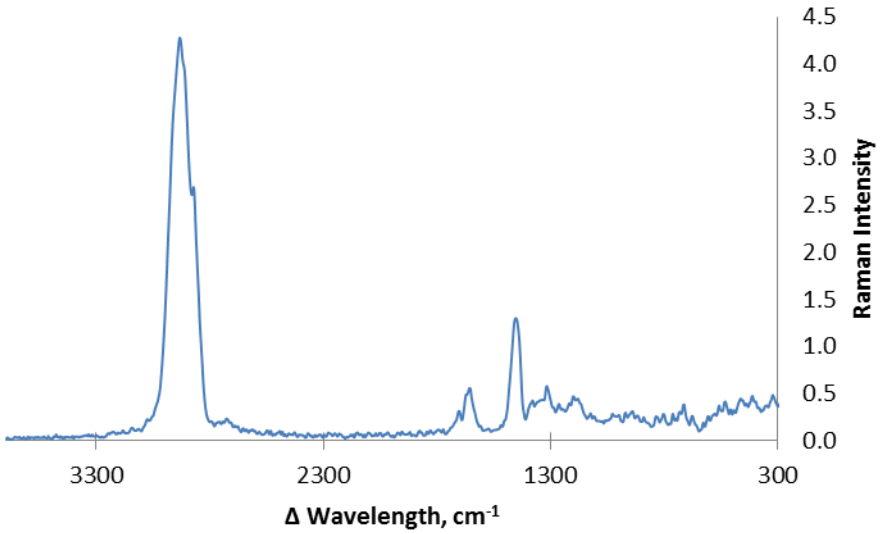


Figure 11: FT-Raman spectrum of Mastic (n = 3).

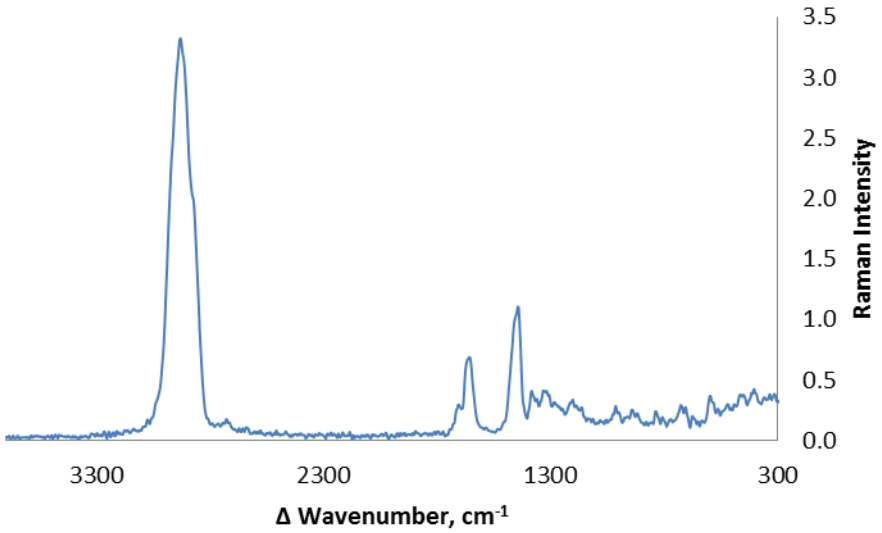


Figure 12: Comparison of FT-Raman spectra of Dammar to Enanthic acid. The Enanthic Acid spectrum is off-set by one intensity unit for clarity.

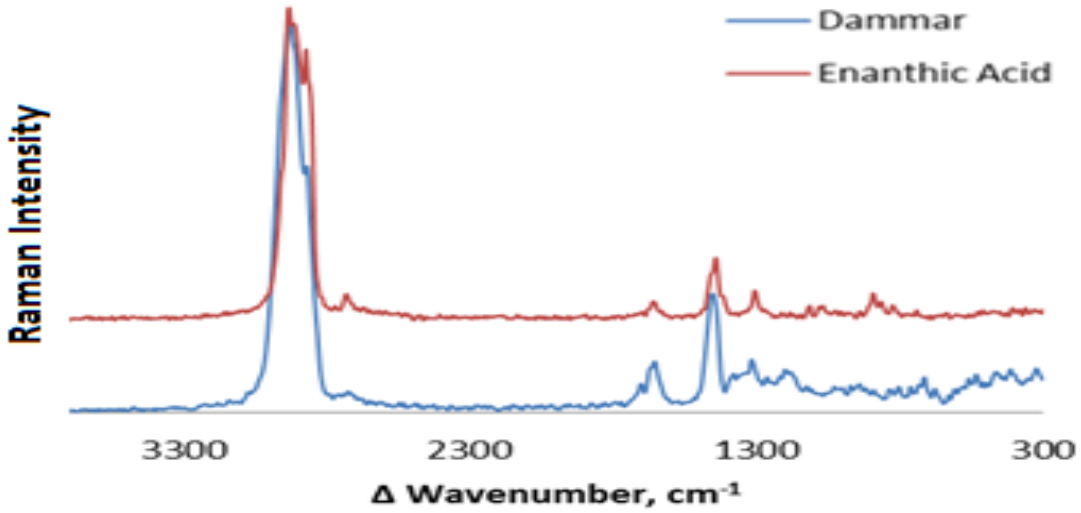


Figure 13: A Scree Plot was created to determine the optimal singular value level to apply to the Raman spectral database of the standard compounds.

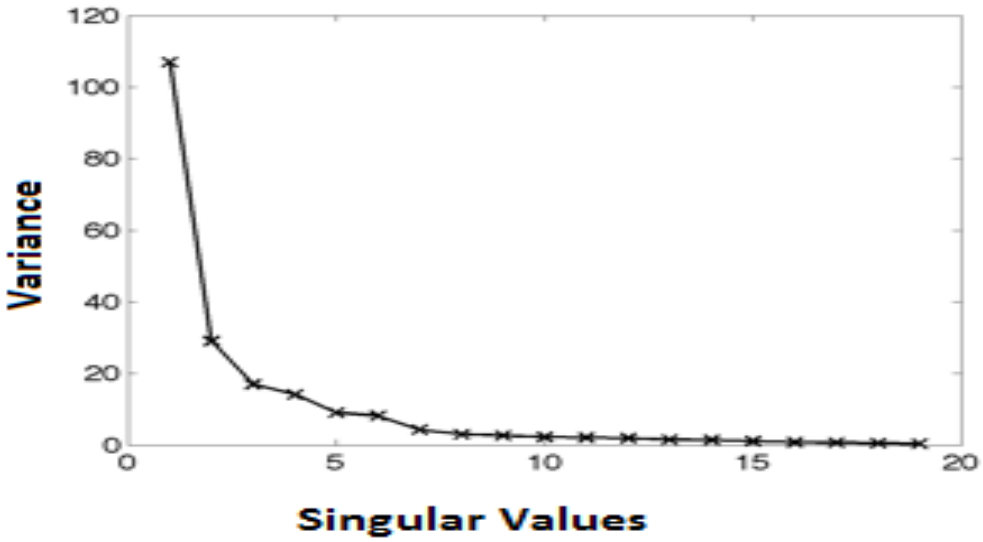


Figure 14: To determine between two singular values a comparison of 6 (red) and 7 (black) singular values were fit to the standards to determine the optimal singular value level to apply to the Raman spectral database of the standard compounds.

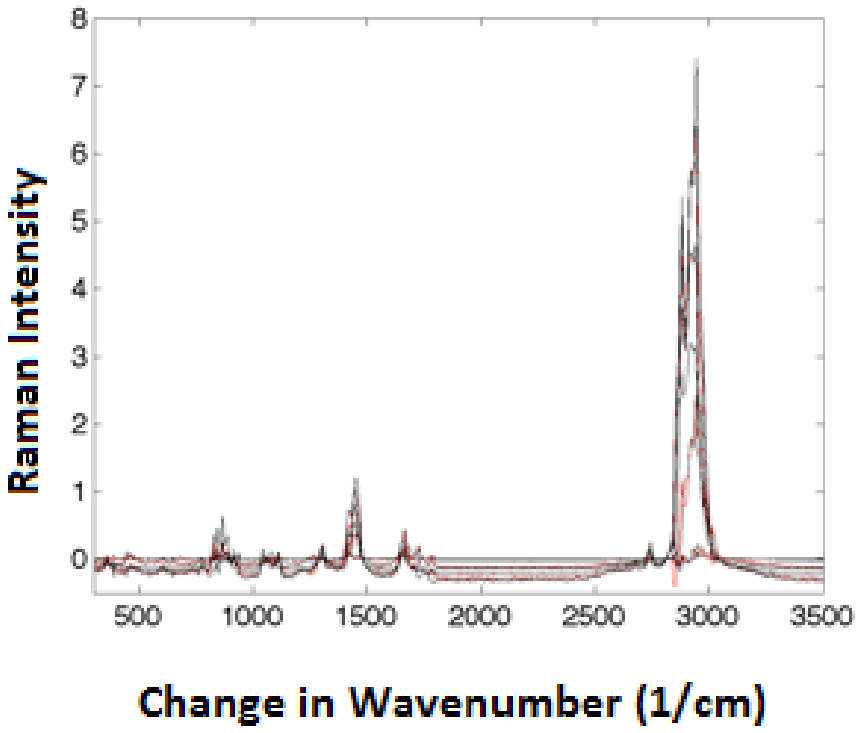


Figure 15: To determine how well the 7 singular values represented the Raman Database they were fit the to the unknowns. The bolded lines are the original unknowns and the thin lines are the modeled fit using 7 singular value.

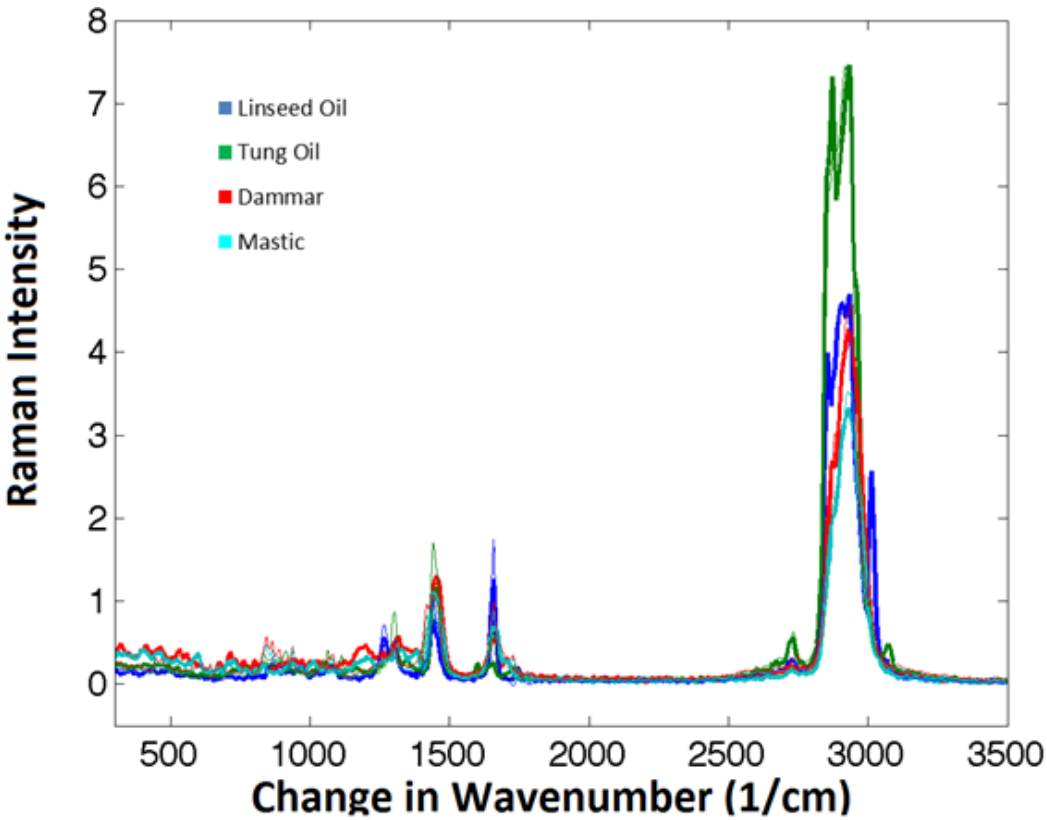


Figure 16: To determine how well the 7 singular values represented the Raman Database they were fit the to the unknowns; specifically focusing on 2600 – 3200 cm^{-1}. The bolded lines are the original unknowns and the thin lines are the modeled fit using 7 singular value.

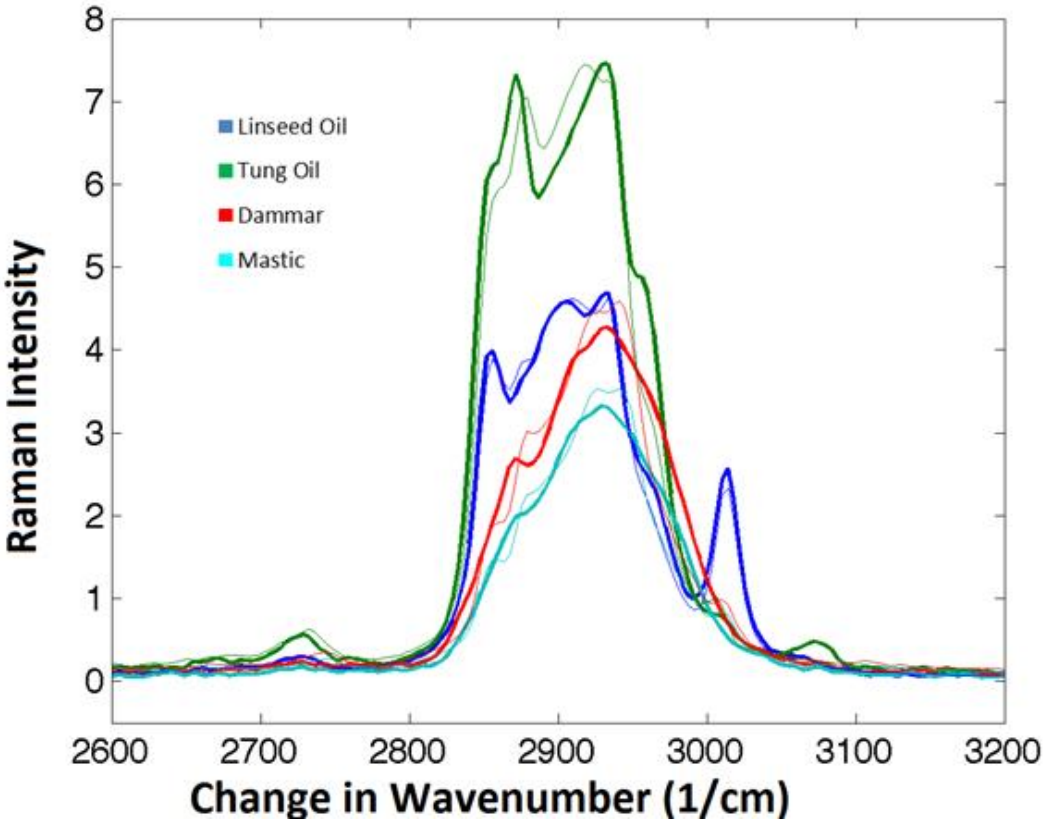


Figure 17: Linear Squares fitting of a known mixture of standards.

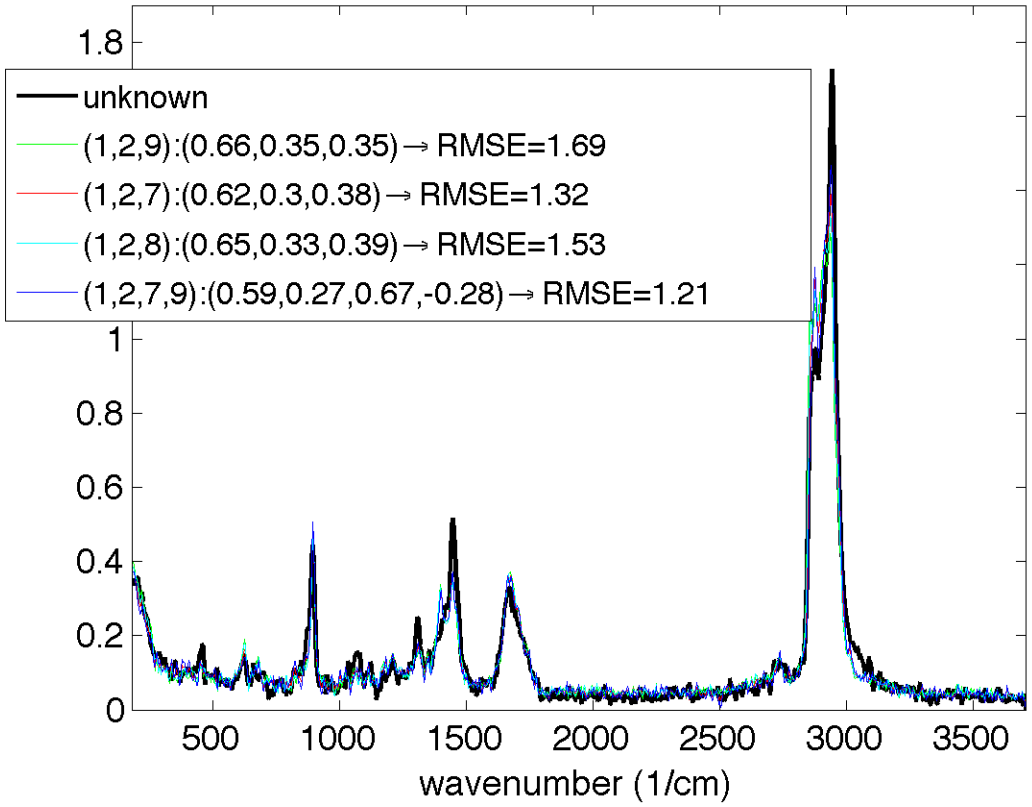


Figure 18: An absorption time study of the Carboxen-PDMS SPME fiber was conducted for the determination of Acetic Acid.

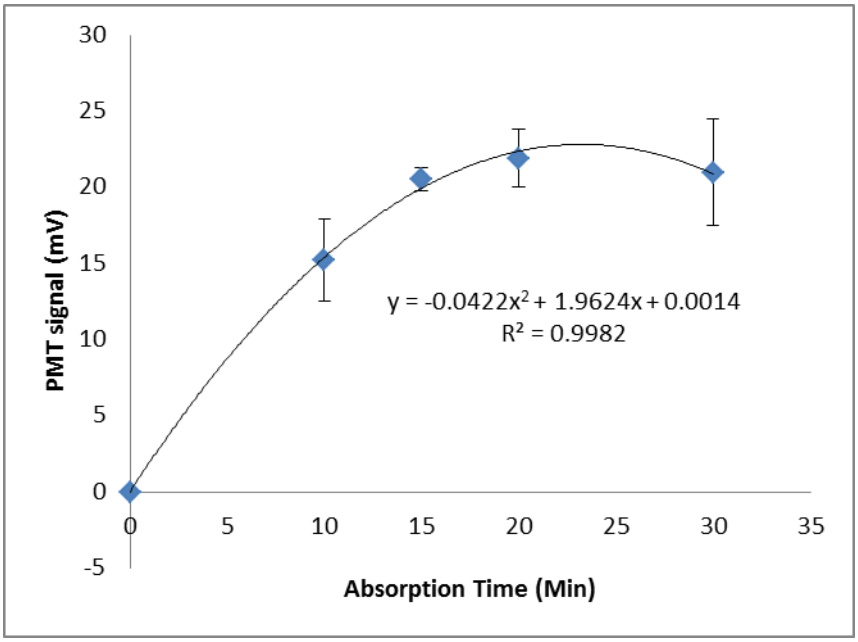


Figure 19: The desorption time for the determination of neat Acetic Acid using a Carboxen-PDMS SPME fiber.

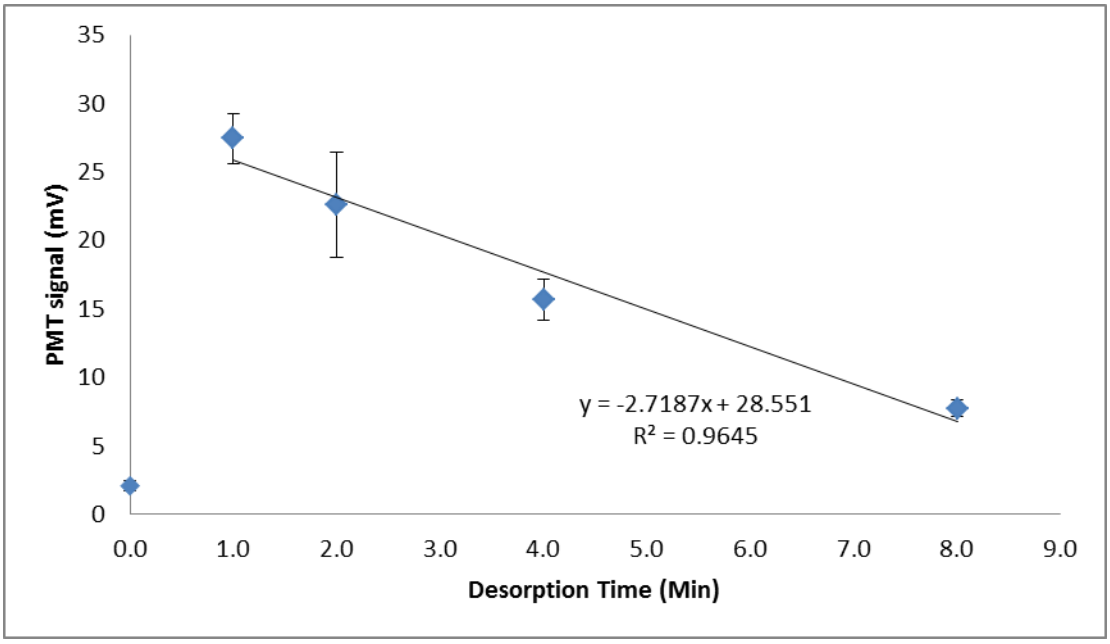
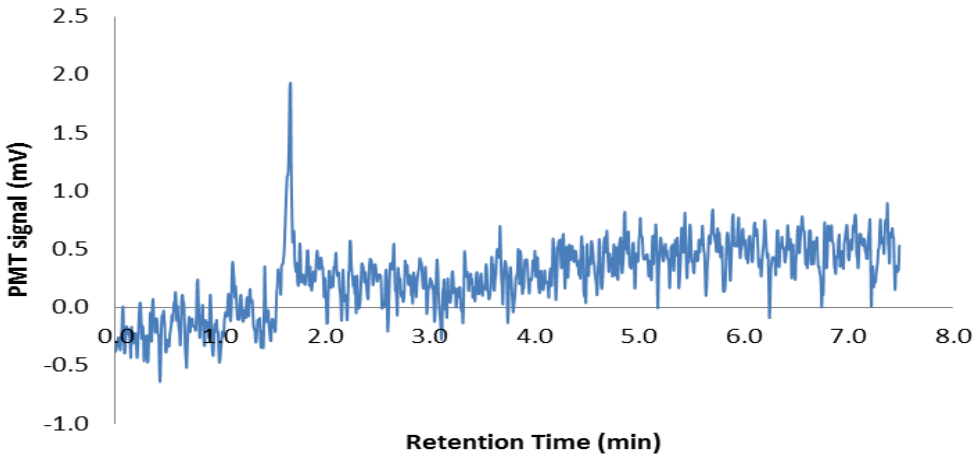


Figure 20: Chromatogram of neat Acetic Acid using a 1:100 split ratio, 1:50 split ratio, and 1:10 split ratio.

a)



b)

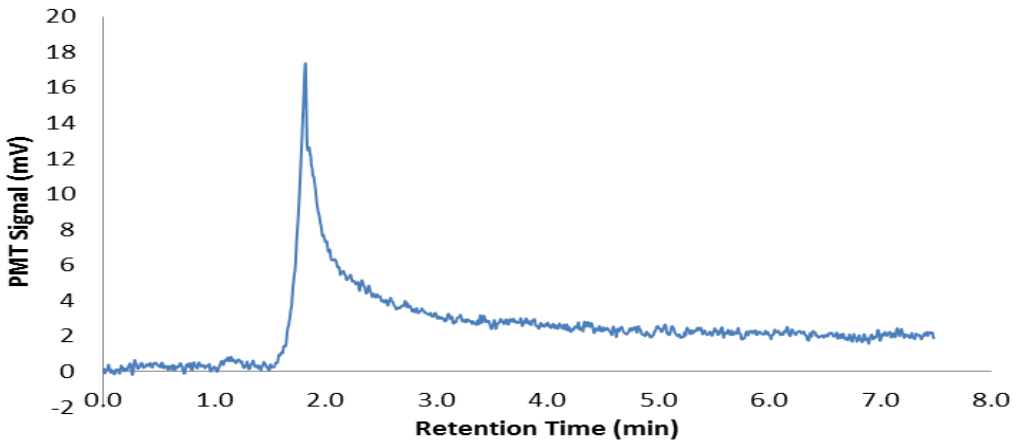


Figure 20: Cont.

c)

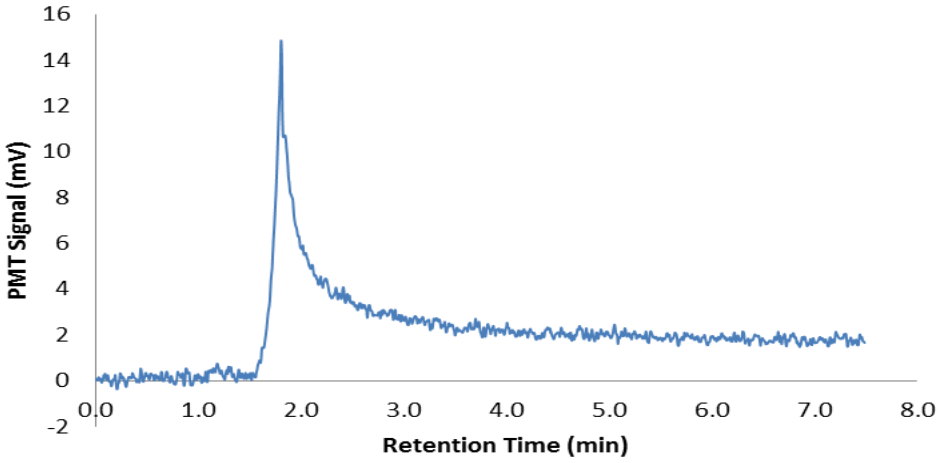


Figure 21: Chromatogram of neat Acetic Acid using a 1:100 split ratio with a 1 min desorption time.

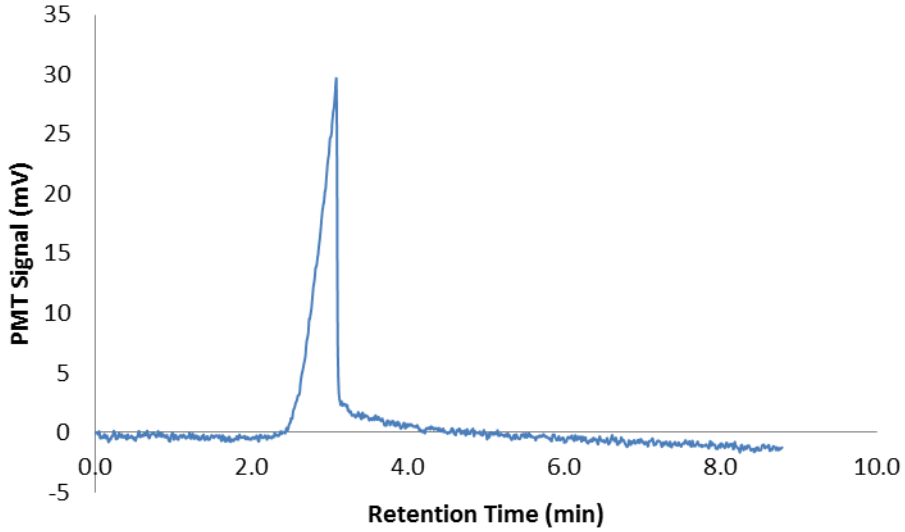


Figure 22: Calibration model of Acetic Acid using the headspace SPME GC-PFPD method.

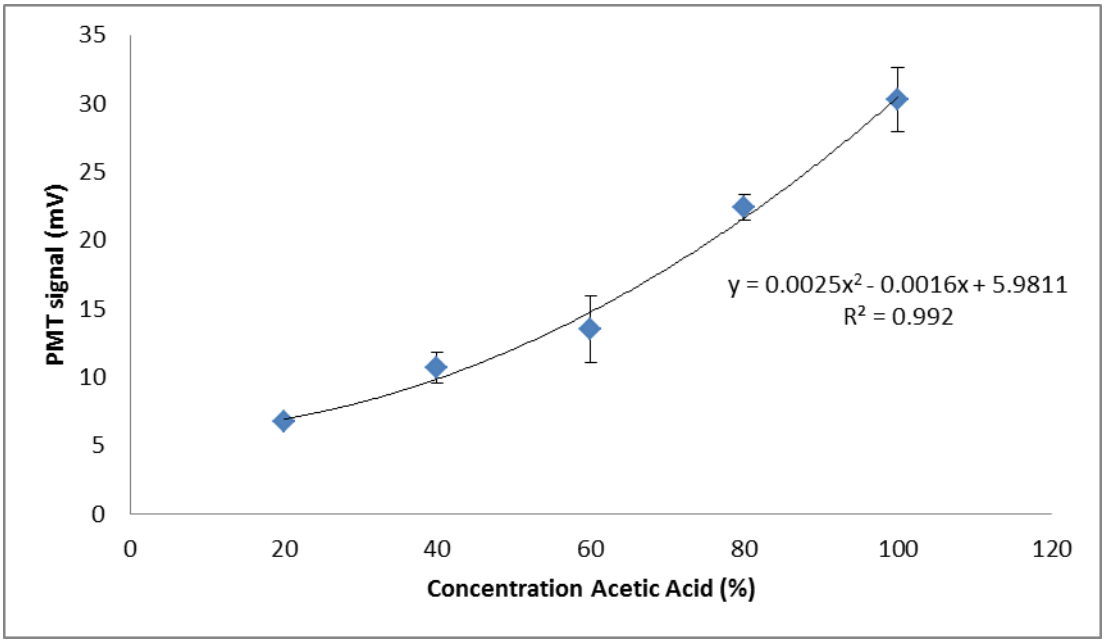


Figure 23: Calibration model of Acetic Acid in 25% (w/v) NaCl using the headspace SPME GC-PFPD method.

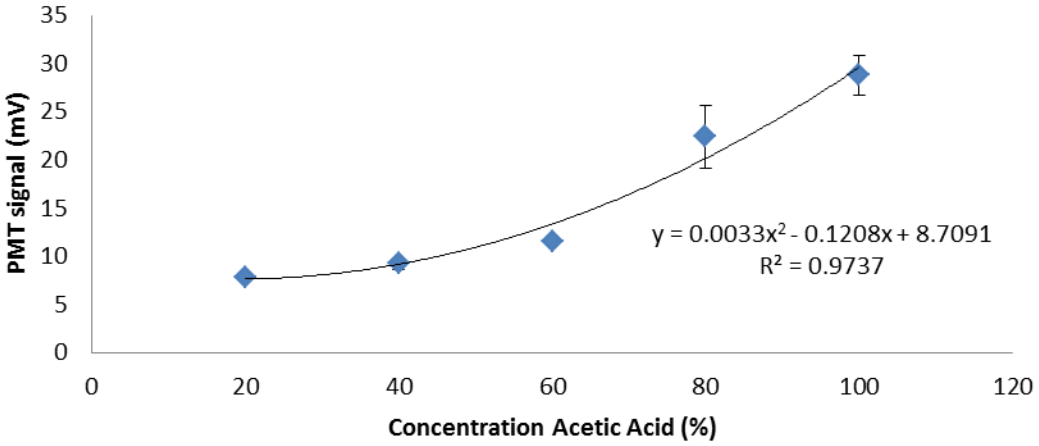


Figure 24: A linear relationship between the number of carbon atoms in the carboxylic acid and the GC retention time was observed.

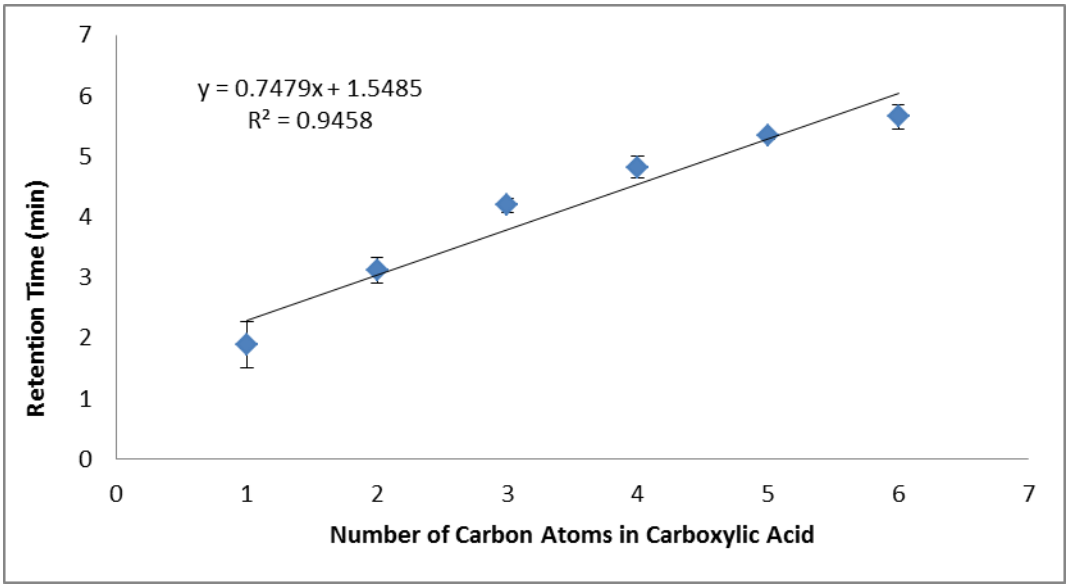


Figure 25: Relationship between retention time of mono carboxylic acid and their boiling points.

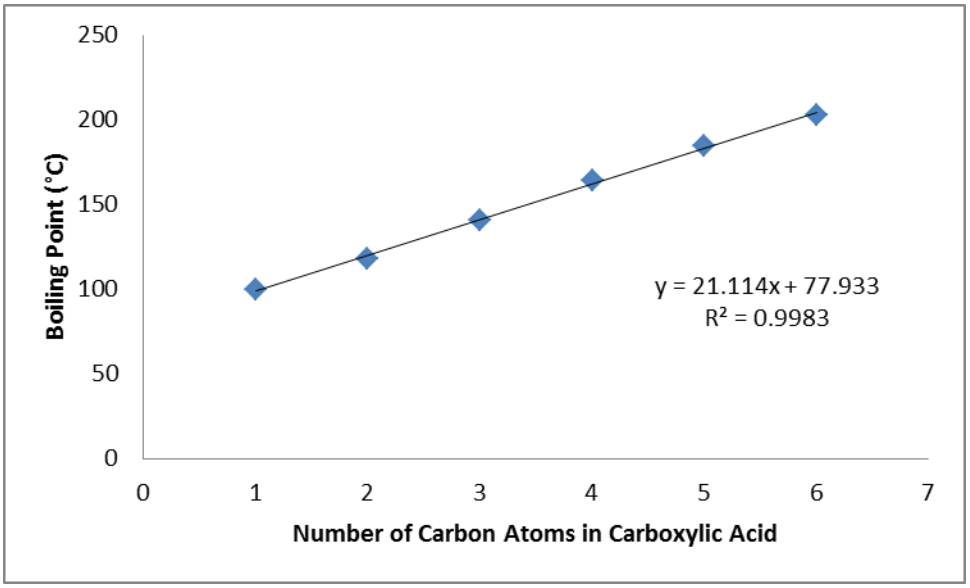


Figure 26: Schematic diagram of the ageing chamber incorporated into the GC oven.

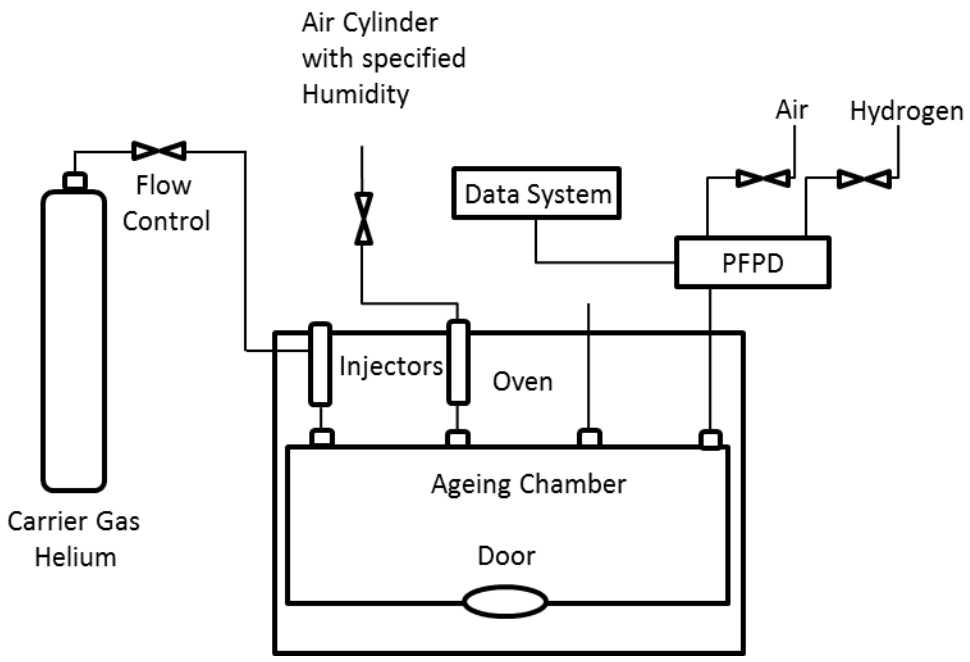


Figure 27: FT-Raman spectrum of "pristine" dried Linseed Oil (n = 2).

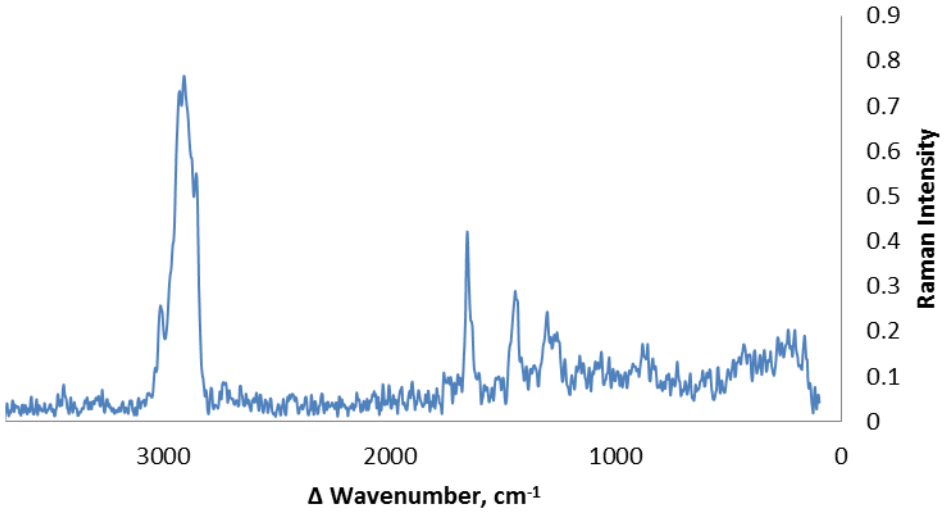


Figure 28: FT-Raman spectrum of authentically aged Linseed Oil (n = 2).

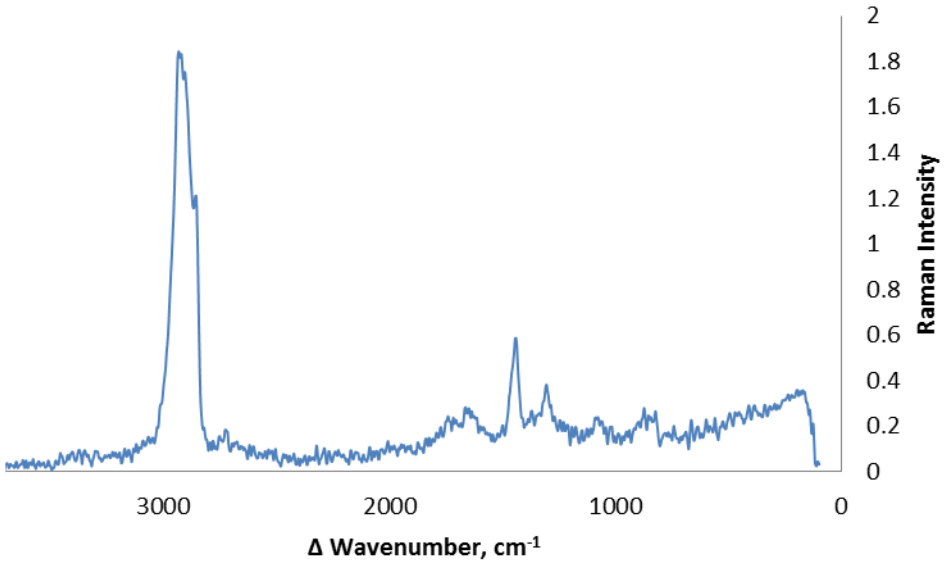


Figure 29: FT-Raman spectra of "pristine" and authentically aged Linseed Oil (n = 2).

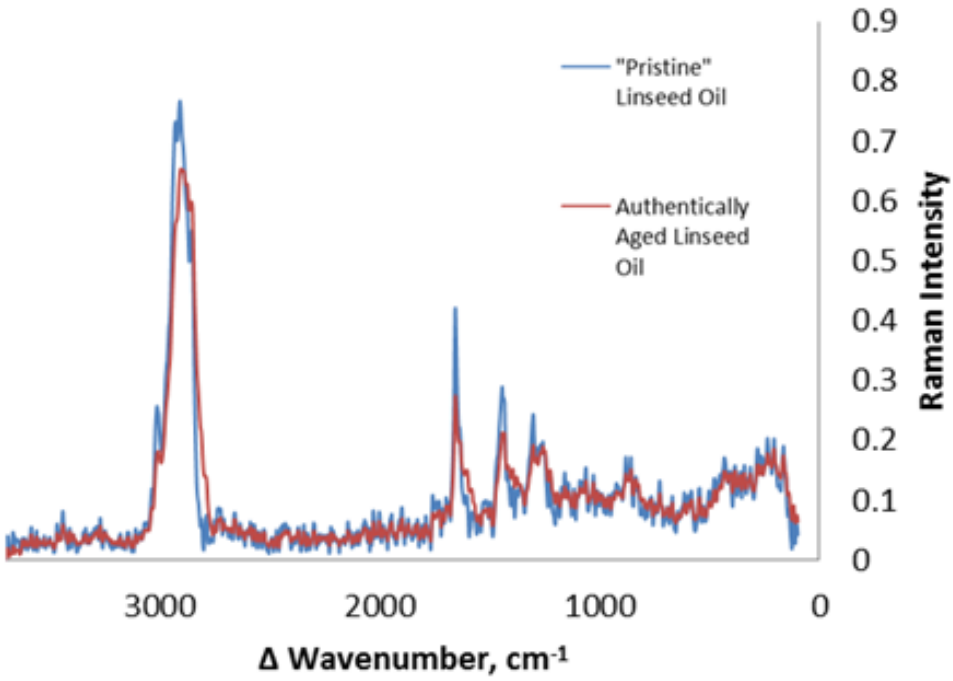


Figure 30: FT-Raman spectra of authentically aged Linseed Oil subtracted from "pristine" Linseed Oil (n = 2).

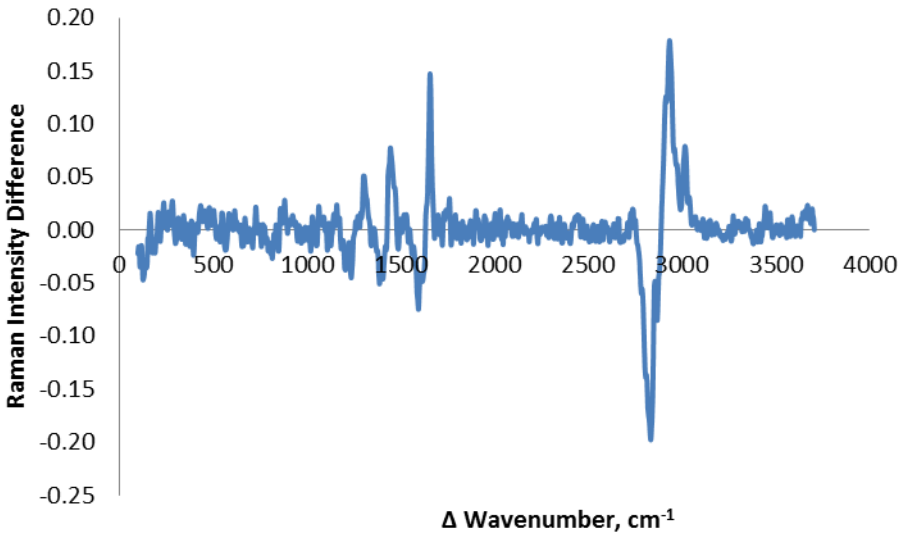


Figure 31: FT-Raman spectrum of artificially aged (24 hr at 100°C and 24 hr UV light) Linseed Oil (n = 3).

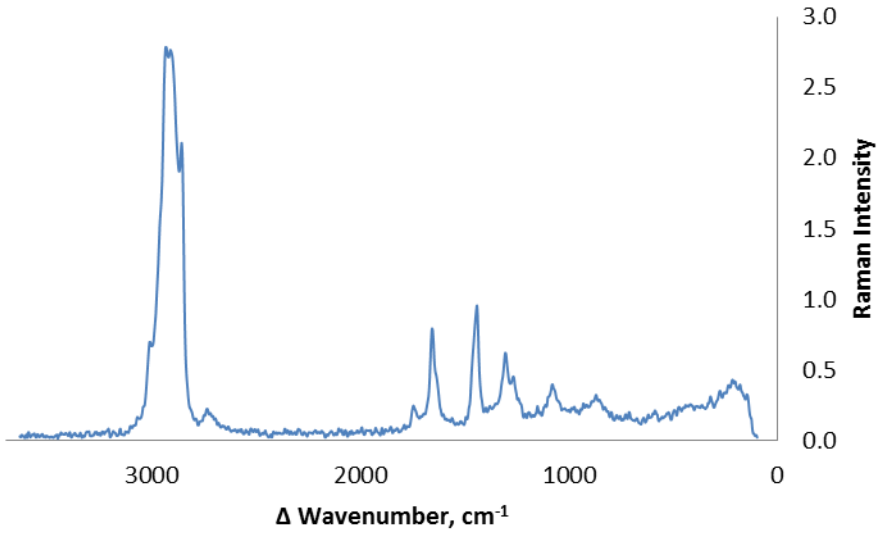


Figure 32: FT-Raman spectra of artificially aged (24 hr at 100°C and 24 hr UV light) Linseed Oil and "pristine" Linseed Oil (n = 2).

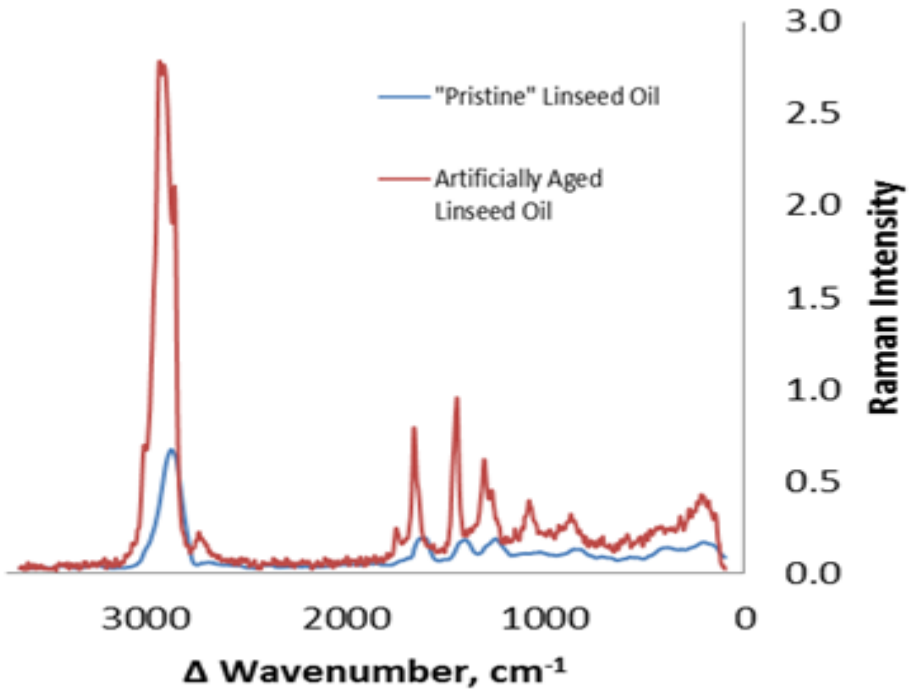


Figure 33: FT-Raman spectra of artificially aged (24 hr at 100°C and 24 hr UV light) Linseed Oil subtracted from "pristine" Linseed Oil (n = 2).

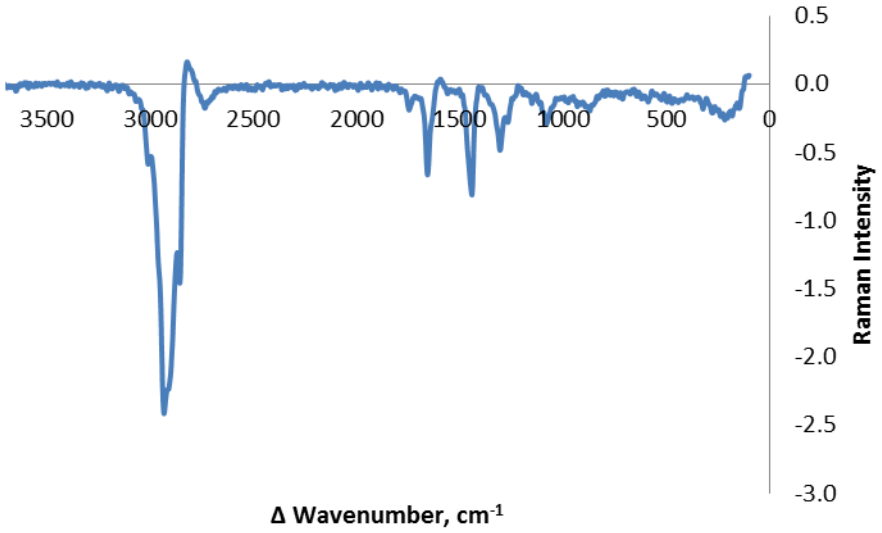


Figure 34: FT-Raman spectra of artificially aged Linseed Oil (at 100°C).

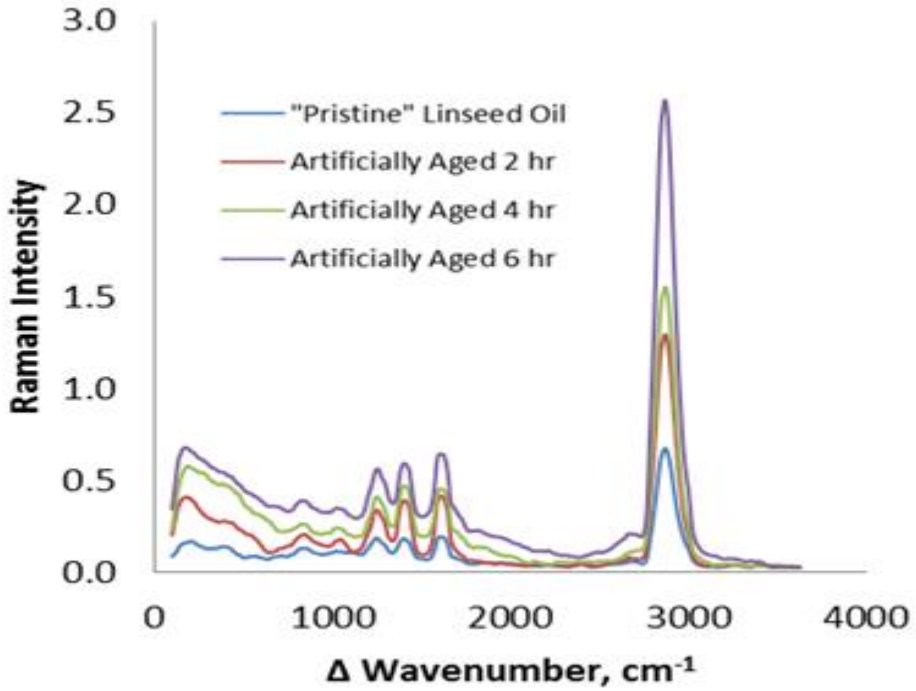


Figure 35: FT-Raman spectra of artificially aged (at 100°C) Linseed Oil subtracted from "pristine" Linseed Oil (n = 2).

References

(1) Beerman, C.; Jelinek, J.; Reinecker, T.; Hauenschild, A.; Boehm, G.; Klor, H.-U., Short term effects of dietary medium-chain fatty acids and n-3 long-chain polyunsaturated fatty acids on the fat metabolism of healthy volunteers, *Lipids in Health and Disease*, 2003, *2*.

(2) Shimanouchi, T., Table of molecular vibrational freqencies consolidated *National Stanard Reference Data Systems,* 1972, *1*, 1-164.

(3) Tomlinson, G. E.; Curnette, B.; Hathaway, C. E., Temperature dependence of the Raman and infrared spectrum of liquid formic acid, *Journal of Molecular Spectroscopy*, 1970, *36*, 26-33.

(4) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.,Infrared Spectroscopy In *Spectrometric Identification of Organic Compounds*; John Wiley & Sons, INC.: Hoboken, N.J., 2005.

(5) Bellamy, L. J.,Carboxylic Acids In *The Infra-red Spectra of Complex Molecules*; Methuen & Co LTD: London, 1958, pp 161-177.

(6) Sherban-Kline. *Infrared Spectroscopy: A Key to Organic Structure*; Yale-New Haven Teachers Institute, 2013.

(7) McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons, Inc.: Canada, 2000; Vol. 157.

(8) Daher, C.; Paris, C.; Le Ho, A.-S.; Bellot-Gurlet, L.; Echard, J.-P., A joint use of Raman and infrared spectroscopies for the identification of natural organic media used in ancient varnishes, *Journal of Raman Spectroscopy*, 2010, *41*, 1494-1499.

(9) Mills, J. S.; White, R. *The Organic Chemistry of Museum Objects*, 2nd ed.; Butterworth Heinemann: London, 1994, p 206.

(10) *Linseed*; Springdale Crop Synergies Ltd, , 2002, p 3.

(11) *Minor oil crops - individual monographs*; Food and Agriculture Organization of the United Nations.

(12) Bonaduce, I.; Odlyha, M.; Di Girolamo, F.; Lopez-Aparicio, S.; Grontoft, T.; Perla Colombini, M., The role of organic and inorganic indoor pollutants in museum environments in the degredation of dammar varnish, *Analyst*, 2013, *138*, 487-500.

(13) Mills, J. S.; White, R., Natural resins of art and archaeology: Their sources, chemistry, and identification, *Studies in Conservation*, 1977, *22*, 12-31.

(14) Lahanier, C.; Preusser, F. D.; Van Zelst, L., Study of conservation of museum objects: Use of classical analytical Techniques, *Nuclear Instruments and Methods in Physics Research*, 1986, *B14*, 1-9.

(15) Baker, K., Singular Value Decomposition; 2005.

(16) Godoi, A. F. L.; Van Vaeck, L.; Van Grieken, R., Use of solid-phase microextraction for the detection of acetic acid by ion-trap gas chromatography-mass spectrometry and application to indoor levels in museums, *Journal of Chromatography A*, 2005, *1067*, 331-336.

(17) Arthur, C. L.; Pawliszyn, J., Solid phase microextraction with thermal desorption using fused silica optical fibers, *Analytical Chemistry*, 1990, *62*, 2145-2148.

(18) Zhang, Z.; Pawliszyn, J., Headspace solid-phase microextraction, *Analytical Chemistry*, 1993, *65*, 1843-1852.

(19) *Solid Phase Microextraction Troubleshooting Guide*; Supelco: Bellefonte, PA, 2004.

(20) *Material Safety Data Sheet*; Sigma-Aldrich Co. LLC.

(21) Leona, M.; Van Duyne, R.; Berrie, B.; Casadio, F.; Ernst, R. R.; Faber, K. T.; Sgamellotti, A.; Trentelman, K.; Whitmore, P., Chemistry and materials research at the interface between science and art; report of a workshop cosponsored by the National Science Foundation and the Andrew W. Mellon Foundation, 2009.

(22) Echard, J. P.; Benoit, C.; Peris-Vicente, J.; Malecki, V.; Gimeno-Adelantado, J. V.; Vaiedelich, S., Gas chromatography/mass spectrometry characterization of historical varnishes of ancient Italian lutes and violin, *Analytica Chimica Acta*, 2007, *584*, 172-180.

(23) Caruso, F.; Saverwyns, S.; Van Bos, a.; Chillura Martino, D. F.; Ceulemans, A.-E.; de Valck, J.; Caponetti, E., Micro-X-Ray Fluorescence and the old masters: Non-destructive in situ characterisation of the varnish of historical low countries stringed musical instruments, *Applied Physics A*, 2012, *107*.

(24) Echard, J.-P.; Bertrand, L., Complementary spectroscopic analyses of varnishes of historical musical instruments, *SpectroscopyEurope*, 2010, *22*, 12-15.

(25) Bernuy, B.; Meurens, M.; Mignolet, E.; Turu, C.; Larondelle, Y., Determination by fourier transform Raman Spectroscopy of conjugated Linoleic Acid in I_2-

photoisomerized Soybean Oil, *Journal of Agricultural and Food Chemistry*, 2009, *57*, 6524-6527.

Chapter 4: Conclusions and Future Work

4.1 Raman Spectroscopy

4.1.1 Raman Database

The goal of the first phase of the study was to determine if constructing a Raman Spectral Database of carboxylic acids would give more insight into the composition of complex organic molecules used for varnishes. Raman spectroscopy proved to be a useful technique to collect high resolution spectra with great detail. Nineteen different standard spectra were collected to comprise the database of mono-carboxylic acids, di-carboxylic acids, and medium- and long- chain fatty acids. Along with the nineteen standards, Raman spectra of four different varnishes were collected. These four varnishes were split into two different classes, the triglycerides and the triterpenoids. The triglyceride varnishes were Linseed Oil and Tung Oil. The triterpenoids were Dammar and Mastic resins. Each spectrum was analyzed for the major frequency shifts and assigned corresponding molecular vibrational stretches. Through both visual observations and comparing the frequency shifts of the carboxylic acid standards, certain patterns emerged. Also using a visual comparison and comparing frequency shifts between the carboxylic acid standards with the unknowns, it was obvious that some of the standards comprised the unknowns more than the others. However, it was difficult to tell exactly which standards were combined and at what quantity to make up the unknowns. For the database to be more useful in determining which standards and how

much of each standard make up the unknown, the mathematical procedure of Singular Value Decomposition was applied.

4.1.2 Singular Value Decomposition

Singular Value Decomposition (SVD) was applied to the 19 standards that comprised the Raman Spectral Database. To optimize the SVD approach, seven singular values were chosen. Singular values are representations of reduced dimensionality of the original data; for this study they represent a feature from the original Raman Database. The seven singular values were then used to model four unknowns: Linseed Oil, Tung Oil, Dammar, and Mastic. The root-mean square (RMS) error was then used as a measure of the model predicted using the seven singular values to that of the unknown spectra. The RMS errors for the unknowns were 0.08, 0.13, 0.21, and 0.21 Raman Intensity units for Linseed Oil, Tung Oil, Dammar, and Mastic, respectively. If those values are compared to the largest peak in the unknown spectra, the % relative RMS errors are 1.7%, 1.7%, 4.9%, and 6.4%. These RMS values are relatively low, thus the seven singular values representing the Raman Spectral Database match well to spectra that were observed for the unknown mixtures.

4.1.3 Future Work for the Raman Database

While the construction of the Raman Database of carboxylic acids is comprehensive there are many additions that could be added to it, such as more medium

and long chain fatty acids. The Raman spectra of the carboxylic acids in the database were collected in the neat phase. By collecting dilute carboxylic acids samples, it will broaden the application of the database and the sample collection. The two mixtures that were collected were Linseed Oil and Tung Oil. If more mixtures and diluted samples were collected the database would be more robust. To transition the database for use with aged varnishes, the sample size will be greatly reduced so it is important to determine the detection limit of the instrument for samples in a dilute state. Also mixtures of the carboxylic acid standards should be collected. Varnishes are mixtures of various components so to get a better understanding of the varnishes, mixtures of the samples of carboxylic acids need to be collected as well.

Once the standard reference materials are created with the ageing chamber they should also be incorporated into the Raman database. The database along with SVD can be used to determine the fragments that are created from the degradation of the ageing varnishes. The database can be used to identify the fragments and SVD can be applied for quantitative analysis.

Finally to make the database more robust, it can be used in conjunction with other applications in analytical chemistry. Carboxylic acids are present in samples such as humic and fulvic acids found in soils and waters. The database can be applied to these samples to determine their carboxylic acid content. Another avenue would be to complex a metal to a representative carboxylic acid and study how the Raman spectrum changes. This could be used study the speciation and mobility of metals in the environment as well as the composition of humic and fulvic acids.

4.2 Gas Chromatography (GC)

4.2.1 GC Characterization

To characterize the products made during the ageing process, a method was developed to identify carboxylic acids. The absorption time, desorption time, the split ratio, and the oven temperature program were optimized. Along with the optimization of these factors, experiments were conducted to determine the selectivity for carboxylic acid. The number of carbons in the carboxylic acid correlated to the retention time. This relationship will be useful when the degradation products (e.g., fragments of fatty acids containing carboxylate functionalities) from the ageing varnishes are measured. In addition to the qualitative utility, a quantitative relationship was determined between the concentration of Acetic Acid and the PMT signal. Therefore when a fragment is detected, both the type of carboxylic acid and the amount of that carboxylic acid can be determined. This optimization, as well as the relationships discovered, will lead to smooth transition for the incorporation of the ageing chamber.

4.2.2. Future Work for the GC Characterization

The quantitative relationship was determined for just a single mono-carboxylic acid. For this to become a more robust method, the relationship needs to be extended to all of the carboxylic acids that were studied. For the carboxylic acids that are not volatile, a methylation process could be developed to convert the fatty acids into their

methyl esters. Because many of the varnishes are not volatile, the methylation will expand the usefulness of the method.

4.3 Ageing Chamber

4.3.1. Ageing Chamber

An ageing chamber was designed, fabricated and tested. The ageing chamber was designed to create standard reference materials by artificially ageing varnishes in a controlled environment. The ageing chamber was designed and fabricated for incorporation directly into the gas chromatograph. This way the heat, use of a UV light, humidity and pollutants can be controlled so the factors that contribute to ageing are known in the exact amounts.

4.3.2. Future work for the Ageing Chamber

The main future work for the ageing chamber would be the creation of SRMs for major varnishes (e.g., Dammar and Mastic). The ageing chamber has been demonstrated as "proof of concept" by incorporation into the gas chromatograph and showing that structural changes can be precisely monitored. The ageing chamber approach can now be studied in detail by simultaneous variation of the key factors that contribute to ageing. By controlling the factors that affect ageing, attempts can be made to create aged SRMs that closely resemble varnishes that have been aged naturally.

Appendix A

Figure 1: FT-Raman spectrum of Formic Acid (n = 3).

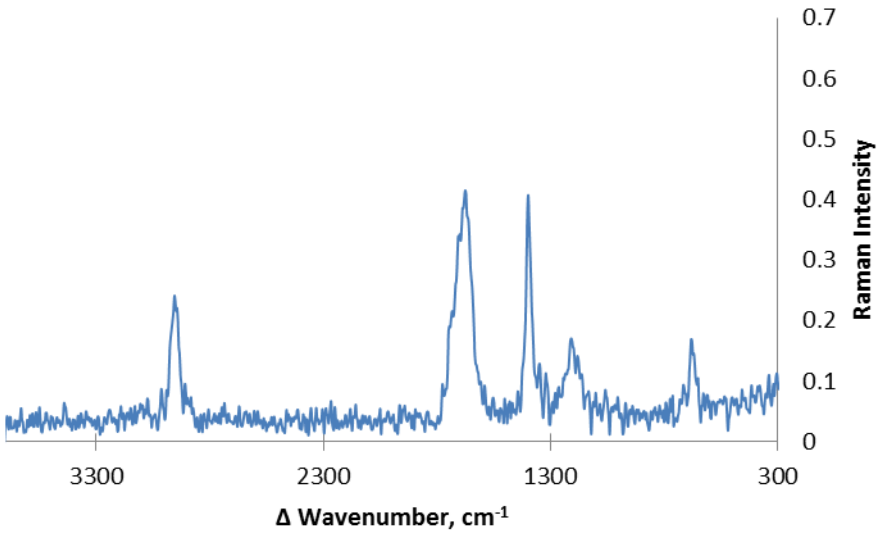


Figure 2: FT-Raman spectrum of Acetic Acid (n = 3).

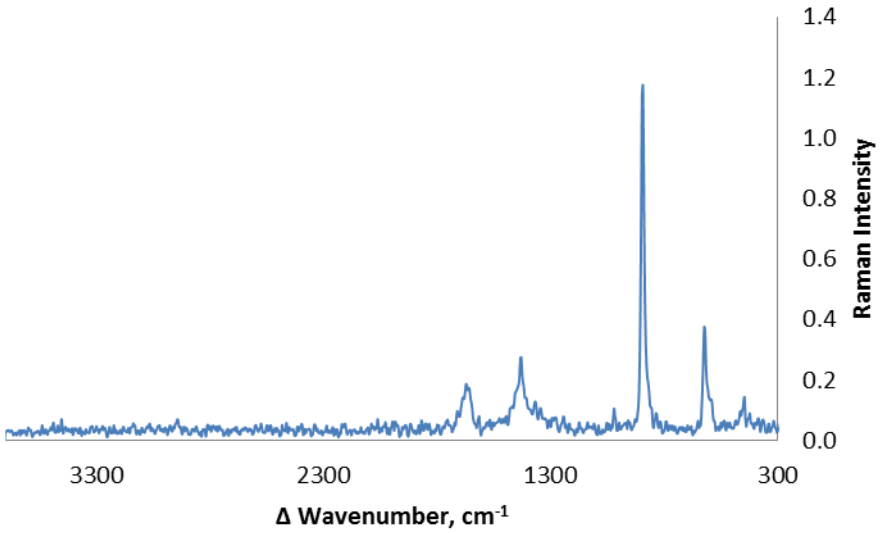


Figure 3: FT-Raman spectrum of Propanoic Acid (n = 3).

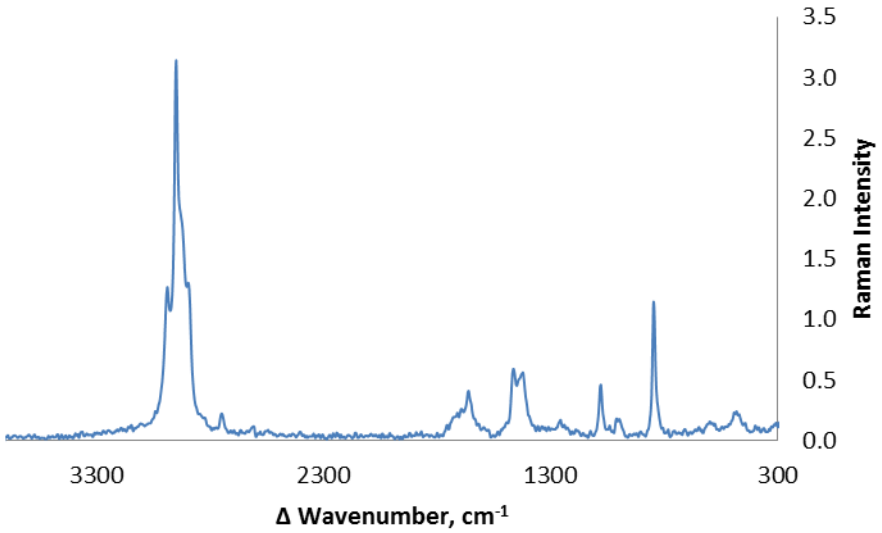


Figure 4: FT-Raman spectrum of Butyric Acid (n = 3).

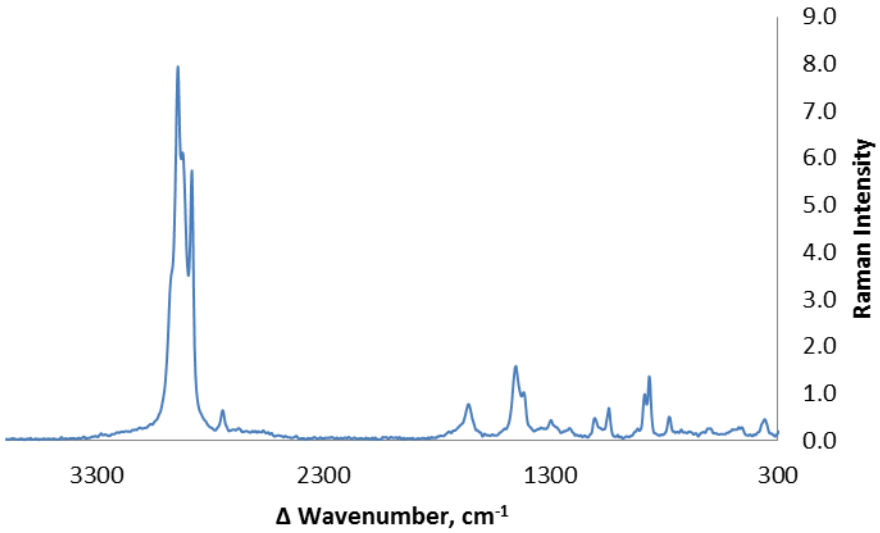


Figure 5: FT-Raman spectrum of Valeric Acid (n = 3).

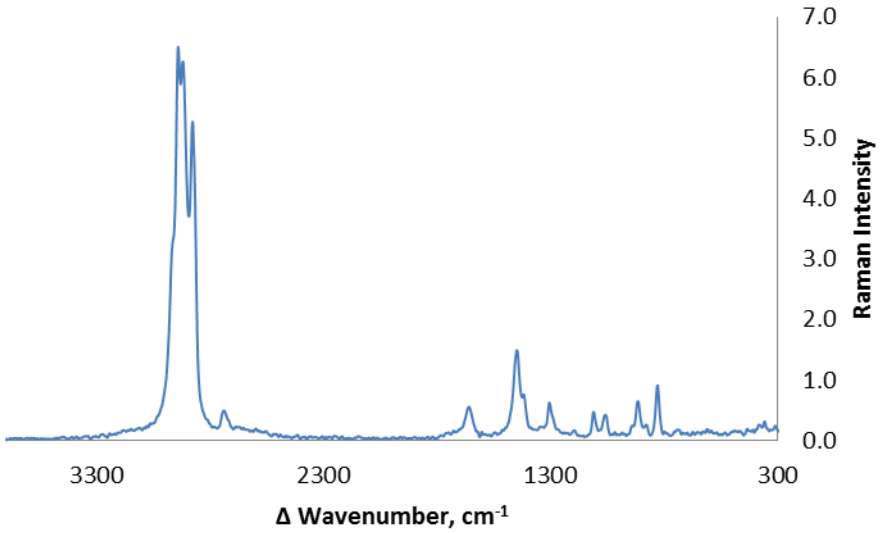


Figure 6: FT-Raman spectrum of Caproic Acid (n = 3).

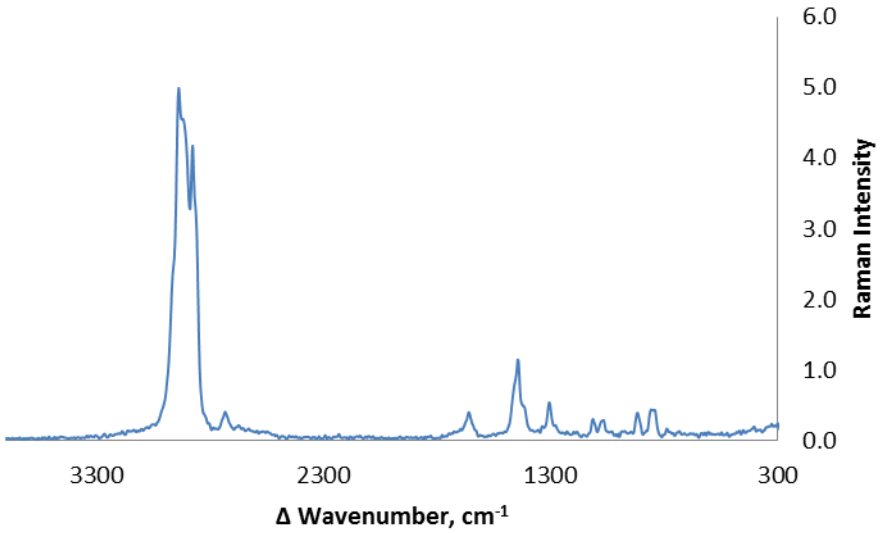


Figure 7: FT-Raman spectrum of Enanthic Acid (n = 3).

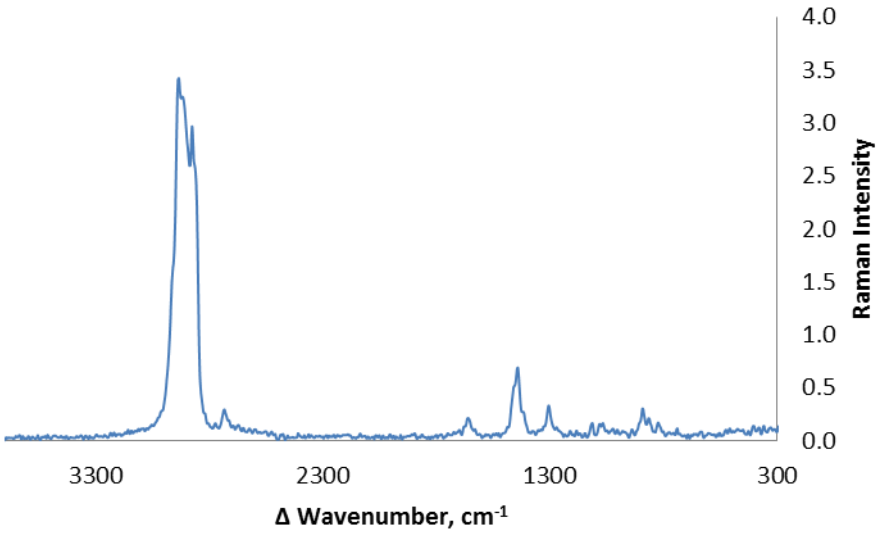


Figure 8: FT-Raman spectrum of Caprylic Acid (n = 3).

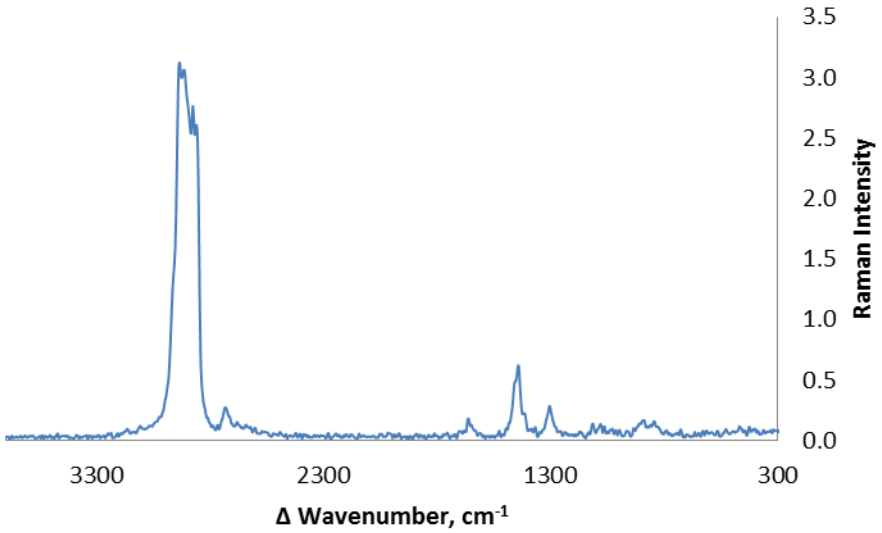


Figure 9: FT-Raman spectrum of Pelargonic Acid (n = 3).

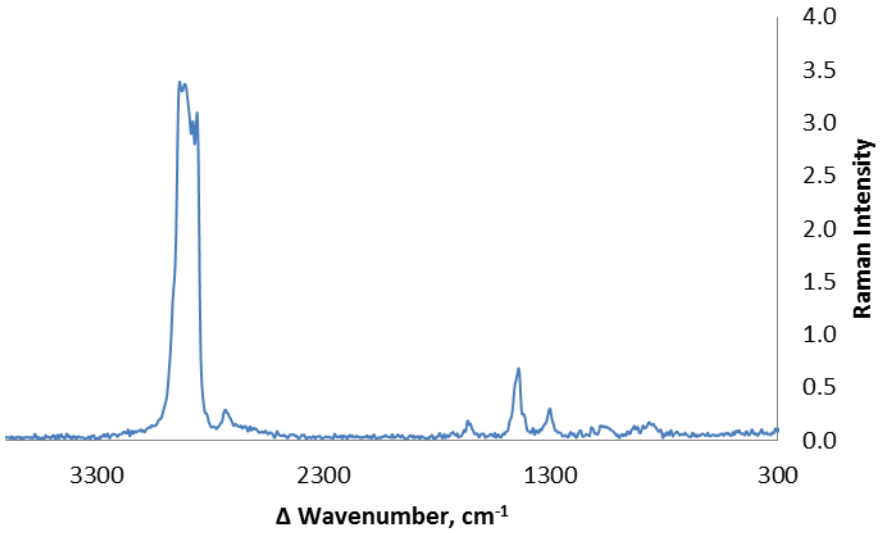


Figure 10: FT-Raman spectrum of Oxalic Acid (n = 3).

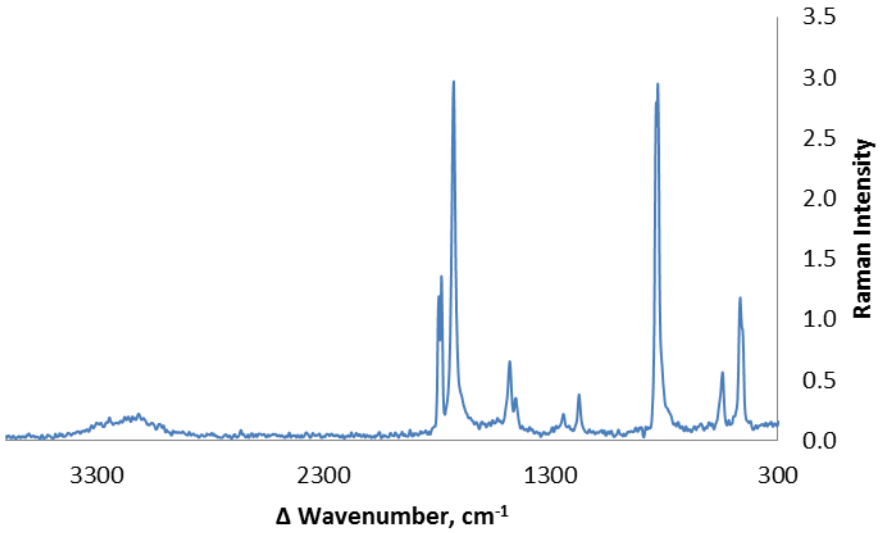


Figure 11: FT-Raman spectrum of Malonic Acid (n = 3).

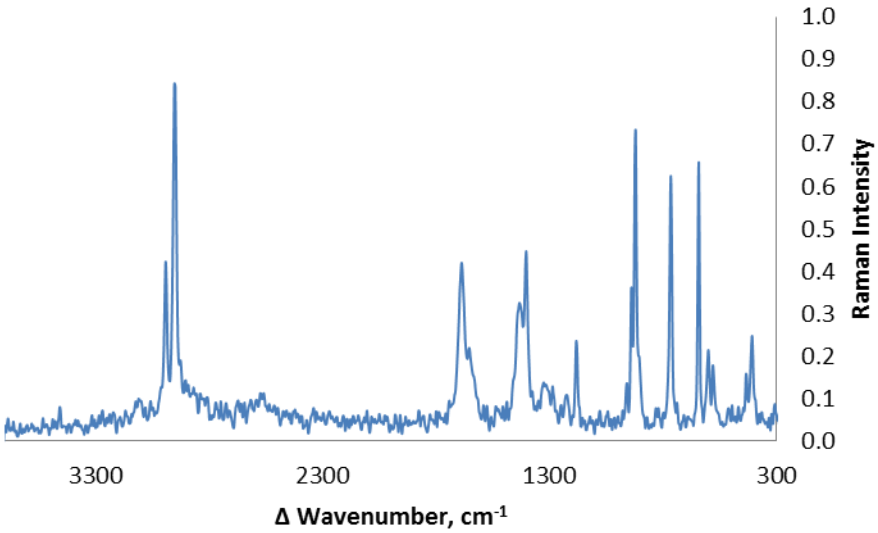


Figure 12: FT-Raman spectrum of Succinic Acid (n = 3).

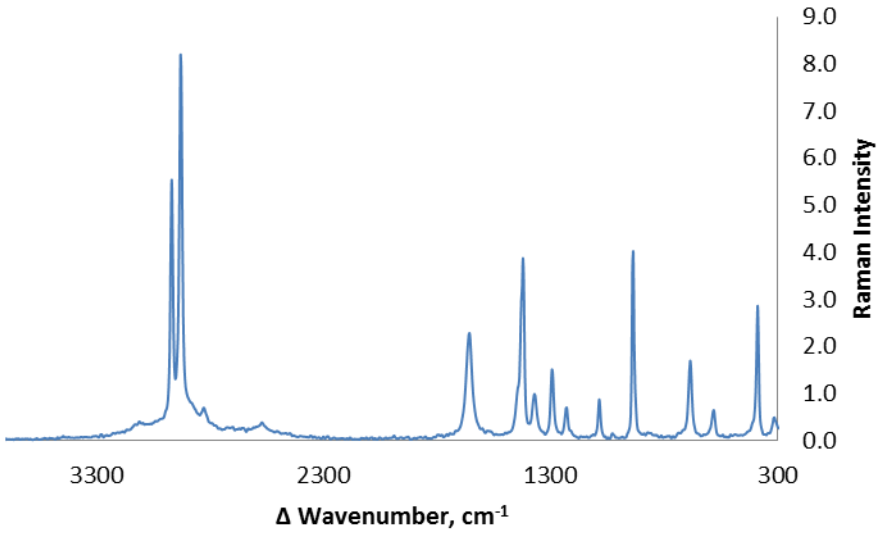


Figure 13: FT-Raman spectrum of Glutaric Acid (n = 3).

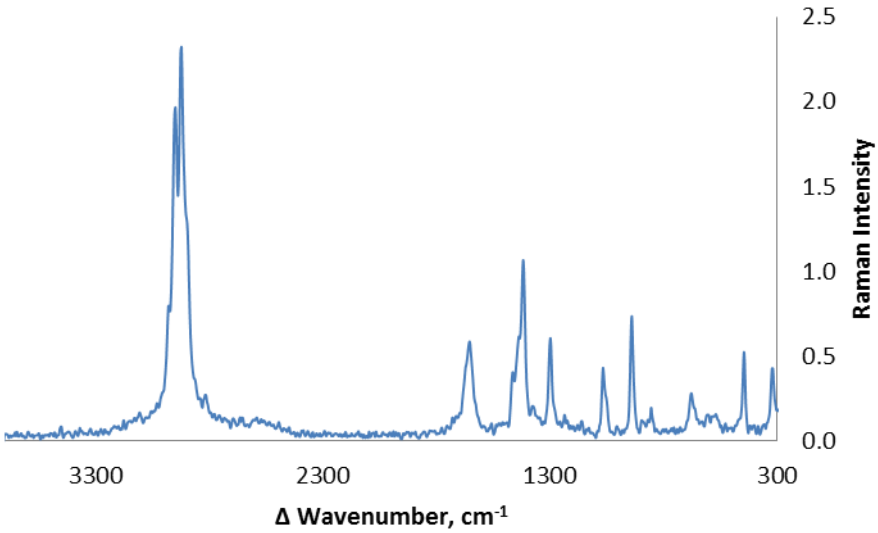


Figure 14: FT-Raman spectrum of Oleic Acid (n = 3).

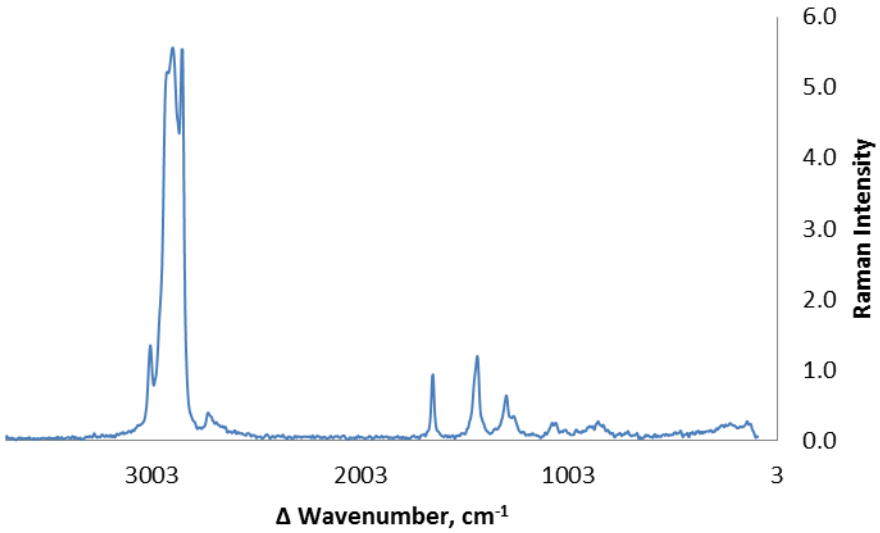


Figure 15: FT-Raman spectrum of Linoleic Acid (n = 3).

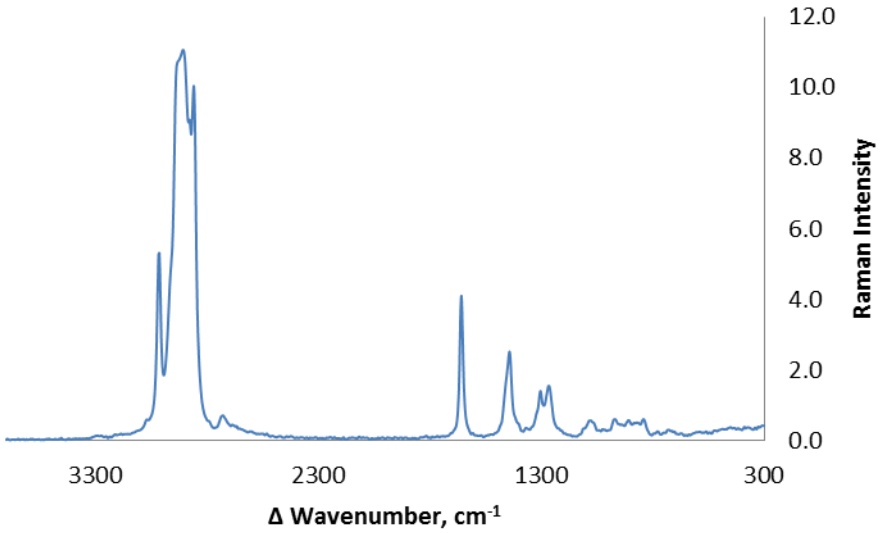


Figure 16: FT-Raman spectrum of Linolenic Acid (n = 3).

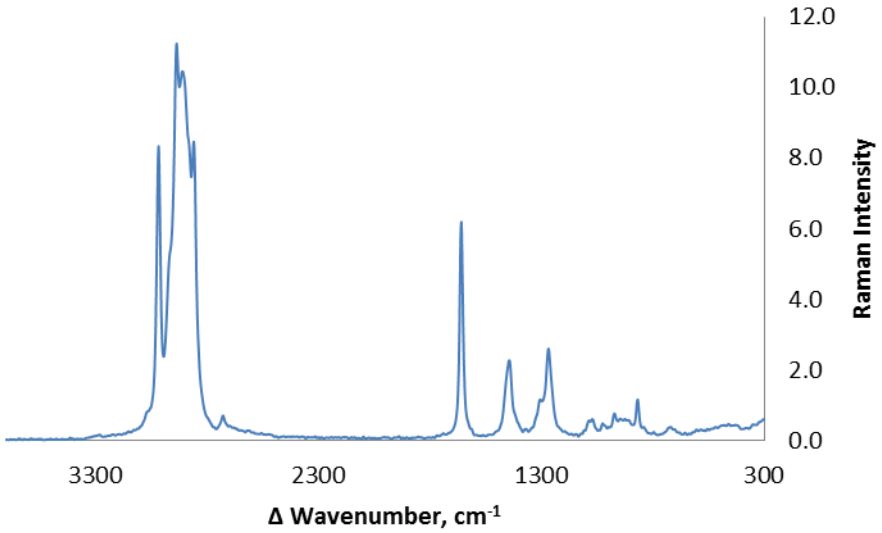


Figure 17: FT-Raman spectrum of Lauric Acid (n = 3).

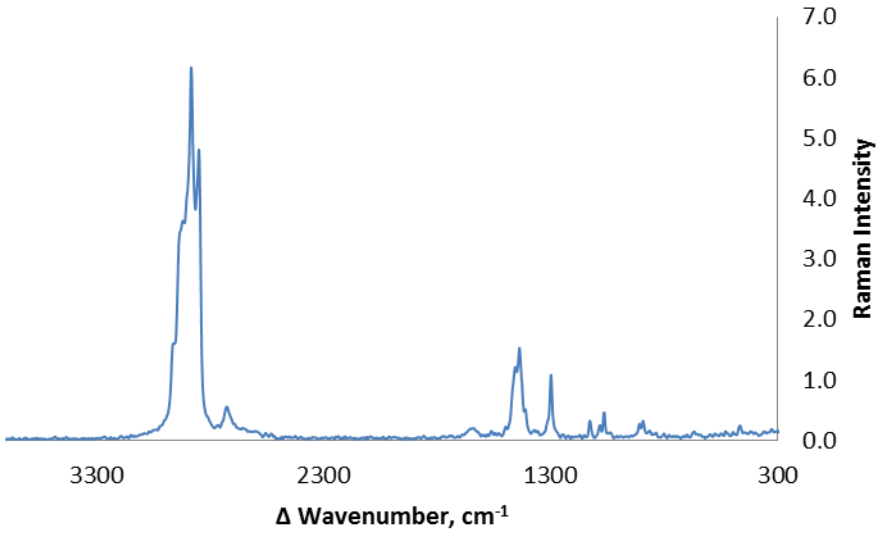


Figure 18: FT-Raman spectrum of Palmitic Acid (n = 3).

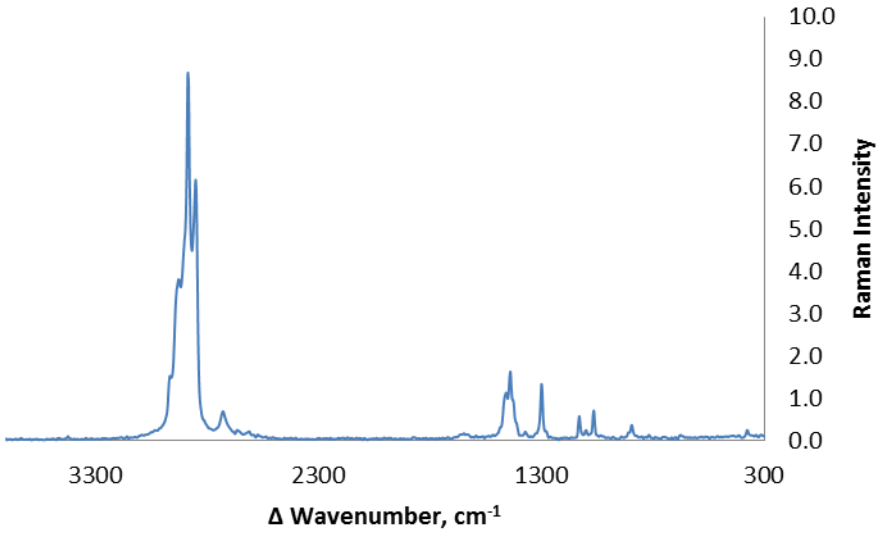
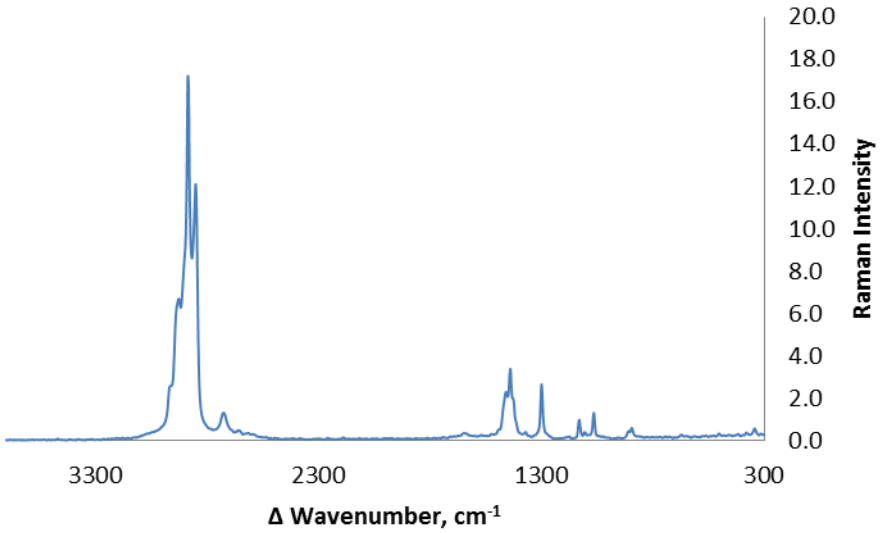


Figure 19: FT-Raman spectrum of Stearic Acid (n = 3).



PART 2: DIFFERENTIAL ITEM FUNCTIONING ON MULTIPLE-CHOICE GENERAL CHEMISTRY ASSESSMENTS.

Chapter 1: Introduction

1.1.1 Assessments

Assessments are given almost every day to students. They can be as simple as a homework assignment, a quiz, an hourly examination (an examination that is given during the course usually with a time limit of one hour), or a standardized examination. Homework may be given to allow students to have a practical application or simply practice of the concepts and ideas that were learned. Depending on the instructor's grading scheme, homework assignments can be lower-stakes assessments, as they don't carry as much weight in the student's final grade. Whereas, hourly examinations or final examinations are usually higher-stakes assessments where they are weighted more in the student's final grade. For high-stakes assessments, the points received can lead to many decisions, such as the score received on the assessment, the grade in the course, or even contribute the student's success in their program. When considering other high-stakes testing such as the American College Testing (ACT) or the SAT (formally known as the Scholastic Aptitude Test), the judgment could influence whether or not a student is granted admission into a particular undergraduate institution or eligible for merit-based scholarships. The Graduate Record Examination (GRE) is often used as a determining factor for accepting students into graduate school, therefore affecting their career. Because high-stakes decisions are based on assessments, it is necessary that these

assessments are as fair and unbiased as possible. Many organizations that provide standardized assessments perform statistical testing to insure their assessments are as fair and unbiased as possible.[1-5] The American Chemical Society's Division of Chemical Education, Examination Institute (ACS-EI) performs many statistical tests to ensure their assessments are as valid and unbiased as possible. Such item-level statistics include item difficulty, item discrimination, and Differential Item Functioning (DIF). DIF is when two subgroups who are match on equal ability perform statistically different on an item on an assessment.[6] This statistical procedure analyzes each item on an assessment for a bias according to the subgrouping. The subgroups can be any of choice such as gender, ethnicity, age, etc. The ACS-EI performs gender DIF on all of its trial assessments with sufficiently large datasets in order to assist in judging which items will be included on a final examination.[1]

1.1.2 Gender Studies

Gender bias can certainly be a controversial issue and differences in gender on overall assessments has been exhaustively studied, but there are still many discrepancies to be accounted for.[4-5, 7] Gender bias is an issue not only in the area of assessment but also in difference in degrees being awarded,[8] the awarding of National Science Foundation (NSF) grants[9] and the academic environment[10] to name a few. While many studies focus on bias on overall assessments, there has been a rise studies examining differential item functioning based on the articles that have been published in the last decade or two. While many of these studies are focused on assessments for high school

students because of the many standardized examination the students take,[11-14] few have examined college assessments and specifically in the area of chemistry.[1, 3]

1.1.3 Phases of Study

The interest of this study was to examine gender DIF on general chemistry I assessments. The study was conducted in four stages. The first stage was to find test items that exhibited DIF. This was conducted by doing a two-stage DIF analysis[15] using the Mantel-Haenszel statistic on two trial tests from ACS-EI. Once items that exhibited gender DIF were identified, the second part of the study was conducted that included cloning these item and retesting them on high-stakes hourly examinations for a general chemistry I course. To determine if the DIF is real, a persistence study was conducted using different external relevant measures of proficiency to determine if the DIF would be persistent by using different matching criteria. While DIF is a real statistical phenomenon, this study was focusing on the DIF being persistent. If the DIF was persistent then it would be considered it to be "real" DIF. The third stage consisted of narrowing some of the possible causes of DIF. Literature has shown that certain formats of items or content areas tend to favor one gender verses another.[4-5, 11-14] Therefore to examine the causes of DIF, the items that exhibited persistent DIF were cloned by their content and/or format. Finally to determine the possible reasons of why DIF is occurring, semi-structured interviews using an eye-tracking instrument were conducted to gain information about the student's problem-solving process.

Together these four stages of the study helped to gain more information about the phenomenon of DIF in an undergraduate college chemistry course. There are very few studies of DIF persistence, as commonly only a one-stage DIF analysis is conducted because of the limited availability of other measures to use as external criteria. By using a more rigorous measure of real DIF items and a larger sample set, the results presented in this study are a better indicator of items that exhibited persistent gender DIF on general chemistry I assessments, along with the possible causes and reasons why it is occurring.

References

1. Kendhammer, L.; Holme, T.; Murphy, K., Identifying Differential Performance in general Chemistry: Differential Item Functioning Analysis of ACS General Chemistry Trial Tests. *Journal of Chemical Education* **2013,** *90* (7), 846-853.

2. Holme, T., Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education* **2003,** *80* (6), 594.

3. Schroeder, J.; Murphy, K. L.; Holme, T. A., Investigating Factors That Influence Item Performance on ACS Exams. *Journal of Chemical Education* **2012,** *89*, 346-350.

4. Cole, N. S. *The ETS Gender Study: How Females and Males Perform in Eductional Settings*; Educational Testing Service: Princeton, NJ, 1997; pp 1-36.

5. Beller, M.; Gafni, N., The 1991 International Assessment of Education Progress in Mathematics and Sciences: The Gender Differences Perspective. *Journal of Educational Psychology* **1996,** *88* (2), 365-377.

6. Holland, P. W.; Wainer, H., *Differential Item Functioning*. Lawrence Erlbaum Associates: Hillsdale, New Jersey, 1993.

7. Maccoby, E. E.; Jacklin, C. N., *The Psychology of Sex Differences*. Stanford University Press: Stanford, California, 1974.

8. S&E Degrees: 1966-2010: National Center for Science and Engineering Statistics. http://www.nsf.gov/statistics/nsf11316/content.cfm?pub_id=4062&id=2 (accessed May 26).

9. Lepkowski, W. I. L., No Gender Bias in its Grant Awards, says NSF. *Chemical & Engineering News Archive* **1997,** *75* (36), 10.

10. Greene, J.; Stockard, J.; Lewis, P.; Richmond, G., Is the Academic Climate Chilly? The Views of Women Academic Chemists. *Journal of Chemical Education* **2010,** *87* (4), 381-385.

11. Zenisky, A. L.; Hambleton, R. K.; Robin, F. *DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices*; University of Massachusetts Amherst: Amherst, MA, 2003b; pp 1-22.

12. Hamilton, L. S.; Snow, R. E. *Exploring Differential Item Functioning on Science Achievement Tests*; 483; National Center for Research on Evaluation, Standards, and Student Testing. : Los Angeles, CA, 1998; pp 1-43.

13. Hamilton, L. S., Gender Differences on High School Achievement Tests: Do Format and Content Matter? *Educational Evaluation and Policy Analysis* **1998,** *20* (3), 179-195.

14. Schmitt, A. P.; Dorans, N. J., Differential Item Functioning for Minority Examinees on the SAT. *Journal of Education Measurement* **1990,** *27* (1), 67-81.

15. Zenisky, A. L.; Hambleton, R. K., Detection of Differential Item Functioning in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach. *Educational and Psychological Measurement* **2003a,** *63* (1), 51-64.

Chapter 2: Theory

2.1 Introduction

With the main objectives of this study including identifying DIF, determining the persistence of DIF, studying the causes of DIF, and lastly why DIF happens there are many statistical and theoretical frameworks to support the research. This chapter is going to give a brief overview on the theories that support each of these objectives starting with identifying gender DIF. To gain a better understanding of gender DIF and its persistence, this section will include background on gender studies and DIF studies. Also, some of the main statistical methods for determining DIF will also be introduced. Next looking at the causes of DIF, the theory of common item equating is explored and how it relates to using cloned items. Finally focusing on why DIF happens, it is important to look at student's problem solving process. This will include brief overviews into Information Processing Theory as well as Dual-Processing Theory.

2.2 Gender Studies

2.2.1 Gender Differences

There have been an extensive number of studies on gender differences in the last 30 years. Maccoby and Jacklin were one of the leaders in the field with their book The

Psychology of Sex Differences.[1] They state that when looking at the intellectual development between groups of individuals (generic for all types of subgroup differentiation) one should consider two things. The first thing to consider is the rate of the intellectual development between the subgroups and if there is a difference between each group's progress and the level that they ultimately reach. The second thing to consider is if one subgroup has a greater representation than the other.[1] The second consideration is of great interest because it takes into account the idea of a group of individuals having an unfair advantage in their intellectual abilities.

The Maccoby and Jacklin review on gender differences included over 1,600 studies featuring many different areas. One conclusion they found was that female students tend to possess greater verbal abilities. It concluded that girls around the age of ten began to outperform boys with regard to their verbal abilities and the gap increases with age. Quantitative and spatial abilities seemed to lay the other direction, favoring male students, along with male students performing better in mathematics and science.[1]

According to the most recent study by the National Science Foundation [11-316] there has been an increase in all degrees (bachelors, masters, and doctorate) earned by women between the years of 1966 and 2010. In fact, in 2010, 50.3% and 45.5% were the bachelors and masters degrees awarded to women in the fields of science and engineering.[2] Specifically in chemistry, 49.9% of the bachelor's degrees in this area were earned by women compared to the 18.5% in 1966.[2] There is obviously an increase in the amount of female students taking science courses so it's imperative that they be on the same level as male students compared to what was found in the Maccoby and Jacklin studies.

A reiteration of the Maccoby and Jacklin study was performed in 1997 by the Educational Testing Service. This study included over 400 tests and 1,500 data sets, and looked at multiple areas of gender difference including the content of the items, format of the items, and age differences. It was found that the gender gap in mathematics and science, as well as quantitative and spatial abilities, has closed although there are more high performing male students in the areas of math and science in 12th grade than there were female students. However, the female students still outperform male students on assessments testing language abilities.[3]

However, a study that investigated performance on the math and science portion of the International Assessment of Educational Progress (IAEP) found that among 13 year olds, male students outperformed female students. There were 20 different countries that participated in the study on the mathematics section and the male students outperformed the female students not only on overall test score but on all five different content areas and three different problems that involved using higher order cognitive processes to solve. For science the gender effect was much larger with the male students still outperforming the female students with a large effect size for the questions in earth and space science as well as physical science.[4]

Another review of sex differences on various assessments found that male students outperform female students not only in mathematics, but also in quantitative and spatial abilities with some of the effects sizes being quite large.[5] While these studies give insight into the types of gender differences in overall test content, they don't give information into the types of questions or the format of the items that favor one gender versus another.

2.2.2 Item Analysis

To investigate the gender difference in item performance one can use the statistical procedure of Differential Item Functioning (DIF). DIF occurs when one subgroup statistically outperforms the other subgroup (when they are matched on equal ability) on an item on an assessment. The subgroups can be based on gender, age, ethnicity, religion, socioeconomic status, education level, and many others. Being matched on equal ability, both subgroups should have the same chance of answering the item correctly, but if the item exhibits possible DIF then one of these subgroups will statistically score better. The matching on equal ability is an extremely important trait because otherwise if one matched a low performing student of one subgroup against a high performing student on the other subgroups a differential is already incorporated before testing the item. Because it is difficult to measure a person's ability, instead the student's proficiency, as measure by the score they received on an exam, is used as a proxy for ability. There are many statistical methods to determine DIF include the Mantel-Haenszel statistic,[6] item response theory,[7] and logistic regression[8] among others. There are two types of DIF that can occur; uniform DIF and non-uniform DIF.[9] Item characteristic curves for both uniform and non-uniform DIF can be seen in Figure 1. The type of DIF can be detected by looking at the item's characteristic curve, which is a plot of the probability of the student answering the item correctly versus their ability level. Uniform DIF is when one subgroup consistently outperforms the other, no matter what ability level as seen in Figure 1a. Non-uniform DIF is when one subgroup is outperforming the other at low ability levels, but upon moving to higher ability levels the subgroups switch and now the other subgroup is outperforming the original subgroup as

seen in Figure 1b. In this case the low performing male students have a higher probability of answering the question correctly, but when moving to high performing students the females have a higher probability of answering the question correctly.

2.2.3 Procedures for Detecting DIF

While there are many procedures to detect DIF, as mentioned before, one of the more common methods is the Mantel-Haenszel procedure.[10] The Mantal-Haenszel procedure is essentially an odds ratio of the probability of one subgroup getting the item correct versus the probability of the other subgroup getting the item correct. The equation is shown in Figure 2, which is a representation taken from Differential Item Functioning.[10]

Another common method for detecting DIF is item response theory. Item response theory is an approach to test theory that uses probabilistic models to compute the likelihood that a student will respond to an item based on the students latent trait ability and the characteristics of the item.[9] Latent traits are variables that are not observable. In the case of this study, the latent trait measured is a person's ability. Because ability is not an observed variable a person's z-score (proficiency on an assessment) is used to represent their ability. There are three different models that are commonly used with item response theory; the one parameter model, the two parameter model and the three parameter model. The standard logistic function for item response theory is shown in equation 2-1.

$$P_g(\theta) = \frac{e^x}{1+e^x} \qquad 2\text{-}1$$

For this equation, *x* is a representative symbol that will change depending on the parameter model. An item response model has three parameters (item discrimination, the item difficulty, and a pseudo guessing parameter) that are used to characterize the models. The discrimination of an item (represented as *a*) is a way to determine how well the item discriminates between students with a high ability level and those with a lower ability level.[9] The difficulty of the item as represented as *b* is the item's location.[9] Lastly, the pseudo-guessing parameter (represented as *c*) accounts for students being able to guess correctly on multiple-choice or true-false type questions. With other types of items, such as open response, the probability of answering the question correctly approaches zero at lower latent trait (θ) values.[11] However, if the students are able to guess the probability would instead approach a higher probability in accordance with the number of available options (i.e. 25% for a multiple-choice item with four responses).

The one parameter model accounts for the item difficulty only as a result of the latent trait ability.[12] For this case the *x* in equation 2-1 would be given by equation 2-2, where *D* and *a* are constants.

$$x = Da(\theta - b_g) \qquad 2\text{-}2$$

The second parameter is much like the first except this time the item's discrimination is accounted for as well as shown in equation 2-3.[12]

$$x = Da_g\left(\theta - b_g\right) \qquad 2\text{-}3$$

The third parameter account for the item's difficulty, discrimination and the pseudo-guessing parameter as seen in equation 2-4.[9] For three parameter fit the equation represented by *x* is the same as the two parameter fit, except there is also the addition of a pseudo-guessing factor.

$$P_g\left(\theta\right) = c_g + \frac{\left(1 - c_g\right)e^{Da_g\left(\theta - b_g\right)}}{1 + e^{Da_g\left(\theta - b_g\right)}} \qquad 2\text{-}4$$

Lastly another procedure that is commonly used to detect DIF is logistic regression. Logistic regression is another mathematical model that works by predicting if the students will get the question right based on an observed variable (usually the score they receive on the assessment).[8] The logistic regression model is shown in equation 2-5, where one is the reference group, two is the focal group and TOT is the scaled item score.[13]

$$Y = b_0 + b_1\left(\text{TOT}\right) + b_2\left(\text{gender}\right) + b_3\left(\text{TOT} \times \text{gender}\right) \qquad 2\text{-}5$$

The Mantel-Haenszel procedure is similar to that of logistic regression, except in logistic regression the ability variable is not discrete and an interaction variable is also added. One big difference is that logistic regression can identify the less common type of DIF: non-uniform DIF.[8]

2.2.4 DIF Studies

Differential Item Functioning analysis has been conducted on many different types of assessments and subjects. For instance, Schmitt and Dorans conducted a study using minority subgroups to identify DIF to study racial equality on items on the Scholastic Aptitude Test (SAT).[14] Their subgroups were African-American students, Hispanic students, Asian-American students, and Caucasian students, and they focused primarily on DIF analysis on items in the verbal and mathematics section of the SATs.[14] Another study investigated differential item functioning using the test color to gain information about both the answer order and the item order.[15] There are many different studies using DIF on various topics, but this study specifically concentrated on chemistry and class-wide analysis.

Often times, Differential Item Functioning analysis is conducted for large standardized examinations. In Canada, nine of the provinces conduct some form of large-scale assessments, and to make sure those assessments are as valid and fair as possible, one of the statistical tests conducted is DIF.[16] The American Chemistry Society, Division of Chemical Education, Examinations Institute also conducts gender DIF analyses on the trial examinations for the standardized assessments they produce.[15] Because most of these assessments are a one-time testing situation, the matching criterion for DIF is usually the score the student received on the assessment. If the assessment has any questions that exhibit possible DIF, then the score received on the assessment could be potentially biased as well. Zenisky, Hambleton, and Robin proposed a two-stage iterative process to account for issues of the assessment score having potential bias.[17] This process takes part in two stages, with the first being to use a statistical approach that

determines DIF using the assessment score as the matching criterion. The second stage involves removing the questions that have the highest DIF, changing the assessment scores accordingly, and then performing the DIF analysis again. By using this two-stage process the assessment score may still be used, but instead of using the raw score, the "purified" score that is reduced of potential bias is used. In the study conducted on the two-stage process, it was found that out of the 18 tests that were studied, the number of DIF items changed with using this two-stage iterative process on 15 of the tests. In fact, more commonly there were more questions that were flagged with DIF after the second stage, than in the first stage – specifically on the mathematics and science tests given at the high school level.[17]

Along with this study, the same group did another study examining gender DIF for a large-scale assessment over multiple age groups, testing both science and technology.[18] They were interested in identifying different classification of items that exhibit DIF including content categories, visual/spatial or reference components, and item type (whether it was multiple choice or open-response). Of the 222 items tested, 60 of them showed DIF. Of those 60, 87% of them favored male students. The items were classified into different levels of DIF: including minor DIF and moderate to high DIF. There were 23 questions that exhibited moderate to high DIF, and they all favored the male students. There were five different content areas the questions were split into including; earth and space sciences, inquiry, life sciences, physical sciences, and technology. For the items that exhibited DIF in the area of earth and space sciences (17 items), 88% favored male students and 12% favored female students. For the items that exhibited DIF in the area of inquiry (eight items), 62.5% favored male students and

37.5% favored female students. For the items that exhibited DIF in the area of life sciences (seven items), 71% favored male students and 29% favored female students. For the items that exhibited DIF in the area of physical sciences (13 items), 100% favored male students. Lastly for the items that exhibited DIF in the area of technology (15 items), 93% favored male students and 7% favored female students.[18] For the visual/spatial or reference component analysis 82 out of the 222 items had either one or two visual/spatial or reference components in the item, with 25 of them flagged with DIF and 23 of them favored male students. Of the 37 open response items, three of them were flagged with DIF that all favored female students. The other 57 flagged items were multiple-choice with 52 of them favoring male students.[18]

Another study was conducted that looked at format and content of the items on science achievement assessments using gender DIF. It was shown that male students tended to do better on critical response items, especially if the item involved a visual or spatial component and made the students use skills and/or knowledge they seemed to acquire outside of school.[19] Hamilton and Snow conducted a study investigating DIF on 12th grades science tests that featured both multiple-choice and constructed-response items and was part of the National Education Longitude Study of 1988.[20] Along with DIF, factor analysis was used (as well as student interviews) to determine different formats into which the items were grouped. Specifically, three factors were found splitting the items into different formats. The factors included Spatial Mechanical Reasoning (SM) which included items with a visual or spatial component, Quantitative Science (QS) which mostly contained items that tested chemistry or physics content knowledge and also involved some type of computation, and lastly Basic Knowledge

Reasoning (BKR) which included items where they students need to know a concept or reason their way through it, mainly in the content of biology and astronomy.[20] Of these items it was found that for SM, five of the questions showed DIF that favored male students and zero questions showed DIF that favored female students. For QS, zero of the items showed DIF that favored male students, whereas, two of the questions showed DIF that favored female students. Lastly, for BKR, seven questions showed DIF with two of them favoring male students and five of them favoring female students.[20]

2.2.5 Stereotype Threat

Testing carries with it the concern that the act of testing in itself may incorporate stereotype threat, thereby inducing lower performance for students who experience the threat.[21] Stereotype threat is defined as "being at risk of confirming, as self-characteristic, a negative stereotype about one's group."[21] Common recommendations emerging from studies of stereotype threat include ensuring that students are not requested to identify defining characteristics (e.g., gender or race/ethnicity) prior to testing and refraining from discussion of stereotypes, particularly expected or measured performance differences.[21-24] Researchers have also developed a Domain Identity Measure (DIM) that may be useful in identifying students at risk of stereotype threat.[22-23] Interestingly, a recent study has identified a problem-solving perseverance in female students who have exhibited stereotype threat.[24] All assessment materials in these studies were administered in such a way that stereotype threat was minimized with demographic information collected from sources other than directly from the student.

2.3 Common Item Equating

When comparing questions from different exams it is important that the exam scores are on the same scale as the scores that are used internally to match the proficiency levels of students. A typical method for scaling non-equivalent exams is common item equating.[9] There are multiple types of common item equating methods including, equating two groups that take different assessments, two groups that take the same assessments but in different orders, and lastly each group takes one common assessment (known as an anchor test) as well as two different assessments to be equated.[9] Depending on the type of assessments and how they are given will depend on which equating method is best to use. With the last type of equating one can use item response theory with the data from both the anchor tests and the assessments to equate the scores. One of the requirements for this type of equating is that all assessments including the anchor test measure the same latent trait.

2.4 Cognitive Models

2.4.1 Introduction

When determining why DIF happens it is essential to explore how students are solving these items. To do that one needs to investigate not only the students' problem solving process, but also to investigate how students solve problems based on how people process information. To better understand these processes, two cognitive

processes will be explored including the Information Processing Model and Dual-Processing Theory.

2.4.2 Information Processing Model

There are many models to explain the way learning happens, specifically concerning memory. Often the memory will be broken down into specific parts including acquisition, storage and retrieval. About 60 years after these parts were introduced they were put into models of how learning and memory work.[25] One model was proposed by Atkinson and Shiffrin, which revolves around the idea of taking in sensory information, storing it in the short-term memory, and finally importing it into the long term memory.[26] Awhile later, Baddeley expanded on this idea, but replacing the label of short-term memory with the working memory and adding a feedback loop (Figure 3).[27] Essentially anything that is perceived goes into the sensory memory and the long-term memory is scanned to see if what was perceived is important enough to put into the working memory. If it is, that information goes into the working memory where some of the information will be encoded into the long term memory and other information will be retrieved from the long-term memory and put into the working memory to interpret the information.[25] This process can be used with any type of learning, including problem solving. Later Johnstone connected the ideas from psychology with that of chemistry, incorporating looking at that learning of chemistry and how that is achieved using the Information Processing Model.[28]

2.4.3 Dual Processing Theory

While the idea of splitting the mind into two different systems is not a novel idea, naming these systems and defining their functions is.[29] Dual Processing Theory suggests that when solving a problem, one can proceed by two different cognitive processes. The first process called system one is automatic, fast, and intuitive.[29] This automated process requires little cognitive resources, attention or thought.[25] This process is often referred to as using a heuristic, which is essentially a mental shortcut that usually bypasses the working memory. The second system is a slower process that is logical, has the person more conscious, and may be limited by the capacity of their working memory. This system is often referred to as analytical reasoning where a person goes through a series of steps to be able to solve the problem similar to the process described by the Information Processing Model (but not exclusive to this model).[29] A recent study by Maeyer and Talanquer investigated the use of Dual Processing Theory on general chemistry II students through the use of a questionnaire and semi-structured interviews.[30] This study focused on the decision portion of problem solving. There are many steps that go into decision making, but generally all follow the same process. Eisenhardt and Zbaracki describe steps of problem identification, development and selection that are involved in making a decision.[31] They then go on to state that these steps are not always linear in process and in fact can be quite cyclical at times. The Maeyer and Talanquer study found that students often use heuristics incorrectly to justify their answers.[30] In fact for one question on the interview, students use a heuristic at least once 82.4% of the time. It was also stated that the students often used more than one heuristic when solving a problem either as a way to further explain their reasoning or when they were stuck after using the

first heuristic. There were four main heuristics that were evident. The first heuristic was recognition which is when the student is familiar with one thing in the problem (i.e. they recognized it) and then base their answer off of that. The second heuristic was representativeness, which is when students would base their answers off of similarities between the choices. The third heuristic was one-reason decision making. For this heuristic the students would pick one thing out of the answers and base their entire reason off of that. These three heuristics are domain independent and can be applied to many different areas. The last heuristic found was specific to chemistry and it was arbitrary trends. This is when the students would base their answers off of perceived trend on the periodic table.[30]

2.5 Eye Tracking

2.5.1 Introduction

While interviewing students, much information can be gained about the student's problem-solving process. By including assessments in the interview process both the student's performance as well as a self-reported mental effort can be collected. The self-reported mental effort is a subjective measure of the load on the student's working memory, but objective measures of the student's working memory cannot be captured by interview alone. The use of an eye-tracking system allows collection of performance, self-reported mental effort, as well as measure of time on task, pupil diameter, and scan paths. Time on task and pupil diameter are both objective measures of a student's load on their working memory while scan paths can give information about the student's

problem-solving process. By collecting information about the student's performance, self-reported mental effort, time on task, and pupil diameter it is possible to determine which system in Dual Processing Theory they are using to solve the item as well as any possible heuristics.

2.5.2 Eye Tracking Studies

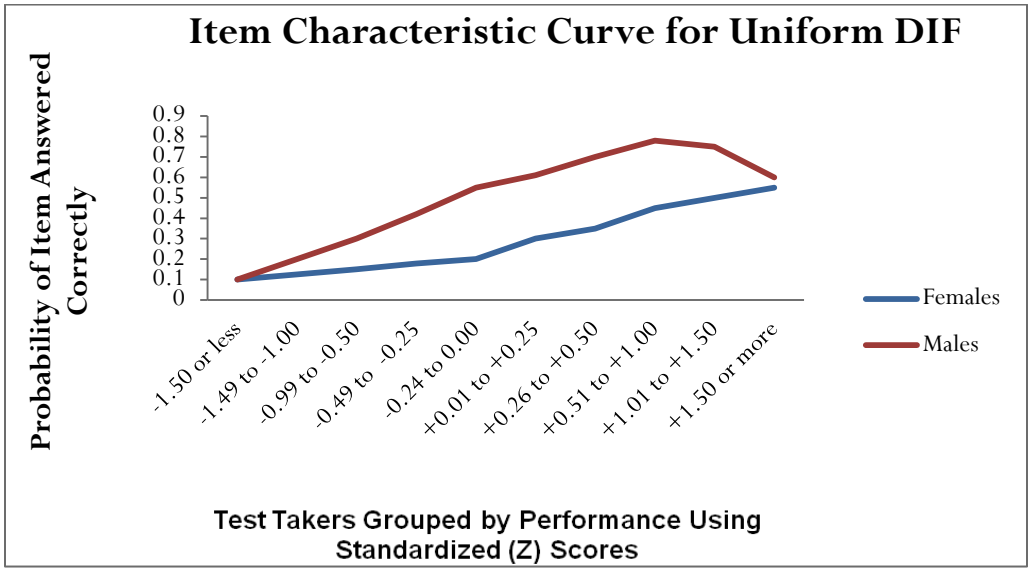
Eye-tracking instruments have been used in many areas of study including market research, medical research, psychology, and sports performance to name a few.[32] There has been a growing popularity in eye-tracking research because of the vast amount of information it can give. Most eye-tracking systems will give the time spent on the task, the pupil diameter, the scan path, as well as information about specific areas on which the subject is focusing. This information is extremely useful because it provides objective measures of the student's load on their working memory as well as insight into their problem solving processes. In a study conducted by Barrett, Tugade, and Engle they looked at how working memory capacity (WMC) is related to the type of processing that occurs.[33] It was found that people with a lower WMC tended to use a more automated (system one) type of processing, whereas students with a larger WMC tend to be slower on their tasks as they are examining multiple interpretations. In their review on all the literature that relates WMC and Dual Processing Theory, it was suggested that how WMC is measured relates to many processes, including Dual Processing Theory. In fact it was stated that, "we suggest that person-level variance captured in dual-process inspired experiments …[can] be meaningfully explained by individual differences in

WMC."[33] A study conducted by Kahneman and Beatty discovered that pupillary dilation is related to the cognitive load on the working memory.[34] In a review on pupillary responses it was stated that usually when set to perform a task there is a short delay in dilation, then the dilation and a rapid decline after finishing the task. It was also stated that once the cognitive load has reached its maximum adding more tasks will no longer increase the dilation of the pupil. Once the load is at a maximum, so is the dilation.[35] Therefore, collecting information about pupil diameter, time on task, and other measures more insight into the students' problem solving process is obtained.

There are many pieces of information that go into identifying DIF items, determining the persistence of the items that exhibit DIF, the possible causes of DIF, and why DIF is happening. By using the information from the literature, as well as tested theories and statistical methods it is possible to answer these questions and collect more information and understanding about the items that exhibit DIF.

Figure 1: Item characteristic curves exhibiting uniform and non-uniform DIF.

(a)



(b)

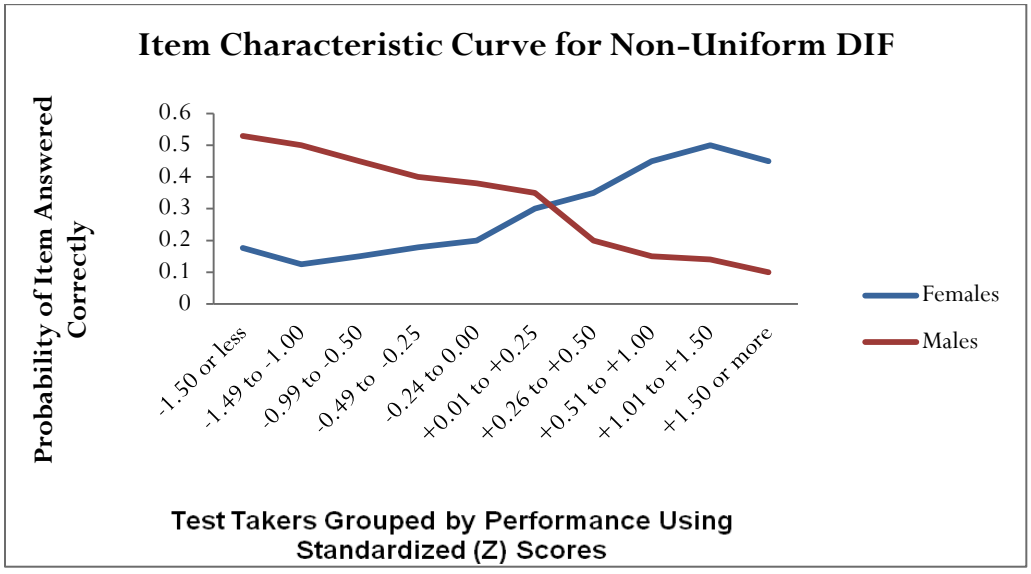


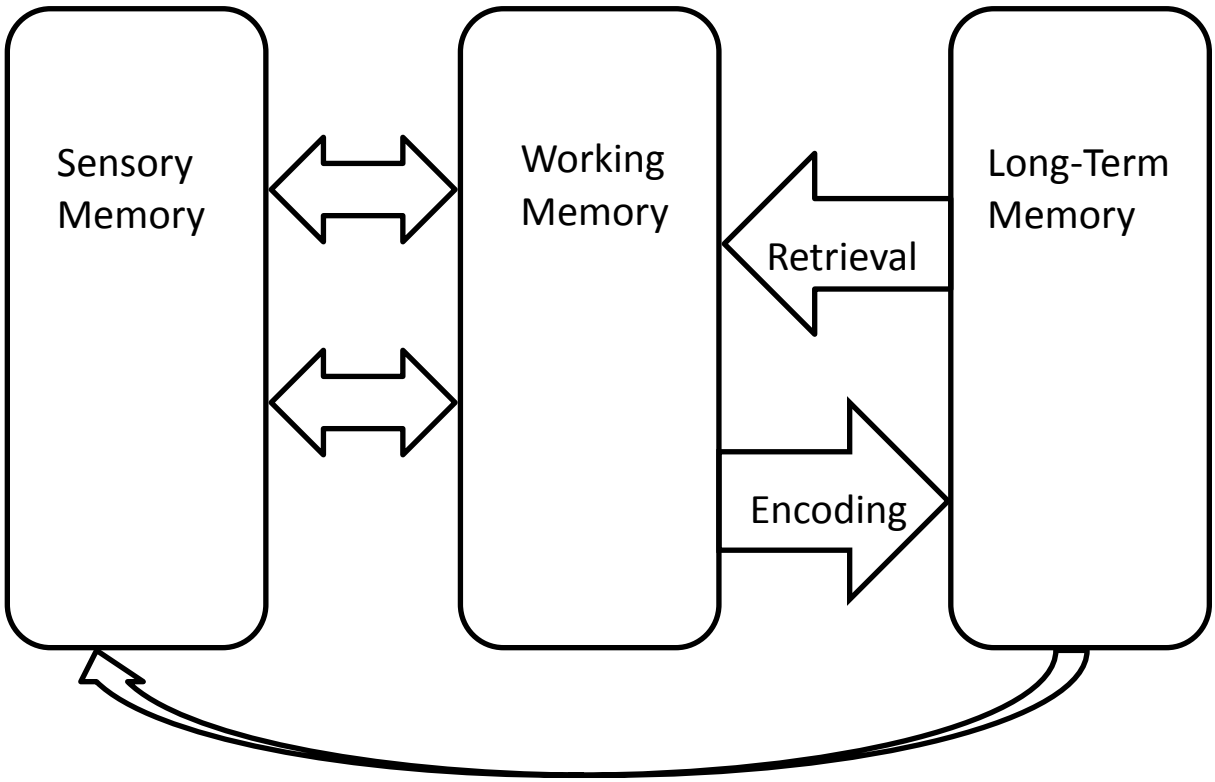
Figure 2: The Mantel-Haenszel statistic.

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

$$\alpha_i = \frac{p_{r_i} / q_{r_i}}{p_{f_i} / q_{f_i}} = \frac{\frac{a_i / (a_i + b_i)}{b_i / (a_i + b_i)}}{\frac{c_i / (c_i + d_i)}{d_i / (c_i + d_i)}} = \frac{a_i / b_i}{c_i / d_i} = \frac{a_i d_i}{b_i c_i}$$

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

Figure 3: The Information Processing Theory.



References

1. Maccoby, E. E.; Jacklin, C. N., *The Psychology of Sex Differences*. Stanford University Press: Stanford, California, 1974.

2. S&E Degrees: 1966-2008: National Center for Science and Engineering Statistics. http://www.nsf.gov/statistics/nsf11316/content.cfm?pub_id=4062&id=2 (accessed May 26).

3. Cole, N. S. *The ETS Gender Study: How Females and Males Perform in Eductional Settings*; Educational Testing Service: Princeton, NJ, 1997; pp 1-36.

4. Beller, M.; Gafni, N., The 1991 International Assessment of Education Progress in Mathematics and Sciences: The Gender Differences Perspective. *Journal of Educational Psychology* **1996,** *88* (2), 365-377.

5. Halpern, D. F.; LaMary, M. L., The Smarter Sex: A Critical Review of Sex Differences in Intelligence. *Educational Psychology Review* **200,** *12* (2), 229-246.

6. Holland, P. W.; Thayer, D. T., Differential Item Functioning and the Mantel-Haenszel Procedure. In *Test Validity*, Wainer, H.; Braun, H. I., Eds. Lawrence Erlbaum Associates, Inc.: Hillsdale, N. J., 1988; pp 129-145.

7. Hambleton, R. K.; Swaminathan, H.; Rogers, H. J., *Fundamentals of Item Response Theory*. Sage: Newbury Park, CA, 1991.

8. Swaminathan, H.; Rogers, H. J., Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement* **1990,** *27* (4), 361-370.

9. Crocker, L.; Algina, J., *Introduction to Classical & Modern Test Theory*. Holt, Rinehart and Winston: New York, 1986.

10. Holland, P. W.; Wainer, H., *Differential Item Functioning*. Lawrence Erlbaum Associates: Hillsdale, New Jersey, 1993.

11. van der Linden, W. J.; Hambleton, R. K., Item Response Theory: Brief History, Common Models, and Extensions. In *Handbook of Modern Item Response Theory*, van der Linden, W. J.; Hambleton, R. K., Eds. Springer: New York, 1997; pp 1-31.

12. Embretson, S. E.; Reise, S. P., *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, INC.: Mahwah, New Jersey, 2000.

13. Zumbo, B. D., A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type Ordinal Item Scores. Directorate of Human Resources Research and Evaluation, Department of National Defense: Ottawa, ON, 1999.

14. Schmitt, A. P.; Dorans, N. J., Differential Item Functioning for Minority Examinees on the SAT. *Journal of Education Measurement* **1990,** *27* (1), 67-81.

15. Holme, T., Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education* **2003,** *80* (6), 594.

16. Gierl, M.; Khaliq, S. N.; Boughton, K., Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications. In *Annual Meeting of the Canadian Society for the Study of Education*, Sherbrooke, Quebec, 1999.

17. Zenisky, A. L.; Hambleton, R. K., Detection of Differential Item Functioning in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach. *Educational and Psychological Measurement* **2003a,** *63* (1), 51-64.

18. Zenisky, A. L.; Hambleton, R. K.; Robin, F. *DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices*; University of Massachusetts Amherst: Amherst, MA, 2003b; pp 1-22.

19. Hamilton, L. S., Gender Differences on High School Achievement Tests: Do Format and Content Matter? *Educational Evaluation and Policy Analysis* **1998,** *20* (3), 179-195.

20. Hamilton, L. S.; Snow, R. E. *Exploring Differential Item Functioning on Science Achievement Tests*; 483; National Center for Research on Evaluation, Standards, and Student Testing. : Los Angeles, CA, 1998; pp 1-43.

21. Steele, C. M.; Aronson, J., Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology* **1995,** *69*, 797-811.

22. Conway-Klaasen, J. M. "Stereotype Threat's Effect on Women's Achievement in Chemistry: The Interaction of Achievement Goal Orientation for Women in Science Majors. University of Nevada, Las Vegas, 2010.

23. Aronson, J.; Justina, M. J.; Good, C.; Keough, K.; Steele, C. M.; Brown, J., When White Men Can't do Math: Necessary and Sufficient Factors in Stereotype Threat. *Journal of Experimental Social Psychology* **1999,** *35* (29-46).

24. Carr, P. B.; Steele, C. M., Sterotype Threat and Inflexible Perseverance in Problem-Solving. *Journal of Experimental Social Psychology* **2009,** *45*, 853-859.

25. Bruning, R. H.; Schraw, G. J.; Norby, M. M.; Ronning, R. R., Sensory, Short-Term, and Working Memory. In *Cognitive Psychology and Instruction*, 4th ed.; Harlan, M., Ed. Pearson Education, Inc.: Upper Saddle River, New Jersy, 2004.

26. Atkinson, R. C.; Shiffrin, R. M., Human memory: A Proposed System and its Control Processes. In *The psychology of learning and motivation: Advances in research and theory*, Spence, K. W.; Spence, J. T., Eds. Academic Press: San Diego, 1968; Vol. 2, pp 89-195.

27. Baddeley, A. D., *Human memory: Theory and Practice*. Allyn & Bacon: Boston, 1990.

28. Johnstone, A. H., The Development of Chemistry Teaching: A Changing Response to Changing Demand. *Journal of Chemical Education* **1993,** *70* (9), 701-705.

29. St. B. T. Evans, J.; Frankish, K., *In Two Minds Dual Processes and Beyond*. Oxford University Press: Oxford, 2009.

30. Maeyer, J.; Talanquer, V., The Role of Intuitive Heuristics in Students' Thinking: Ranking Chemical Substances. *Science Education* **2010**, 1-22.

31. Eisenhardt, K. M.; Zbaracki, M. J., Strategic Decision Making. *Strategic Management Journal* **1992,** *13*, 17-37.

32. SMI Gaze & Eye Tracking Systems. http://www.smivision.com/ (accessed May 27).

33. Barrett, L. F.; Tugade, M. M.; Engle, R. W., Individual Difference in Working Memory Capacity and Dual-Process Theories of the Mind. *Psychological Bulletin* **2004,** *130* (4), 553-573.

34. Kahneman, D.; Beatty, J., Pupil Diameter and Load on Memory. *Science* **1966,** *154*, 1583-1585.

35. Beatty, J., Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin* **1982,** *91* (2), 276-292.

Chapter 3: Experimental

3.1 Introduction

The experimental section is broken down into three main sections. The first section contains experimental methods to identify DIF persistence, mainly identifying items that were flagged for DIF that favors one gender verses the other and investigating the degree to which the DIF was persistent. The second part contains experimental methods to examine the causes of DIF by examining content and construct clones. The final section comprises methods for an eye tracking study to investigate the reasons why DIF is happening by looking at students' problem solving processes. The IRB approval number obtained through the Institutional Review Board at the University of Wisconsin - Milwaukee for the eye tracking study is 09.047.

To be consistent with the literature on DIF, the use of the term "gender" to differentiate subgroups based on sex will be used throughout this dissertation. The socially constructed associations of gender were not collected in this study, but rather the self-reported (when available) biological difference of sex was collected for each student.[1] The demographic data was obtained through a request to University institutional research and obtained approximately two weeks after the start of each semester. In addition to gender, student ACT scores and sub-scores, year in school and intended or declared major were requested as some were used as external criteria.

3.2. DIF Persistence

3.2.1 Identifying Items that Exhibit DIF

The research was conducted over six semesters at a large, public, doctoral university in the Midwest with a population of 30,000 students.[2] To identify items that were flagged with DIF, a two-stage[3] DIF analysis using the Mantel-Haenszel statistic[4] was performed on two trial tests from the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS DivCHED EI). For an item to be considered as having DIF the significance had to be lower than 0.051 using the Mantel-Haenszel statistic. The two-stage DIF analysis included a purification process, where the results of the first stage of the two-stage DIF analysis were used for the purification step. As mentioned in Chapter 2, the first stage of the analysis uses the score the participants received on the assessment as their matching criterion for DIF analyses. The issue with using the score the participants receive on the assessment is that if there are items on the assessment that are flagged with DIF, then the score on the assessment could be biased as well. The second stage of the analyses removes the items with the highest DIF and recalculates the score, then runs the DIF statistical analysis again using the "purified" score.[3] Only questions that were flagged with DIF in both stages of the analyses were considered for the analysis of persistent DIF. The Mantel-Haenszel statistic was used because it is a common method in literature for determining DIF and due to its accessibility through SPSS software.

The two trial examinations had 70 unique items each that were written by a committee of volunteers usually teaching in the field that the exam is being constructed.

The committee first meets to discuss what content should be on the exams and then writes various multiple choice questions as possible exam items. After much discussion and revisions, the items are ready to be tested by students throughout the U.S. as trial examinations.[5] For the two trial tests used in this study, the number of participants is shown in Table I. The trial examinations were tested at seven different institutions, typically for a final examination or equivalent high-stakes testing. The item performance for each student, along with their gender, was reported to the Examinations Institute. In the case that the gender was not reported, an attempt was made to assign gender based on student's name by using a program that generates the probability of the name being assigned for each sex.[6] The gender was only assigned if the probability of the name was greater than 50% for one gender over the other. For instance, Kelley is only 1.081 times more likely to be a girl, so their data would have excluded in the analysis.

3.2.2 Classwide DIF Persistence

The questions that were flagged with having DIF for both stages of the two-stage[3] analyses were then use in the original format (if they were not used on the final version of the standardized examination) or cloned by content and placed on the multiple choice section of different hourly, high-stakes assessments throughout six different semesters of general chemistry I. After these assessments were taken by the students the items were again examined for the presence of DIF by performing a two-stage DIF analysis using the Mantel-Haenszel statistic. The items that statistically showed DIF support the claim that there are certain items that have significant DIF based on gender subgroup. To then

support the claim that DIF is real, the persistence of the DIF items was investigated beyond using an internal matching criterion. As mentioned before, for an item to exhibit DIF, the subgroups must be matched on equal proficiency as to avoid any potential bias from matching a high performing student of one gender versus of low performing student of the opposite gender. For many DIF analyses, the matching criterion used is the score the student receives on the exam. However, to study the degree to which DIF is persistent, other relevant matching criteria were used to re-match the groups. By using other relevant measures of proficiency as matching criteria it gave more validation into determining the persistence of DIF. The other relevant measures of proficiency can be found in Table II. The Mantel-Haenszel statistic was run again to determine if the DIF is still statistically significant. For a DIF item to be considered as being persistent it had to be statistically significant against both stages of the two stage analyses and the majority of the relevant measures of proficiency unless otherwise stated. Figure 1 shows both an item that would be considered to show persistent DIF (a) and an item that doesn't show persistent DIF (b). Notice that the item in Figure 1b was flagged for DIF in both stages of the two-stage analyses but wasn't flagged for the majority of the other relevant measure of proficiency; therefore, it wasn't considered to show persistent DIF. To determine which subgroup was favored for the items with persistent DIF, a simple difference in difficulty was performed (the number of students who answered the item correctly in that subgroup versus the number of students in the subgroup who were given the opportunity to answer the question).[7] Table III shows the number of participants for the assessments that were included in the analysis. These numbers vary from

examination to examination depending on which students took the exam or dropped the course.

3.2.3 Standardized Examination DIF Persistence

A DIF persistence analysis was conducted on the ACS DivCHED EI Toledo Chemistry Placement Examination,[8] the ACS DivCHED EI First Term General Chemistry Paired Questions final examination,[9] and the ACS DivCHED EI General Chemistry Conceptual final examination[10]. The number of participants for each exam over the six semesters is shown in Table IV. The procedure was the same for these examinations as listed above with the exception of having different matching criteria. The relevant measures of proficiency that were used as matching criterion for this analysis are shown in Table V. The last three examinations in the table were not used as an external criterion if the items being tested were from that examination. For example, if the items that were being tested for persistent DIF were from the ACS DivCHED EI Toledo Chemistry Placement Examination the score received on the assessment was used as the internal matching criterion in the two-stage analyses and was not applicable as an external matching criterion.

3.3 Cloning Items

3.3.1 Categorizing Persistent DIF Items

To investigate the possible causes of DIF, each item that was flagged with persistent DIF was categorized by both its general chemistry content area (Table VI) as well as its construct (format of the item). The four format categories were visual-spatial, calculation, reasoning, and specific chemical knowledge. These were adapted from a study of DIF on a science achievement examination where factor analysis was performed to determine different formats of items.[11] Both the content and the format were assigned to each item by two separate raters. If there were any discrepancies between the two raters, a discussion was had until full agreement had been reached. A third rater then periodically checked items for full agreement. If there wasn't full agreement, again the raters discussed the item until full agreement was reached. Once the items were categorized into groups, it was easier to see patterns emerge as to which content areas or formats had a higher number of items flagged with DIF that favored one gender over the other.

3.3.2 Cloning Persistent DIF Items

To gather more information about the causes of DIF, several items were cloned either based on the content or format (i.e. to clone the content of the item, the format of the item was kept the same and only the content changed or vice versa). See Figure 2 as an example. The original item is highlighted and on the left and the cloned item based on

format is shown on the right. The content of this item stayed exactly the same, only the format changed. Figure 3 shows how the same item was cloned base on content, making sure the format of the item was exactly the same. The clones were placed on the multiple-choice section of the hourly examination throughout four semesters and analyzed using a two-stage DIF analysis using the Mantel-Haenszel statistic along with using different relevant measures of proficiency to determine the persistence of DIF on the clones. These items again were classified by two raters to determine the content and construct of the items and checked by a third rater for agreement. Lastly, all questions given (both those that exhibited persistent DIF and those that did not) were classified by their content and construct to see if patterns emerged over different semesters and different clones of the items.

3.3.3 Common Item Equating

The best way to compare these items that exhibit persistent DIF would be to give the same items over and over again to the same sample of students. Because that is not possible, due to students becoming familiar with these items, cloning the items was a necessity, as was giving the items to different samples of students. These items were given over 6 semesters, so while the population stayed the same (first term general chemistry I students) the samples changed every semester. As a way to compare the results of a DIF analysis for one cloned item to another (comparing the item's Chi-Squared values calculated through the Mantel-Haenszel statistic), it was necessary to ensure that the samples were the same. To do this common item equating with an anchor

test was performed using item response theory.[7] Common item equating is a way to make sure that the scores from two different assessments are measuring the same thing. Essentially, scores from two different tests can be considered to be equal if they are measuring the same latent trait. Another condition includes the latent trait needs to be of percentile rank and variability of the scores needs to be considered equal.[7] There are different ways to do common item equating; because for each semester there is a different sample of students who take different hourly exams, in addition to a common test that is the same every semester (the anchor test), it was decided to do common item equating with Item Response Theory using an anchor test.[7] This was done with the PARSCALE program by Scientific Software International.[12] For information on the PARSCALE command file see Appendix A. The anchor test used was the ACS DivCHED EI First Term General Chemistry Paired Questions final examination and every test and semester was equated to the Fall 2009 semester. In order to obtain the most valid measure of equating students, the final examination for the course was selected as the anchor test because it was given after a semester of the same instruction to all students under the same conditions of testing. A 2-parameter fit was chosen because of the sample size of students. Also the use of priors was incorporated to help calibration for they give a little extra stability to estimation but doesn't bias or distort the estimates. To use common item equating with a 2-parameter fit in item response theory, the location values (*b*) were calculated for all the anchor items for each semester. Then the difference in *b* values for each item from the anchor test was calculated comparing each semester back to Fall 2009. For example, the location value for question 1 on the Spring 2009 ACS DivCHED EI First Term General Chemistry Paired Questions final examination was subtracted from

the location value for question 1 on the Fall 2009 ACS-EI First Term General Chemistry Paired Questions final examination, and so on for each question. Then the average of the differences (*m*) was calculated. This average of the differences was then added to each students' latent trait score (Z-score) for each hourly examination given. A Z-score is a way to define the student's score in relation to how they performed compared to the group.[7] The equation is shown below in 3-1, where *Z* is the Z-score, *X* is the score the student received on the assessment, μ is the group average, and σ is the group standard deviation.

$$Z = \frac{X-\mu}{\sigma} \qquad (3-1)$$

This was done individually by gender subgroup to equate the scores for each exam so the examinations would be on the same scale that was used for the Fall 2009 semester. Once the student's new latent trait scores were calculated into the equated scaled score for each student, a one-stage DIF analysis was performed using the Mantel-Haenszel statistic to examine for DIF using the equated scaled scores.

3.3 Eye Tracking Studies

3.3.1 Assessment-like Interviews

To investigate why gender DIF is happening, it was important to understand how the students were solving the tasks. To investigate this, 76 questions (comprised both of DIF items and non-DIF items) were encoded into a SensoMotoric Instruments (SMI) Eye tracking instrument (Figure 4) using Experiment Center software to construct the

interviews and BeGaze software to analyze the results. There were 62 undergraduate students and 13 graduate students who participated in the interviews. See Table VII for a breakdown of gender for the participants. For these interviews the students sat in front of a desk mounted remote eye tracking device (RED), performed a calibration, and then answered the 76 multiple choice questions by clicking their choice on the screen and then advancing to the next question. While this gave information about the students solving the tasks such as their time on task, the pupil diameter, and scan path to see what they were looking at, for how long and in what order, this didn't give specific information into how they were solving the task so another set of interviews was conducted.

3.3.2 Semi-Structured Interviews

Semi-structured interviews using an eye tracking system were conducted in the Spring and Fall of 2012 at the end of the semesters. For these interviews the students not only solved the multiple choice items that were shown on the eye-tracking instrument, but they also verbalized their thought process as they were working the tasks to give more information into the student's problem-solving process.[13-14] For the Fall 2012 interviews, there was an extra question and an extra exercise included. The extra exercise involved a set of numbers flashing on the screen and then the student having to either recite the set of numbers or to recite a transformed set of numbers. For example, if the set of numbers 0 5 8 6 were shown on the screen the student would see these numbers and then a new screen would pop up and the student would have to recite these numbers out-loud while looking at the screen. For the transformed set of numbers the student

needed to add the digit one to each number upon recitation, so in the case listed above the student would recite 1 6 9 7. They performed this exercise including both the plain recitation and the transformed recitation at the beginning and end of answering the multiple choice general chemistry questions. The string of numbers was the same for each student and started with just one digit and progressed up to ten. They were allowed to view each string of numbers for a time of a half of second per digit. For the string of digits listed above the participants would see them for two seconds before the recite screen would pop up. This exercise was performed to get information about each student's maximum pupil diameter to use as a comparison for pupil diameter dilation (as the pupil diameter should be maxed out during the recitation of seven to eight digits and lower for the transformation part) during the multiple choice general chemistry questions.[15]

The Spring 2012 interviews consisted of 19 general chemistry questions. There were 24 participant; five of which were male students and 19 female students. The Fall 2012 interviews contained 20 general chemistry questions (19 of which were the same from the Spring 2012 interviews). There were 25 participants; ten of which were male students and 15 female students, for a total of 15 male students and 34 female students participating in both sets of interviews. Using BeGaze, the data collection program for the eye-tracking instrument, the participants' answers to the items, their self-reported mental effort, their total time on task, and their average pupil diameter were collected. For their self-reported mental effort the prompt is shown in Figure 5. This is a question they were all familiar with as it was often asked in their course lecture. All of the items and mental effort pages were self-advanced by the students giving them as much time as

they needed on each page. Most of the interviews lasted approximately 45-60 minutes. These interviews were also video taped and transcribed to capture the participants' problem-solving process for each item. Students' t-test was used to examine for differences in the male students and female students performance, mental effort, and total time on task. Instead of using the participant's average pupil diameter per fixation, the maximum pupil diameter was more of interest. To ensure that the maximum pupil diameter was a result from solving the item and not an anomaly, each average pupil diameter versus fixation start time was graphed along with using a set of criteria to ensure the correct maximum pupil diameter was used. The set of criteria is shown in Figure 6. If all of these criteria were met the average pupil diameter for that fixation was included to determine the maximum pupil diameter for the task. As a way to filter all of the results from the semi-structured interviews it was decided to look only at the participants for each item whose results suggest the use of an incorrect heuristic. To determine that only the participants who got the question wrong, had a low mental effort, a low time on task, and a small pupil diameter were examined. If the participant met the criteria for that item, the transcript was read to determine the student's problem-solving process and if there was evidence of an incorrect heuristic. This was determined for each participant that fit the criteria for each item by two separate raters. If there were any discrepancies between the two raters, a discussion was had until full agreement had been reached. A third rater then periodically checked items to see if there was full agreement. If there wasn't a full agreement, again the rater discussed the item until a full agreement was reached.

Table I: Participants of the two trial test given by The American Chemical Society's Division of Chemical Education, Examinations Institute.

	Trial Test A	Trial Test B
Number of Question	70	70
Male Participants	735	758
Female Participants	504	518
Total Participants	1239	1279
Number of Institutions	7	7

Table II: The external relevant measures of proficiency for the hourly examinations.

Relevant Measure of Proficiency	Raw Score (points (pts))
Total Exam Score	100 pts*
ACS DivCHED EI Toledo Chemistry Placement Examination[8]	60 pts
Scale Literacy Skills Test (SLST)	45 pts
Cumulative Examination	44 pts**
ACS DivCHED EI First Term General Chemistry Paired Question Final Examination[9]	40 pts
ACS DivCHED EI Conceptual Exam Final Examination[10]	40 pts

* This is the only exam where raw scores were not used. Also there were extra credit options given on most hourly examinations so it was possible for students to receive over 100 pts.

** Each cumulative exam was out of 44 pts, except the Spring 2010 exam and Spring 2013 were out of 40 pts and the Fall 2010 exam was out of 43 pts.

Table III: The participants for the six semesters of hourly examinations.

Semester	Exam Number	Number of Multiple Choice Items	Male Participants	Female Participants	Total Participants
Fall 2009	1	24	167	207	374
	2	24	152	191	343
	3	24	140	178	318
	4	44	132	168	300
Spring 2010	1	24	173	188	361
	2	24	158	172	330
	3	24	149	162	311
	4	40	145	153	298
Fall 2010	1	24	158	191	349
	2	24	139	170	309
	3	24	132	163	295
	4	43	120	158	278
Fall 2011	1	24	171	162	333
	2	24	153	144	297
	3	24	150	130	280
	4	44	143	126	269
Fall 2012	1	24	167	176	343
	2	24	156	165	321
	3	24	147	153	300
	4	44	141	149	290

Table III: cont.

Semester	Exam Number	Number of Multiple Choice Items	Male Participants	Female Participants	Total Participants
Spring 2013	1	24	162	171	333
	2	24	155	165	320
	3	24	134	152	286
	4	40	128	153	281

Table IV: The participants for the six semesters of standardized examinations.

Semester	Number of Multiple Choice Items	Male Participants	Female Participants	Total Participants
ACS DivCHED EI Toledo Chemistry Placement Examination[8]	60	975	1,100	2,075
ACS DivCHED EI First Term General Chemistry Paired Question Final Examination[9]	40	811	929	1,740
ACS DivCHED EI Conceptual Exam Final Examination[10]	40	810	928	1,738

Table V: The external relevant measures of proficiency for the standardized examinations.

Relevant Measure of Proficiency	Raw Score (points (pts))
ACT Math Score	36 pts
ACT Science Score	36pts
ACT Composite Score	36pts
ACS DivCHED EI Toledo Chemistry Placement Examination[8]	60 pts
ACS DivCHED EI First Term General Chemistry Paired Question Final Examination[9]	40 pts
ACS DivCHED EI Conceptual Exam Final Examination[10]	40 pts

Table VI: The chemistry content areas that the persistent DIF questions were categorized as.

Content (general)
Classification of Matter
Physical & Chem Properties of Matter
Scale
Measurement
Nuclear Symbols
Periodic Table (Classify)
Isotopes
Chemical Formulas General
Nomenclature
Avogadro's Number
Greatest/Least Number of Atoms
Molecular Mass
How Many Grams/Moles/Etc.
Empirical Formula
Empirical Formula-% Comp
% Composition
% Composition by mass of O_2/H_2
Chemical Reactions
Stoichiometry-General
Limiting Reagents
Percent Yield
General Properties of Aq Solns
Precipitation
Precipitation
Acid Base
Oxidation-Reduction Reactions
Concentration of Solutions
Dilutions
Solution Stoichiometry
Gas Laws
Ideal Gas Equation
Gas Stoichiometry

Table VI: Cont

Content (general)
Solution Stoichiometry
Gas Laws
Ideal Gas Equation
Gas Stoichiometry
Partial Pressures
Kinetic Molecular Theory
Deviations from Ideal Behavior
Thermodynamics (General)
Enthalpy
Content (general)
Calorimetry
Heat
Standard Enthalpy of Reaction
Electromagnetic Radiation
Photoelectric Effect
Bohr's Theory
Quantum Numbers
Electron Configuration
Valence Electrons
Electron Configuration of Ions
Atomic Radius
Ionic Radius
Ionization Energy
General Trends in Chemical Properties
Covalent Bond
Lattice Energy
Electonegativity
Lewis Structures
Formal Charge
Resonance
Bond Enthalpy
Molecular Geometry

Table VI: Cont.

Content (general)
Bond Angles
Polarity
Hybridization in Molecules
Molecular Orbital Theory
Organic Molecular Formulas
KMT of Liquids and Solids
Intermolecular Forces
Properties of Liquids
Crystal Structures
Phase Changes
Phase Diagrams

Table VII: Undergraduate and graduate students who participated in the eye-tracking interviews in an assessment only format.

	Undergraduate Students	Graduate Students
Number of Questions	76	76
Male Participants	26	9
Female Participants	36	4
Total Participants	62	13

Figure 1: Examples of questions that were analyzed for persistent DIF. (a) An item that was flagged with persistent DIF that favored male students. (b) An item that was flagged with internal DIF, but not for persistent DIF.

(a)

Favors	Item #	Internal Measures of Proficiency						
			Total Exam Score	Toledo Exam	SLST	Cumulative Exam	Paired Question Final	Conceptual Final
Male	1	X	X	X	X	X	X	X

(b)

Favors	Item #	Internal Measures of Proficiency						
			Total Exam Score	Toledo Exam	SLST	Cumulative Exam	Paired Question Final	Conceptual Final
-	2	X				X		X

Figure 2: An item cloned based on the changing the format of the item while keeping the content the same. The question on the left was flagged with persistent DIF that favored female students, whereas the question on the left had no DIF.

According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?	
(A)	A bonding σ molecular orbital and an antibonding σ molecular orbital
(B)	A bonding σ molecular orbital and an antibonding π molecular orbital
(C)	A bonding π molecular orbital and an antibonding σ molecular orbital
(D)	A bonding π molecular orbital and an antibonding π molecular orbital

According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?

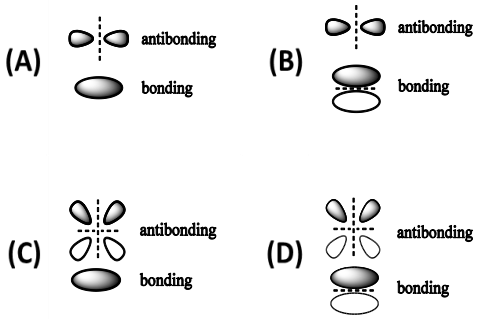


Figure 3: An item cloned based on the changing the content of the item while keeping the format the same. The question on the left was flagged with persistent DIF that favored female students, whereas the question on the left had no DIF.

According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?	
(A)	A bonding σ molecular orbital and an antibonding σ molecular orbital
(B)	A bonding σ molecular orbital and an antibonding π molecular orbital
(C)	A bonding π molecular orbital and an antibonding σ molecular orbital
(D)	A bonding π molecular orbital and an antibonding π molecular orbital

	According to molecular orbital theory, if the two 1s atomic orbitals of hydrogen are mixed, what are the resulting lowest two molecular orbitals?
(A)	2 electrons in the sigma bonding and 0 electrons in the sigma antibonding
(B)	2 electrons in the pi bonding and 0 electrons in the sigma antibonding
(C)	2 electrons in the sigma bonding and 0 electrons in the pi antibonding
(D)	2 electrons in the pi bonding and 0 electrons in the pi antibonding

Figure 4: A photograph of the SMI eye tracking instrument, including both the remote eye tracking device (RED) and the analysis computer.

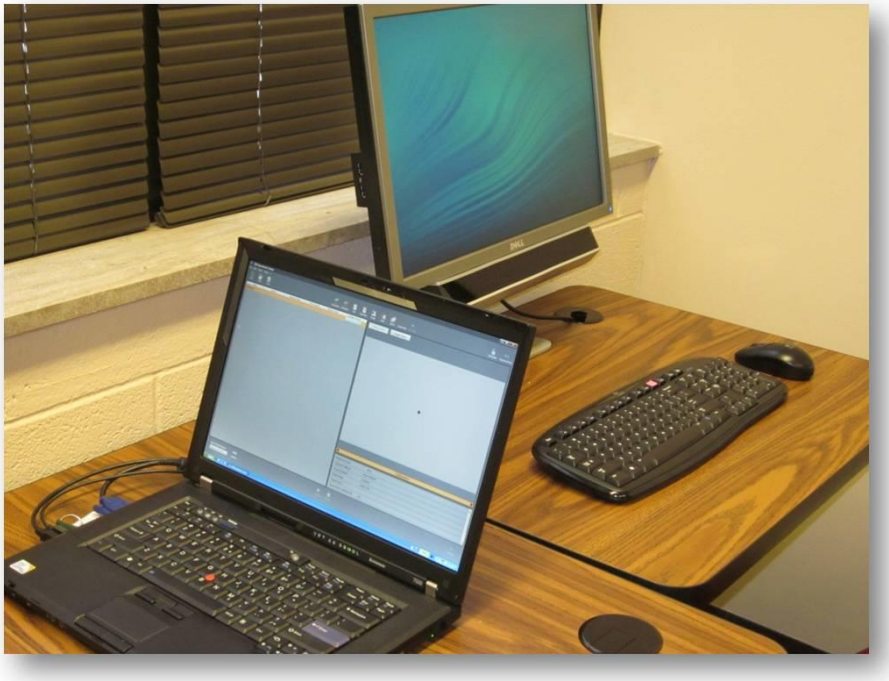
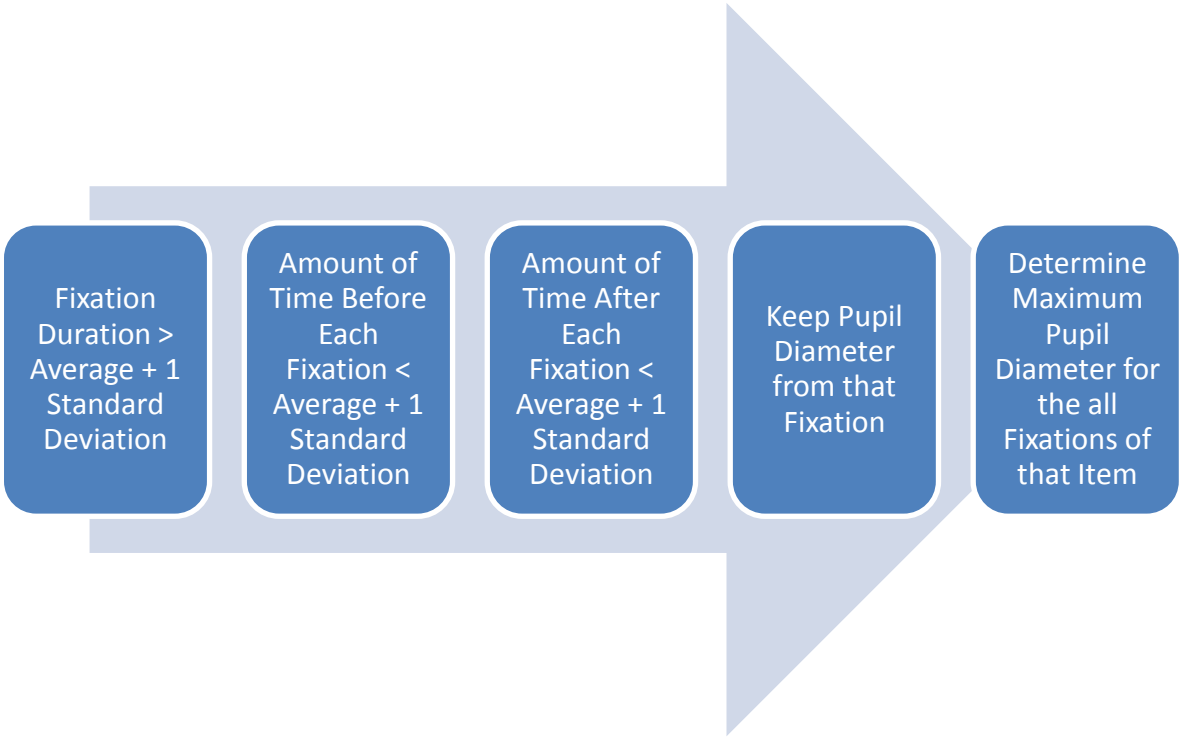


Figure 5: The prompt for students to report their mental effort.

How much mental effort did you expend on the previous question?

o Very low amounts
o Low amounts
o Moderate amounts
o High amounts
o Very high amounts

Figure 6: The set of criteria for determining the maximum pupil diameter per item on the interview.



References

1. Deaux, K., Sex and Gender. *Annual Review of Psychology* **1985,** *36*, 49-81.

2. About UWM. http://www4.uwm.edu/discover/about.cfm (accessed May 8).

3. Zenisky, A. L.; Hambleton, R. K., Detection of Differential Item Functioning in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach. *Educational and Psychological Measurement* **2003a,** *63* (1), 51-64.

4. Holland, P. W.; Thayer, D. T., Differential Item Functioning and the Mantel-Haenszel Procedure. In *Test Validity*, Wainer, H.; Braun, H. I., Eds. Lawrence Erlbaum Associates, Inc.: Hillsdale, N. J., 1988; pp 129-145.

5. Holme, T., Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education* **2003,** *80* (6), 594.

6. Baby Name Guesser. http://www.gpeters.com/names/baby-names.php (accessed May 8).

7. Crocker, L.; Algina, J., *Introduction to Classical & Modern Test Theory*. Holt, Rinehart and Winston: New York, 1986.

8. Toledo Chemistry Placement Examination. ACS DivCHED, Examinations Institute: 1992.

9. First Term General Chemistry Paired Questions ACS DivCHED, Examinations Institute: 2005.

10. Conceptual Exam ACS DivCHED, Examinations Institute: 2008.

11. Hamilton, L. S.; Snow, R. E. *Exploring Differential Item Functioning on Science Achievement Tests*; 483; National Center for Research on Evaluation, Standards, and Student Testing. : Los Angeles, CA, 1998; pp 1-43.

12. Muraki, E.; Bock, D., IRT from Scientific Software International: PARSCALE. 2002.

13. Bowen, C. W.; Bodner, G. M., Problem-Solving Processes Used by Graduate Students While Solving Tasks in Organic Synthesis. *International Journal of Science Education* **1991,** *13*, 143-158.

14. Herron, J. D.; Greenbowe, T. J., What Can We Do About Sue: A Case Study of Competence. *Journal of Chemical Education* **1986,** *63* (6), 528-531.

15. Kahneman, D.; Beatty, J., Pupil Diameter and Load on Memory. *Science* **1966,** *154*, 1583-1585.

Chapter 4: Results

4.1 Introduction

The results section will be split into four main sections. The first section will focus on the identification of items that exhibited potential Differential Item Functioning (DIF). Here the results for the two trial tests from the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS- EI) will be presented. The next section focuses on the results from the classwide analysis using both hourly examinations and ACS-EI standardized assessments. This section will show that certain items have persistent DIF, supporting the hypothesis that DIF is real and that certain types of items favor one gender versus the other. The third section will highlight the results to determine the possible causes of DIF, including cloning the items by content and format, as well as, using common item equating to compare clones given to different samples of students. The last section focuses on the results of the eye-tracking interviews which gives insight into why DIF is happening by focusing on students' problem-solving processes.

4.2 Identifying Items that Exhibit DIF

4.2.1 Trial Examinations

For the DIF analysis of the two trial tests from the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS-EI), 33 items were flagged for DIF using a two-stage DIF analysis. There were 14 items that were flagged with DIF on Form A, where seven items that were flagged favored female students and seven items that were flagged favored male students. For Form B there were 19 items that were flagged with DIF, where eight items that were flagged favored female students and 11 items that were flagged favored male students. Table I shows the categorization of the items by both content area and format. Of the 24 content areas only three overlapped between those with items flagged with DIF that favored male students and those with items flagged with DIF that favored female students. Therefore, there were nine unique content areas that featured items flagged with DIF that favored female students, and 12 unique content areas that featured items flagged with DIF that favored male students. The three content areas that featured items flagged with DIF that favored both female students and male students were oxidation-reduction reactions, standard enthalpy of reaction, and electromagnetic radiation. Figure 1 shows how many of the flagged items had formats of visual-spatial (VS), specific chemical knowledge (SCK), reasoning (R), and computation (C). There was an almost even distribution of the formats between the items that were flagged that favored both genders. One interesting thing to note is that in the literature visual-spatial items tended to favor male students. However, it was found that eight items that were flagged with DIF having the format of visual-spatial favored

female students whereas, only six items that were flagged with DIF having the same format favored male students. While this may not be consistent with literature,[1-3] it is consistent for what was found with other trials tests from the ACS-EI[4]. While there was a distribution of content areas by gender for the items that were flagged with DIF, there was no clear pattern to the format of the items. This analysis using the trial tests was not conducted to necessarily gain insight into the possible causes of DIF; instead it was conducted as a starting point for a DIF analysis on introductory chemistry exams. These 33 items that exhibited DIF were used as the originals (if not used on the final standardized test) or cloned and put on differently hourly examinations to determine if the DIF was persistent, suggesting that DIF is real.

4.3 Determining the Persistence of DIF Items

4.3.1 Measures of Proficiency and Item Selection

For an item to be considered as exhibiting persistent DIF, it must not only exhibit statistically significant DIF for both stages of a two-stage DIF analysis, but also exhibit statistically significant DIF using the majority of the applicable relevant measures of proficiency. The relevant measures of proficiency included the total examination score, the ACS-EI Toledo Chemistry Placement Examination[5], the Scale Literacy Skills test, the cumulative examination, the ACS- EI First Term General Chemistry Paired Questions final examination[6], and the ACS- EI General Chemistry Conceptual final examination[7]. If the item did not match these criteria it was not considered to exhibit persistent DIF.

For the Fall 2009 semester items the exhibited DIF, as well as some that did not exhibit DIF from the two trial tests, were taken as the originals (if they were not used on the final examination) or cloned and were put on the hourly examination to determine if the items that originally exhibited DIF still were flagged with DIF and those that did not exhibit DIF still were DIF free. This was conducted to determine if the results from the hourly examinations were consistent with the results found from the two trial examinations. Items that were put on the hourly examinations for the Spring 2010, and Fall 2010 semesters included some items from the trial tests that were not tested in the Fall 2009 semester, as well as a couple of new items that exhibited persistent DIF that arose on the Fall 2009 examinations. Some of the items that exhibited persistent DIF on the Fall 2009 semesters were also cloned, mainly on their content, and placed on these exams as well. For the Fall 2011, and Fall 2012 semesters the items included on the hourly examinations mainly consisted of clones of items that previously exhibited persistent DIF. These were items that were cloned based on their content or format and were used to determine if changes to the content area or format affected the persistence of DIF. For the Spring 2013 semester, the items that were included on the hourly examinations were mainly clones of items that were either not previously cloned (overall), or not cloned by only content or format. If any items exhibited persistent DIF from new items (ones that were not intentionally included) those were put on the following semester's hourly examinations and studied for persistence.

4.3.2 Fall 2009 Semester

The main objective of this semester was to study items that exhibited DIF and ones that did not exhibit DIF from the two trial examinations to determine if the results were consistent. Table II is the persistent DIF chart for Fall 2009. In this semester there were eight items that exhibited persistent DIF. Of those eight items, two items that were flagged with persistent DIF favored female students, and six items that were flagged with persistent DIF favored male students. There were seven unique content areas, with the two items that were flagged with persistent DIF that favored female students in the content area of nomenclature, and the other six items that were flagged with persistent DIF that favored male students having different content areas which included percent composition, heat, crystal structures, limiting reagents, properties of liquids, and chemical reactions. When considering the format of the items, both items that were flagged with persistent DIF that favored female students had the format of specific chemical knowledge, whereas the items that were flagged with persistent DIF that favored male students had the formats of visual-spatial, reasoning, and computation.

4.3.3 Spring 2010 Semester

The main objectives of this semester was to study items that exhibited DIF and ones that did not exhibit DIF from the two trial examinations that were not included on the Fall 2009 hourly examinations, to study new persistent DIF items that arose from the Fall 2009 semester, and to clone items that exhibited persistent DIF by their content areas. Table III is the persistent DIF chart for Spring 2010. In this semester there were

four items that exhibited persistent DIF. Of those four items, three items that were flagged with persistent DIF favored female students, and one item that was flagged with persistent DIF favored male students. There were four unique content areas. The three content areas with items that favored female students were nomenclature, hybridization in molecules, and molecular orbital theory. The item that was flagged with persistent DIF that favored male students was in the content area of greatest/least number of atoms. Considering the format of the items, of the three items that were flagged with persistent DIF that favored female students, two items had the format of specific chemical knowledge and the other item that both the formats of visual-spatial and computation, whereas the item that were flagged with persistent DIF that favored male students had the formats of reasoning and computation.

4.3.4 Fall 2010 Semester

The main objectives of this semester was to study items that exhibited DIF and ones that did not exhibit DIF from the two trial examinations that were not included on the Fall 2009 or Spring 2010 hourly examinations, to study new persistent DIF items that arose from the previous semesters, and to clone items that exhibited persistent DIF by their content areas. Table IV is the persistent DIF chart for Fall 2010. In this semester there were five items that exhibited persistent DIF. Of those five items, two items that were flagged with persistent DIF favored female students, and three items that were flagged with persistent DIF favored male students. There were five unique content areas. The two content areas that included items that were flagged with persistent DIF that

favored female students were quantum numbers and polarity. The three items that was flagged with persistent DIF that favored male students had the content areas of measurement, greatest/least number of atoms, and intermolecular forces. Considering the format of the items, of the two items that were flagged with persistent DIF that favored female students, one item had the formats of specific chemical knowledge and the other item had both the formats of specific chemical knowledge and reasoning, whereas the items that were flagged with persistent DIF that favored male students had the formats of visual-spatial, reasoning, and computation.

4.3.5 Fall 2011 Semester

The main objectives of this semester was to study new persistent DIF items that arose from the previous semesters, and to clone items that exhibited persistent DIF by their content areas and their formats. Table V is the persistent DIF chart for Fall 2011. In this semester there were three items that exhibited persistent DIF. Of those three items, one item that was flagged with persistent DIF favored female students, and two items that were flagged with persistent DIF favored male students. There were three unique content areas. The one content area that included items that were flagged with persistent DIF that favored female students was measurement. The two items that were flagged with persistent DIF that favored male students had the content areas of oxidation-reduction reactions and bond enthalpy. Considering the format of the items, the one item that was flagged with persistent DIF that favored female students had the format of

computation, whereas the items that were flagged with persistent DIF that favored male students had the formats of visual-spatial, specific chemical knowledge, and reasoning.

4.3.6 Fall 2012 Semester

The main objectives of this semester was to study new persistent DIF items that arose from the previous semesters, and to clone items that exhibited persistent DIF by their content areas and their formats. Table VI is the persistent DIF chart for Fall 2012. In this semester there were nine items that exhibited persistent DIF. Of those nine items, five items were flagged with persistent DIF that favored female students, and four items that were flagged with persistent DIF favored male students. There were nine unique content areas. The five content areas that included items that were flagged with persistent DIF that favored female students were isotopes, nomenclature, oxidation-reduction reactions, precipitation, and concentration of solutions. The four items that was flagged with persistent DIF that favored male students had the content areas of measurement, crystal structures, polarity, and nuclear symbols. Considering the format of the items, all of the formats were represented at least once in the items that were flagged with persistent DIF that both favored female students and male students.

4.3.7 Spring 2013 Semester

The main objectives of this semester was to study new persistent DIF items that arose from the previous semesters, and to further study any anomalies that were found in

the data. Table VII is the persistent DIF chart for Spring 2013. In this semester there were four items that exhibited persistent DIF. Of those four items, two items were flagged with persistent DIF favored female students, and two items that were flagged with persistent DIF favored male students. There were four unique content areas. The content areas that included items that were flagged with persistent DIF that favored female students were general stoichiometry and molecular orbital theory. The two items that was flagged with persistent DIF that favored male students had the content areas of periodic trends (atomic radii), and percent composition. Considering the format of the items, the two items that were flagged with persistent DIF that favored female students both had the formats of visual-spatial and computation, whereas the items that were flagged with persistent DIF that favored male students had the formats of specific chemical knowledge, reasoning, and computation.

4.3.8 Classwide Analysis

A more helpful way to determine trends by direction of favor of persistent DIF, the content of these items or the format of the items is to look at the aggregate data from all six semesters. For the class-wide analysis, six semesters of data were analyzed including Fall 2009, Spring 2010, Fall 2010, Fall 2011, Fall 2012, and Spring 2013. There were a total of 1,716 students, which is comprised of the students who took the last hourly examination of the semester. By using the number of participants who took that last hourly examination of the semester ensures that participants were not "double counted". On the 24 hourly examinations there were 687 items total; 33 (5%) of these

items had a significant value using the Mantel-Haenszel statistic exhibiting persistent DIF. Of those 33 items, 15 of the items that were flagged with persistent DIF favored female students and 18 of the items that were flagged with persistent DIF favored male students. Table VIII shows the categorization of the items by both their content area as well as their formats. Of the 22 content areas, only three overlapped between those with items flagged with persistent DIF that favored male students and those with items flagged with persistent DIF that favored female students. Therefore, there were eight unique content areas that featured items flagged with persistent DIF that favored female students, and 11 unique content areas that featured items flagged with persistent DIF that favored male students. The three content areas that featured items flagged with persistent DIF that favored both female students and male students were measurement, oxidation-reduction reactions, and polarity. This suggests that certain content areas could potentially favor one gender versus another gender, because so few content areas overlapped. Some content areas, such as nomenclature that contains only items that exhibited persistent DIF that favored female students had four items in it. That means that for these items they all have something about them that tend to favor female students over male students. The reasons for the DIF findings, however, are not provided through these analyses. The reasons for this could be how the content area is taught, how it is presented in the textbook, how it is tested or even how the student is learning the material. Figure 2 shows how many of the flagged items had formats of visual-spatial (VS), specific chemical knowledge (SCK), reasoning (R), and computation (C). There were eight items that were flagged with persistent DIF that favored male students and had the format of VS, whereas there were only six items that were flagged with persistent DIF

that favored female students and had the format of VS. The formats of visual-spatial, reasoning, and computation all included more items that were flagged with persistent DIF that favored male students than items that were flagged with persistent DIF that favored female students. For the format of reasoning there were 13 items that were flagged with persistent DIF and of those, three favored female students and ten favored male students. For the format of computation there were 16 items that were flagged with persistent DIF and of those, six favored female students and ten favored male students. The format of specific chemical knowledge was the only format that had a greater number of items that were flagged with persistent DIF that favored female students. There were eight items that were flagged with persistent DIF that favored female students and three items that were flagged with persistent DIF that favored male students with the format of specific chemical knowledge. Some interesting things related to the format of the items are starting to arise. These trends suggest that items that were flagged with persistent DIF that favored female students are more likely to have to format of specific chemical knowledge versus the other three formats. For items that were flagged with persistent DIF that favored male students, they are more likely to have a format of reasoning, as well as visual-spatial and computation. However, for the formats of visual-spatial and computation there is less of a gap between how many items there were that had these formats separated by each gender. This part of the analysis was done to determine the persistence of DIF, to suggest that DIF is real. It is not something that arises from different instructors or different samples of students. There are in fact, certain items, that when matched against several relevant measures of proficiency continuously have a statistical significance in favoring one gender versus another. The results of these

analyses must be considered in the next phase of the project: examining these items to first extricate the format from the content (and determine the degree to which each contributes to the DIF) and the reason why this is occurring.

4.3.9 Standardized Examination Analysis

Over the course of the semester the students took three different standardized assessments: the ACS-EI Toledo Chemistry Placement Examination[5], the ACS-EI First Term General Chemistry Paired Questions final examination[6], and the ACS-EI General Chemistry Conceptual final examination[7]. To examine if these standardized examinations also contained persistent DIF items, a two-stage DIF analyses along with analyses using different relevant measures of criteria was conducted. Because these items stay the same, a sample of three semesters (Fall 2009, Spring 2010, and Fall 2010) were used for the analysis with a total of 900 students comprising the sample. The filter for selecting a (random) sample was based on using students who took the final exams rather than the placement exams from these three semesters of testing. On the three standardized examinations there were 140 items total; 19 (14%) of them had a significant value using the Mantel-Haenszel statistic exhibiting persistent DIF. Of those 19 items, two of the items that were flagged with persistent DIF favored female students and 17 of the items that were flagged with persistent DIF favored male students. Table IX shows the categorization of the items by both their content area as well as their formats. Of the 14 content areas, none of them overlap between those with items flagged with persistent DIF that favored male students and those with items flagged with persistent DIF that

favored female students. Accordingly, there was one unique content area that featured items flagged with persistent DIF favoring female students, and 13 unique content areas that featured items flagged with persistent DIF favoring male students. By having no overlap in content areas this suggests again that certain content areas could potentially favor one gender versus another gender. Figure 3 shows how many of the flagged items had formats of visual-spatial (VS), specific chemical knowledge (SCK), reasoning (R), and computation (C). Although there are differences between the formats as seen in the figure, no conclusions can be reached based on gender because of the inequality in the number of items that were flagged with persistent DIF that favored each gender. The ACS-EI does perform a two-stage DIF analysis by gender on all of the trial tested items when the datasets are sufficiently large, but not all items that are flagged due to potential DIF are automatically excluded from the released exam[8]. The ACS-EI is not provided with data needed to determine persistence on the trial tests and so all of the decisions are based on both stages of the two-stage DIF analysis, other statistical measures, and the expertise of the committee as to which items should be included.[4, 8-9] In the case of high DIF (uniform, $p < 0.001$), exam committees who write exams are not permitted to use the items. However, in the case of other items that are flagged with potential DIF, exam committees are sometimes permitted to use items in pairs (if they select one item that favors one gender, then they must also select one item that favors the other gender).

4.4 Determining the Possible Causes of DIF

4.4.1 Clones

While DIF was studied over six semesters, the last three semesters (Fall 2011, Fall 2012, and Spring 2013) were mainly used to test items that were cloned from the persistent DIF items to determine how the change effected the DIF. To make sure that more DIF items were not created from cloning items, items that exhibited persistent DIF, as well as, non-DIF items were cloned. Out all the items given over the six semesters, 170 of them were cloned, either by content, format, or in some cases both. Of those 170 items, 62 (36%) of them were DIF items, either items flagged with persistent DIF from the hourly examinations or items flagged with DIF identified with a two-stage analysis from the trial examinations. (For this section both of these types of DIF items will be classified as persistent DIF so as not to confuse them with the hourly examination items that exhibited DIF for both stages of the two-stage analyses, but wasn't persistent.) Many items were cloned more than once. In fact, 29 out of the 62 (47%) DIF items had more than one clone, and one item specifically had 12 clones over the six semesters. There were also 20 items (12%) out of the 170, which were flagged with DIF, but were not cloned. Those items mostly fell into the category of being present on standardized examinations; the items could not be cloned because of the security of the standardized examinations. The few uncloned DIF items that were not present on the standardized examinations arose from being present on the Spring 2013 examination. These will likely be further examined in later semesters.

Table X presents which content area and format the DIF item and the clone(s) had. Besides the general content areas, specific content areas were assigned to help further differentiate the items. There are 50 sets of clones on the table. Of those 50, eight (16%) of them have more than one item that exhibited DIF. These were in the content areas of measurement, nomenclature, greatest/least number of atoms, limiting reagents, oxidation-reduction reactions, ideal gas equation, molecular orbital theory, and crystal structures.

4.4.2 Common Item Equating

Because these items and their clones were given to the same population but different samples of students, it was important to ensure these samples were equivalent to compare the absolute results of the item analysis. To do this common item equating was performed by equating each hourly exam for each semester to the Fall 2009 exam, using the ACS-EI First Term General Chemistry Paired Questions final examination[6] as the anchor exam. Common item equating will essentially compare the exams to set them to the same difficulty, thereby adjusting the scores received to reflect this difference in difficulty. Table XI shows by what number of points the students total scores for each exam changed. Even though some of the exam total scores changed, this had no effect on the Differential Item Functioning, because the equating for each gender was equal. Now that it was determined that the scores are equal, the chi-squared value as calculated through the Mantel-Haenszel statistic can be compared for each of the DIF items to determine how the cloning affected the item.

6 4.4.3 Set of Cloned Items with the Content Area of Measurement

While all of the cloned items give information into possible causes of DIF, the items that have two or more DIF items per clone are of the most interest, because they lend more information to how cloning by content or format can affect DIF. The first set of clones is in the content area of measurement. There were nine questions in total in this set of clones and two exhibit persistent DIF. In both cases the items flagged with persistent DIF favored male students. Table XII gives the identity of these items as well as significance values for both stages of the two-stage analyses and chi-squared values. Figure 4 (A-I) show each of these items. Figure 4A is the original item from the trial examination; because of security reasons this question was blackened out, but had the format of computation. The following item (Figure 4B) is a content clone where the format was the same, but the content of the item changed. This item was also blackened out due to security reasons. Neither of these questions exhibited DIF. Figures 4C-4F are content and format clones of the original item. The format changed to include a visual spatial item besides the computation. For Figures 4C-4F and 4I the visual spatial item includes a figure, whereas for Figures 4G-4H they contain a table. The two items that exhibited persistent DIF are Figures 4F and 4H. This series of clones makes it difficult to determine if the content area or the format of the item is the reason for both Figure 4F and 4H exhibit persistent DIF and they are both content and format clones. Another way to compare these items is to look at the chi-squared values. Figure 5 shows a comparison of each of these items' chi-squared values. These results are inconclusive for the underlying cause of the DIF or the ability to separate a cause due to format or content. For instance, besides the two items that exhibited persistent DIF, the next closest were the

items in Figure 4B and 4D. One of these items is asking for the density and the other is asking for the volume, one has a visual-spatial item and the other doesn't. For this set of clones the results were inconclusive, so they were used in the next phase of the study.

4.4.4 Set of Cloned Items with the Content Area of Nomenclature

The next set of clones is in the content area of nomenclature. There were 12 questions in total in this set of clones and four of them exhibit persistent DIF. In all cases the items flagged with persistent DIF favored female students. Table XIII gives the identity of these items as well as significance values for both stages of the two-stage analyses and chi-squared values. Figure 6 (A-L) show each of these items. Figure 6A is the original item from the trial examination with the format of specific chemical knowledge that did show persistent DIF. The following item (Figure 6B) is a retest of the same item. Both of these items exhibited DIF. Figures 6C-6D are the same item and are content clones of the original item. Interestingly Figure 6C exhibited persistent DIF, but Figure 6D did not. Figures 6E-6G are content clones of the original item that contain a multivalent cation with a positive two or positive three charge, with Figure 6F exhibiting persistent DIF. Out of the 12 clones, seven of the items contain a multivalent cation with a charge of positive two or positive three. Out of those seven items, four (57%) of them exhibit persistent DIF. Interestingly, if the item did not have a multivalent cation that had a charge of positive two or positive three, it did not exhibit DIF. While it would be irresponsible to say that a nomenclature item that features a multivalent cation with a positive two or positive three charge combining with a polyatomic anion with a negative

two or a negative three charge and no visual-spatial component is going to exhibit DIF favoring female students, based on these results the probability is higher that it may. This is not always the case as seen in Figures 6C and 6D which are the exact same item and follow the trend from the previous statement and yet one items exhibits persistent DIF and the other does not. Figure 6H is a content clone of the original item that contains a multivalent cation of a positive four charge. Figures 6I-6K are content clones of the original item but do not contain a multivalent cation. Figure 6L is a format clone of the original item adding a visual-spatial item.

Looking at the comparison in chi-squared values (Figure 7) it can be seen that most of the chi-squared values are close to what would be considered to exhibit DIF for the first stage of the two-stage analyses (above 3.8), except for the items in Figure 6D and 6L. These items could have small chi-squared values because the students may have become familiar with this type of problem by having seen these types of items on different exams and homework assignments. Also Figure 6L is the only item in this set of clones that also includes a visual-spatial component, making this be an item with a content that typically favors female students and a format that has be shown to typically favor male students, therefore "cancelling" out the gender DIF effect.

4.4.5 Set of Cloned Items with the Content Area of Greatest/Least Number of Atoms

The next set of clones is in the content area of greatest/least number of atoms. There are ten questions in total in this set of clones and three of them exhibit persistent

DIF. In all cases the items flagged with persistent DIF favored male students. Table XIV gives the identity of these items as well as the significance values for both stages of the two-stage analyses and chi-squared values. Figure 8 (A-J) show each of these items. Figure 8A is the original item from the trial examination with the format of computation that did show persistent DIF. Figures 8B and 8C are content clones of the original item. Neither of these items exhibited DIF. Figures 8D-8H are content clones and format clones of the original item. The format that was added was reasoning. Two of these items (Figures 8E and 8G) exhibited persistent DIF and were content clones of one another. The results are inconclusive as to if it was the content or format that caused DIF. Figures 8I and 8J were also content and format clones of the original item, but this time the format was changed to visual-spatial and reasoning with no computation. Neither of these items exhibited persistent DIF. By adding the visual-spatial component, one would expect the item to have a stronger tendency for DIF towards male students, but that didn't seem to be the case suggesting that format has a lower contribution to persistent DIF compared to content. Looking at the comparison in chi-squared values (Figure 9) it can be seen that the chi-squared values are varied with one being very large and another being quite small. Additionally, the item in Figure 8I exhibited DIF for both stages of the two-stage analyses, but did not prove to be persistent. In fact, the only external relevant measure of proficiency that produced a significant value was the ACS-EI Toledo Chemistry Placement Examination[5] score. There is also something unique about this set of clones. There are two different times that two cloned items were on the same hourly examination. Here the first item (Figure 8D) does not exhibit persistent DIF, but the second item (Figure 8E) one the same item does. They both have the same format

with a small variation in content. The second occurrence was for Figures 8H and 8J; here both of the items do not exhibit DIF and have different formats. For this set of clones the results were inconclusive, so they were used in the next phase of the study.

4.4.6 Set of Cloned Items with the Content Area of Limiting Reagents

The next set of clones is in the content area of limiting reagents. There are five questions in total in this set of clones and two of them exhibit persistent DIF. In all cases the items flagged with persistent DIF favored male students. Table XV gives the identity of these items as well as the significance values for both stages of the two-stage analyses and chi-squared values. Figure 10 (A-E) show each of these items. Figure 10A is the original item from the trial examination with the format of visual-spatial and computation that exhibited persistent DIF. The following item (Figure 10B) is the same as the original item and was a retest of that item. Interestingly, Figure 10B did not exhibit persistent DIF. The chi-squared value is quite large for the item in Figure 10B, but the item does not exhibit persistent DIF (Figure 11). Figure 10B exhibited DIF for both stages of the two-stage analyses, as well as is flagged with DIF against the external criteria of the ACS-EI Toledo Chemistry Placement Examination[5] and the ACS-EI First Term General Chemistry Paired Questions final examination[6]. Figure 10C – 10E are content clones where the format was the same as the original, but the content of the item changed. Figure 10D exhibited persistent DIF. Because of how similar each of these clones are it is difficult to determine in the causes of DIF are determined by changes in the content or

format. For this set of clones the results were inconclusive, so they were used in the next phase of the study.

4.4.7 Set of Cloned Items with the Content Area of Oxidation-Reduction Reactions

The next set of clones is in the content area of oxidation-reduction reactions. There are seven questions in total in this set of clones and two of them exhibit persistent DIF. One item that was flagged with persistent DIF favored male students and the other item that was flagged with persistent DIF favored female students. Table XVI gives the identity of these items as well as the significance values for both stages of the two-stage analyses and chi-squared values. Figure 12 (A-G) shows each of these items. Figure 12A is the original item from the trial examination with the formats of visual-spatial and computation that did show persistent DIF that favored male students. Figures 12B is the same as the original item and was retested to see if the results were similar with what was found for the trial examination. Figure 12B, however, did not exhibit persistent DIF. Figures 12C-12F are content clones of the original item. None of these items exhibited DIF. Figures 12C-12E are all the same item. Looking at their chi-squared (Figure 13) values even though these two sets have the same items within the set they test differently. The item from Figure 12B is close to being significant for DIF for the first stage of the two-stage analyses, and is significant for the second stage of the two-stage analyses, but it is does not exhibit persistent DIF. Also the significance and chi-squared values (Table XVI) are very similar for the items in Figures 12D and 12E, but are quite different for the

item in Figure 12C. It looks that there is something particular about the reaction featured in Figures 12A and 12B that make it more likely to exhibit DIF. However, Figure 12G which is a content and format clone of the original item with the format changing to visual-spatial and specific chemical knowledge does exhibit persistent DIF favoring female students and includes the same reaction found in Figures 12C-12E. From the analysis of the hourly examination, it was shown that of the items that exhibited DIF and had the format of specific chemical knowledge tended to favor female students. This information combined with the fact that when the format was kept the same and the content changed the item no longer exhibited persistent DIF suggests that the cause of DIF for this item is the format.

4.4.8 Set of Cloned Items with the Content Area of the Ideal Gas Equation

The next set of clones is in the content area of the ideal gas equation. There are two questions in total in this set of clones and both of them exhibit persistent DIF that favored male students. Because both of these questions came from Standardized Examinations these items were not cloned and these items cannot be released. Table XVII lists the significance values for both stages of the two-stage analyses and chi-squared values. Both of these items have a specific content area of looking at the density using the Ideal Gas Equation. The original question had the format of specific chemical knowledge and computation, whereas the clone has the format of specific chemical knowledge and reasoning. From what was observed on the hourly examination analysis, the format of specific chemical knowledge typically favored female students. This

suggests, along with the fact that the format changed and the DIF was still persistent, that the cause of DIF for these items has more to do with the content than with the format.

4.4.9 Set of Cloned Items with the Content Area of Molecular Orbital Theory

The next set of clones is in the content area of molecular orbital theory. There are five questions in total in this set of clones and two of them exhibited persistent DIF. Both of the items that exhibited persistent DIF favored female students. Table XVIII gives the identity of these items as well as the significance values for both stages of the two-stage analyses and chi-squared values. Figure 14 (A-E) shows each of these items. Figure 14A is the original item from the trial examination that had a format of specific chemical knowledge and exhibited persistent DIF that favored female students. Figures 14B and 14C are the same item as the original that was retested to see how the results would compare. Figure 14C exhibited persistent DIF that favored female students, but Figure 14B did not. Because the items in Figures 14A and 14C are all the same item and two of them exhibit persistent DIF, it suggests there is something about this particular item that exhibits DIF. Figure 14D was a content clone of the original item and Figure 14E was a format clone (adding the format of visual-spatial) of the original item. Neither of these clones exhibited persistent DIF.

4.4.10 Set of Cloned Items with the Content Area of Crystal Structures

The last set of clones is in the content area of crystal structures. There are five questions in total in this set of clones and two of them exhibited persistent DIF. Both of the items that exhibited persistent DIF favored male students. Table XIX gives the identity of these items as well as the significance values for both stages of the two-stage analyses and chi-squared values. Figure 16 (A-E) shows each of these items. Figure 16A is the original item with the format of visual-spatial and reasoning that exhibited persistent DIF. The following items (Figure 16B and 16C) are the same as the original item and were a retest of that item. Neither of these items exhibited persistent DIF. The chi-squared value is quite small for these items compared to Figure 16A (Figure 17). Figure 16D is testing the same content as the original item and had the same format class but the specific aspect of the format changed. In the original item there was a pictorial representation of both the salt and SiC crystals. This was not present in Figure 16D. Figure 16E is a format clone of the original item, changing the format to only reasoning. This item did exhibit persistent DIF. This suggests that the DIF is not caused by the format of the item; instead it may be caused by the content. However, one cannot conclude that to always be the case, because of the first three items being the exact same and only one of them exhibiting persistent DIF.

4.4 Eye Tracking Interviews

4.4.1 Introduction

The cloning of the items based on the content and/or format led to some suggestions as to the causes of DIF, but some of the results were inconclusive. Those results, along with the fact, that it is still unknown why DIF is happening, supported moving to the final phase of the project of capturing the students' problem-solving process. It is hypothesized that some of the DIF could be caused by one subgroup using an incorrect heuristic more commonly than the other.

4.4.2 Assessment-like Interviews

The first set of interviews included having the students solve 76 general chemistry questions in an assessment format. The students performed this interview on an eye-tracking instrument as if they were doing online homework. There were 62 undergraduate students (finishing the semester in either general chemistry I or general chemistry II) and 13 graduate students. Because of the size of the sample, there are too few participants to be able to do a DIF analysis, but some other statistical results were examined. First considered was the overall performance by the participants and by gender or expertise subgrouping. Figure 18 shows the performance of the participants separated by undergraduate versus graduate. There was a significant effect for gender, $t(75) = 3.880$, $p<0.050$. To determine if there was a gender biased an independent *t*-test was performed separating out the participants by gender (Figure 19). There was no

statistically significant difference between the genders in each expertise group (graduate vs undergraduate), but there was a significant effect for male graduate and undergraduate students, $t(35) = 3.141$, $p<0.050$. However, there was not a statistical difference between the female graduate students and the female undergraduate students' scores. This could be because there was a much smaller sample of females graduate students than females undergraduate students, and for some reason the variance in the performance of the female graduate students is quite large (which again could be because there was such a small sample). There is more information to be gained by the eye-tracking instrument, but upon analysis it became obvious that what was really needed to determine how the students were solving the problems was to have the students think-aloud and describe their problem-solving process in connection with the information collected by the eye-tracking instrument. The instrument can give information about where they were looking, but it is ultimately what the student was thinking about as they were looking at the problem or an area of the problem that was wanted.

4.4.3 Semi-Structured Interviews

4.4.3.1 Introduction

As a way to combine both the information gathered from the eye-tracking instrument as well as gain information about what the students are thinking about as they solve the items, two sets of semi-structured interviews were performed. The participants of these interviews were only students who were completing general chemistry I, and they were conducted in the final weeks of the semester. One can look at the tradition

measures using an eye tracking instrument like performance, time on task, and a self-reported mental effort. Besides studying these measures, it was of interest to use them as a filter to help determine which students were using an incorrect heuristic to solve a task. This section will look at both the tradition measures, as well as the investigation of heuristics to help determine possible reasons of why DIF is happening. Also there were over twice the amount of female students who participated in the interview versus male students, so there was not a representative sample. Therefore, the following results for the independent *t*-test are reported only as a measure of what may be happening. It would be irresponsible to use any of the statistical evidence from the independent *t*-test to make conclusions about the data. Because of the similarity in the items on the interview to items on standardized assessment, the items will not be shown, but they will be described according to their content and format.

4.4.3.2 Performance, Time on Task, and Mental Effort

Figure 20 compares the overall performance for the semi-structured interview. Table XX gives the item performance by gender and Table XXI shows the overall performance, overall time on task, and overall mental effort by gender. There is no significant difference between the performances of the male students on the interview compared to that of the female students. Because the interview has both questions that exhibited persistent DIF and those that didn't (Table XXII), the performance per item was also analyzed to determine if there was any statistical significance. Out of the 20 items, there were two items that had a statistically significant difference between the male

students' performance and the female students' performance. The two items that had a statistically significant effect for gender were item two $t(49) = 2.098$, $p<0.050$ and item 14 $t(48) = 2.798$, $p<0.050$. Item two had a content area of measurement (density) and a format of visual-spatial and reasoning. This item did not previously show any DIF, though a clone of the item did. For this interview all 15 male participant answered the item correctly whereas only 30 out of the 34 (88%) female students answered this item correctly. Item 14 had a content area of quantum numbers and a format of specific chemical knowledge. This was an item that had previously exhibited DIF that favored female students. For this interview six out of the 15 (40%) male students answered this item correctly, whereas 26 out of the 33 (79%) female students answered this item correctly. Because of the sample in the study, any conclusions based solely on statistical analysis of item performance would be premature and irresponsible. More important though, the performance is the first window into finding students who may be processing the item differently. It is through this lens that these items will be further examined for an explanation of the observed DIF from the classwide analysis. Regardless, replicating results from a classwide analysis with a laboratory experiment with a smaller sample for even one item is encouraging.

Figure 21 compares the overall average time on task for the semi-structure interviews. There is no significant difference between the average times it took for the male students on each item on the interview compared to that of the female students. While this gives us information about overall how long it took the students to answer the items, it is of more interest to study the time on task for each item. For the 20 items in the interview only one item (item 16) had a statistically significant difference between the

time on task for the male students compared to that of the female students, $t(46) = 2.158$, $p<0.050$. The average time on task for male students was 3.07 min, whereas the average time on task for female students was 2.08 min. This item did not previously exhibit DIF and had a content area of bond angles and a format of visual-spatial and reasoning. For this item the male students took significantly longer on this task than the female students. Out of the 15 male students, eight students (53%) got this item correct, whereas out of the 32 female students only 12 students (38%) answered this item correctly. The lower time on task for female students along with the lower performance would suggest that an incorrect heuristic was used. At this point the students' problem solving process was then examined to determine if an incorrect heuristic was used or not.

Figure 22 compares the overall average mental effort for the semi-structure interviews. There is no significant difference between the average mental effort of the male students on the interview compared to that of the female students. Both of these numbers are just below three indicating that on average the students self-reported mental effort was between low amount (two) and moderate amounts (three). There was also no significant difference between the average mental effort of the male students compare to that of the female students per item either.

This gives us information about the interview and the participants in general, but to determine why DIF is happening the students' problem solving processes as reported by the students as they solve the tasks needed to be investigated.

4.4.3.3 Filter for Heuristic Determination

A heuristic is a fast, automated process that people use when making decisions.[10] It is sometimes known as a rule of thumb and is performed usually unconsciously.[11] It is hypothesized that one of the reasons as to why DIF is happening is that a subgroup of students is using an incorrect heuristic to solve certain problems. The interviews were examined for the usage of four incorrect heuristics: recognition, representativeness, one-reason decision making, and arbitrary trends. Later in this chapter will be examples of how students used these heuristics to incorrectly solve the item. An example of using one of these heuristics correctly could be a person looking a molecular structure of H_2O and being able to say it is water. They could use recognition in that this molecular structure is very common and it is likely they have seen it before. As a way to determine when an incorrect heuristic was used a filter system was set up, using objective and subjective measures of cognitive load, which have been shown to be related to the type of processing used.[12] The filter consisted first of performance. Only students who got the questions wrong were considered for the next step. Following this, mental effort was considered; only students who answered that they had moderate amounts (three) of mental effort or lower were included. The next step consisted of time on task, if the participant had a lower time on task then the average time on task for the students who got the item correct, they were analyzed for consideration of an incorrect heuristic. The last filter was initially pupil diameter, however this measure was found not to be the most stable indicator for a filter. An exercise was performed at the beginning of each interview that had been previously shown to maximize the participants' pupil diameter.[13] Therefore, the filter for the pupil diameter was set so that if the pupil diameter was less

than the maximum pupil diameter for the pupil diameter exercise (which should have been all of them) they would be considered for use of an incorrect heuristic. However, after more investigation it was shown that many times the students had larger pupil diameters when they were working the chemistry items on the interview than they did in the maxing out the pupil diameter exercise. Additionally, the pupil diameter filter was integrated into the interview protocol for the second group of participants only. The following will present the heuristic analysis without using the pupil diameter filter, as well as some preliminary results with using it.

4.4.3.4 Heuristic Analysis

Table XXIII shows the amount of participants whom were likely to have a heuristic according to the filter (not using pupil diameter) compared to that of how many of the flagged participants actually did use a heuristic. The filter was correct 70% of the time, which while it is not perfect it is useful to reduce the number of transcripts for analysis. If the pupil diameter would have been added as a fourth stage to the filter, 27 participants who had heuristics would have been missed (Table XXIV). Concluding from our results and the way the pupil diameter measure was used, either it is not a good measure for a heuristic or it needs to be used in a different way in the filter to determine incorrect heuristics.

Using the filter without the pupil diameter there were 150 times that at least one heuristic was used with a total 189 times that heuristics used overall. Of the 189 times heuristics were used, 184 (97%) of them were identified by reading the transcripts. The

remaining five instances were all on item nine, which was an item that exhibited DIF that favored male students and had the content area of stoichiometry-general and a format of visual-spatial, reasoning, and computation. For this item, it was difficult to discern on five of the transcripts if the students were using the heuristic of recognition or representativeness. More information would be needed to confidently classify those into one heuristic or another. As for the heuristics that were used, recognition was used 14.9% of the times, representativeness was used 35.9% of the time, one-reason decision making was used 47.5% of the time, and arbitrary trends was used 1.7% of the time. Table (XXV) shows the percentage that each heuristic was used per item on the interview. This information was not counting the other heuristics from item nine that were not able to be classified as recognition or representation. The amount of time each gender used a heuristic per item is represented in Table XXVI.

For item one, the most common heuristic that was used was one-reason decision making, with only one person using a combination of one-reason decision making and representativeness and the remaining students using one-reason decision making alone. Item one was a question that previously exhibited DIF that favored male students. There were nine male students who were flagged as using a heuristic with evidence that seven did use at least one heuristic. There were 23 female students who were flagged as using a heuristic with evidence that 18 used one or more heuristics. There were variance in the process that students used to try to solve the density task, but the most common heuristic was using mass as the only reason an object would sink, neglecting to consider volume or density. Repeatedly the students focused solely on the mass of the objects and made their decision for their response accordingly. As an example of these conclusions, below are

two different portions of two different transcripts for this item exhibiting the use of this heuristic:

> "at first I was looking for size, and then I realized that size of the container doesn't matter but it was the mass that mattered"
>
> "I used my Sherlock Holmesian inference of logic (laughs), … but if its heavier I would assume that it would sink simply because it is heavier"

These are representative statements as examples of how the students based their response using the reason that the mass determines whether an object will sink or float regardless of the size or density of the object.

For item three, the most common heuristic used was recognition, with one-reason decision making being used about 31% of the time. Both of these heuristics were used separately by all students and not in conjunction with another heuristic for this item. Item three is a density question that had previously exhibited DIF that favored male students that had the content area of measurement (density) and the format of visual-spatial and computation. There were 4 male students who were flagged with using a heuristic to solve the item and all 4 of them used the heuristic of recognition. There were nine female students who were flagged with using a heuristic to solve the item and four of them used one-reason decision making and five of them used recognition. For this item it is possible that the difference may have arose from the different types of heuristics that were used because only the female participants used the heuristic of one-reason decision making. For the students who used recognition most saw the values (more importantly, focusing on the units) and recognized they would give a response using the density formula, without realizing that more information was needed from the figure to be able to

obtain the correct response. One example of a student who used the heuristic of recognition is:

> "umm, in the equation there's given the mass of the sample and then it gives you the density. It just asks for the volume so automatically I thought of the density equation and I just plugged in, err, rearranged the equation and plugged the numbers in"

From the above excerpt, one can tell that the student is just taking the numbers from the problem, recognizing they go into a formula they know and using it (which was also confirmed by the response they selected and the scan path map from the eye-tracker that shows them not fixating on the figure). Additionally, they are not thinking about if their answer makes sense or any other information that is given in the item, also supporting the lack of analytical processing. Considering the other type of heuristic used on this task, below is an example of a student who used one-reason decision making:

> "so we know the initial volume from the picture, which is 29 mL. Now we need the final volume, the volume probably be density times volume [and] will give us the mass. So then the volume will be mass over density…(the student is asked what they used to help them get to their answer) ok I first used, umm, I was thinking about the initial and final volume, but then I did not actually use that….it wasn't really helpful. The only thing I should have think about is the density formula, which is mass over volume and then after that I will get volume as mass over density and then just put the calculation"

This student started to realize they needed to use the figure, but as soon as they didn't know how to use it, they focused on the other information in the problem and decided not to use the information from the figure. They even state that the figure wasn't helpful and they just needed the numbers and to put them into the calculation, thus supporting the conclusion that this student used one-reason decision making.

For item seven, the most common heuristic used was one-reason decision making, with arbitrary trends being used about 17% of the time, and representativeness 8% of the time. Item seven is a greatest/least number of atoms question that had previously exhibited DIF that favored male students and had the format of reasoning and computation. There were only two male students who were flagged with using a heuristic to solve the item and both of them used the heuristic of one-reason decision making. There were nine female students who were flagged with using a heuristic to solve the item and eight of them did use a heuristic. Of the eight, five of them used only one-reason decision making, two of them used both one-reason decision making and arbitrary trends, and one used representativeness only. The heuristics of this item suggest that the reason for the persistent DIF is because the female students are using an incorrect heuristic. In fact, out of the heuristics used for this item the female students used a heuristic approximately five to one compared to the male students. This item did exhibited persistent DIF that favored male students and one possible reason for the DIF is that the female students are using an incorrect heuristic. Interestingly, this is the first time the heuristic of arbitrary trends has been used. Here is an example of how it was used:

> "I'm…gonna say the .5 grams (clicks on answer) cuz it has the lowest number of grams, and, ahh yep, since all the ones have higher gram substance they probably have greater amount of atoms, and I'm just getting a feeling because the sodium (laughs) is on the far left side of the periodic table compared to the other ones"

In this instance the student is using both one-reason decision making, by picking the 0.5 grams and basing their answer off that choice as well as using arbitrary trends by looking for a trend on the periodic table to support their answer without any justification.

For item 14, recognition was the only heuristic used. Item 14 is a quantum number question that had previously exhibited DIF that favored female students and had the format of specific chemical knowledge. There were only two male students who were flagged with using a heuristic to solve the item. The heuristics of this item suggest that the reason for the persistent DIF is because the male students are using an incorrect heuristic; this is also one of the items that had statistically significant difference between the performance of the male students and the female students. This item did exhibit persistent DIF that favored female students and one possible reason for the DIF is that the male students are using an incorrect heuristic to solve the item. Here is an example of a student who used recognition:

> "I believe that's n and that's, I guess I'm, uhh, that's what I had stated in the last problem myself again that's my reason…(the student is asked what they used to help them solve the item) Again, the memory of quantum numbers are"

The student recognized that n was a quantum number and based their decision off of that.

Items 16 and 18 are of interest because of the large amount of times heuristics were used. Together these two items make up 34% of the times heuristics were used for the entire 20 items in the interview. For item 16, 30 heuristics were used, which accounts for 17% of all the heuristics used. For item 18, 32 (18%) heuristics were used. Also, the heuristic of representativeness was used 47 (72%) times in these two items. Item 16 is a question about bond angles with the format of visual-spatial and reasoning. This question previously did not exhibit DIF. Primarily, the students looked at the figure in 2D space and answered based on this representation or from their knowledge of "geometry." They did not consider how the structure would look in 3D space or how the angles could be

more correctly represented rather than as they were in the figure. Here are excerpts from the interviews:

> "ohh this is going back to geometry knowing that a 90 degree angle umm makes a corner and then 180 degree angle makes a straight line so its linear"
>
> "a full circle is 360 so if it's a straight line that's half of a circle so half of 360 is 180 that's for B, and then umm uhh full like right corner which is like a fourth of a circle 390 divided by four is 90, and I just used math skills not really chemistry"

Here, the students are interpreting the shape of a molecule that a tetrahedral carbon as 90° because it has four atoms attached and a carbonyl group as linear because that is how it is depicted in the figure.

Item 18 had a content area of hybridization in molecules and formats of visual-spatial and computation. It was an item that had previously exhibited DIF that favored female students. If the hypothesis of one of the potential reasons why DIF happens is because one subgroup is using an incorrect heuristic, this item would not support the hypothesis and another explanation for the reason for DIF would need to be investigated. There were nine male students who used an incorrect heuristic on this item and 23 female students. For this item the reason behind DIF is not from the male students using an incorrect heuristic compared to the female students, but instead there is another unknown factor that is affecting the DIF.

One way to check if the results of this interview support the hypothesis of the reasons why DIF is happening is because of one subgroup using an incorrect heuristic is to look at which items had previous exhibited DIF and the percentage of each subgroup

that used an incorrect heuristic. This is shown in Table XXVII. If this hypothesis were correct any item that had persistent DIF that favored female students should have a larger amount of male students who used an incorrect heuristic, and any item that had persistent DIF that favored male students should have a larger amount of female students who used an incorrect heuristic. On items 1, 3, 7, 8, 9, 14, and 19 the data supports the hypothesis with the participants of the unfavored subgroup using an incorrect heuristic in a greater percentage. However, on items 5, 6, 15, 17, and 18 the data refutes the hypothesis by either showing a lack of heuristics used overall or incorrect heuristics used by the subgroup that was previously showing favor on the items by DIF analysis. For these items more analysis would have to be done to determine the causes and reasons for DIF.

Overall, items with DIF have been discovered by using the two trial tests from the ACS-EI. By doing further studies and using external relevant measures of proficiency, it has been show that DIF is real. There are certain items that are persistent, meaning there is something about those items that are favoring one subgroup versus another regardless of the sample tested or the measure of proficiency used as criteria. Using clones of the items, it has been determined that some of the causes of persistent DIF are based on content or format of the items, but separating the cause of DIF by content or format was not possible for all items. Indeed, it is more likely that a synergistic relationship between content and format exists. Also through semi-structured interviews using an eye-tracking instrument it has been shown that the use of incorrect heuristics can be used to explain the reason why DIF occurs for some of the items. This can then be used twofold: first to propose and test interventions (either practice of teaching and learning or testing) to offset the DIF or to continue investigating why DIF occurs for the remaining items.

Table I: Items that were flagged with DIF on the two trial tests from the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS DivCHED EI).

Number of Items	Content Area	Favors	Format[†]
2	Nomenclature	Female	SCK
1	Stoichiometry (general)	Female	VS, C
1	Oxidation-Reduction Reactions	Female	SCK
2	Gas Laws	Female	VS, C
1	Equal Gas Equation	Female	VS, C
2	Standard Enthalpy of Reaction	Female	VS, SCK, C
1	Bohr's Theory	Female	R
1	Electromagnetic Radiation	Female	VS, R
1	Electron Configuration	Female	SCK, R
1	Electron Configuration of Ions	Female	C
1	Formal Charge	Female	VS, C
1	Molecular Orbital Theory	Female	SCK
1	Classification of Matter	Male	VS,SCK
2	Measurement	Male	VS, SCK
1	Greatest/Least Number of Atoms	Male	C
1	Limiting Reagents	Male	VS, C
1	Dilutions	Male	C
1	General Properties of Aq Solns	Male	SCK
2	Oxidation-Reduction Reactions	Male	VS, SCK, C
1	Precipitation	Male	SCK
1	Heat	Male	SCK, R
1	Standard Enthalpy of Reaction	Male	VS, C
1	Electromagnetic Radiation	Male	C
1	Polarity	Male	VS, SCK
1	Empirical Formula-% Comp	Male	R, C
1	Ideal Gas Equation	Male	SCK, C
2	Phase Changes	Male	SCK, R

[†]Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table II: Persistent DIF chart for the Fall 2009 semester.

Content	Favors	# of Items	Format†	Internal Measures of Proficiency	External Measure of Proficiency					
					Total Exam Score	Placement Exam	Scale Exam	Cumulative Exam	Final Part I	Final Part II
Nomenclature	Female	2	CK	x	x*	x	x	x*	x	x
% Composition	Male	1	VS, C	x	x	x	x	x	x	x
Heat	Male	1	C	x		x	x	x		x
Crystal Structures	Male	1	VS, R	x	x	x	x	x	x	
Limiting Reagents	Male	1	VS, C	x	N/A	x	x	N/A	x	x
Properties of Liquids	Male	1	R	x	N/A	x	x	N/A	x	x
Chemical Reactions	Male	1	VS, R	x	x	x	x	x	x	

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

*Represents when there are two items in a content area that favor one gender and flagged with DIF when the external measure of proficiency is applicable.

Table III: Persistent DIF chart for the Spring 2010 semester.

Content	Favors	# of Items	Format	Internal Measures of Proficiency	External Measure of Proficiency					
					Total Exam Score	Placement Exam	Scale Exam	Cumulative Exam	Final Part I	Final Part II
Nomenclature	Female	1	CK	x	x	x	x	x	x	x
Hybridization in Molecules	Female	1	VS, C	x	N/A	x	x	N/A		x
Molecular Orbital Theory	Female	1	CK	x	N/A	x		N/A	x	x
Greatest/Least Number of Atoms	Male	1	R, C	x	x	x	x	x	x	x

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table IV: Persistent DIF chart for the Fall 2010 semester.

Content	Favors	# of Items	Format†	Internal Measures of Proficiency	External Measure of Proficiency					
					Total Exam Score	Placement Exam	Scale Exam	Cumulative Exam	Final Part I	Final Part II
Quantum Numbers	Female	1	CK	x	x	x	x	x	x	x
Polarity	Female	1	CK, R	x	N/A	x	x	N/A	x	x
Measurement	Male	1	VS, C	x	N/A	x		N/A	x	x
Greatest/Least Number of Atoms	Male	1	R, C	x	N/A			N/A	x	x
Intermolecular Forces	Male	1	R	x	N/A	x		N/A	x	x

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table V: Persistent DIF chart for the Fall 2011 semester.

Content	Favors	# of Items	Format	Internal Measures of Proficiency	External Measure of Proficiency					
					Total Exam Score	Placement Exam	Scale Exam	Cumulative Exam	Final Part I	Final Part II
Measurement	Female	1	C	x	x			x	x	x
Oxidation-Reduction Reactions	Male	1	VS, R	x	x	x	x		x	x
Bond Enthalpy	Male	1	CK, R	x	N/A	x		N/A		x

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table VI: Persistent DIF chart for the Fall 2012 semester.

Content	Favors	# of Items	Format†	Internal Measures of Proficiency	External Measure of Proficiency					
					Total Exam Score	Placement Exam	Scale Exam	Cumulative Exam	Final Part I	Final Part II
Isotopes	Female	1	VS, R	x	x	x	x	x	x	x
Nomenclature	Female	1	CK, R	x	x	x	x		x	
Oxidation-Reduction Reactions	Female	1	VS, CK	x	x	x		x	x	x
Precipitation	Female	1	VS, C	x	N/A	x	x	N/A	x	x
Concentration of Solutions	Female	1	C	x	N/A		x	N/A	x	x
Measurement	Male	1	VS, C	x	N/A	x	x	N/A	x	x
Crystal Structures	Male	1	R	x	N/A	x		N/A	x	x
Polarity	Male	1	CK, C	x	N/A	x	x	N/A		
Atomic Number, Mass #, Etc.	Male	1	VS, C	x	N/A	x	x	N/A	x	x

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table VII: Persistent DIF chart for the Spring 2013 semester.

Content	Favors	# of Items	Format†	Internal Measures of Proficiency	External Measure of Proficiency					
					Total Exam Score	Placement Exam	Scale Exam	Cumulative Exam	Final Part I	Final Part II
Stoichiometry-General	Female	1	VS, C	x	N/A	x	x	N/A	x	x
Molecular Orbital Theory	Female	1	VS, C	x	N/A		x	N/A	x	x
Atomic Radius	Male	1	CK, R	x	x	x		x	x	
% Composition	Male	1	C	x	N/A	x		N/A	x	x

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table VIII: Items that were flagged with DIF on the 24 hourly examinations given over six semesters.

Number of Items	Content Area	Favors	Format†
1	Measurement	Female	C
1	Isotopes	Female	VS, R
4	Nomenclature	Female	CK, R
1	Stoichiometry-General	Female	VS, C
1	Concentration of Solutions	Female	C
1	Oxidation-Reduction Reactions	Female	VS, CK
1	Precipitation	Female	VS, C
1	Quantum Numbers	Female	CK
1	Hybridization in Molecules	Female	VS, C
2	Molecular Orbital Theory	Female	VS, CK, C
1	Polarity	Female	CK, R
2	Measurement	Male	VS, C
2	% Composition	Male	VS, C
2	Greatest/Least Number of Atoms	Male	R, C
1	Limiting Reagents	Male	VS, C
1	Oxidation-Reduction Reactions	Male	VS, R
1	Heat	Male	C
1	Atomic Radius	Male	CK, R
1	Polarity	Male	CK, R
2	Crystal Structures	Male	VS, R
1	Properties of Liquids	Male	R
1	Atomic Number, Mass #, Etc.	Male	VS, C
1	Chemical Reactions	Male	VS, R
1	Bond Enthalpy	Male	CK, R
1	Intermolecular Forces	Male	R

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table IX: Items that were flagged with persistent DIF on the three standardized examinations given over three semesters.

Number of Items	Content Area	Favors	Format†
2	Chemical Reactions	Female	VS,C
1	Classification of Matter	Male	SCK
2	Measurement	Male	VS, SCK, R
1	Empirical Formula	Male	R
1	Limiting Reagents	Male	VS, R
1	General Properties of Aq Solns	Male	VS, SCK
1	Oxidation-Reduction Reactions	Male	SCK
1	Ideal Gas Equation	Male	SCK, R
2	Kinetic Molecular Theory	Male	SCK, R
1	KMT of Liquids and Solids	Male	SCK
2	Properties of Liquids	Male	SCK, R
1	Stoichiometry-General	Male	VS, R, C
2	Kinetic Molecular Theory	Male	SCK, R
1	Heat	Male	VS, C

†Formats of the items include Visual-Spatial (VS), Specific Chemical Knowledge (SCK), Reasoning (R), and Computation (C).

Table X: The DIF items and their clones.

Item	DIF	Content (general)	Content (specific)	Construct
Original	x	Classification of Matter	Classification of Element	SCK
Clone		Classification of Matter	Classification of Element	SCK
Clone		Classification of Matter	Classification of Element	VS, SCK
Original	x	Classification of Matter	Classification of Compound	VS, SCK
Clone		Classification of Matter	Classification of Compound	VS, SCK
Clone		Classification of Matter	Classification of Compound	VS, SCK
Clone		Classification of Matter	Classification of Compound	VS, SCK
Original	x	Measurement	Temp Conversion	C
Clone		Measurement	Temp Conversion	VS, R
Original	x	Measurement	Vol Grad Cylinder	VS, SCK
Clone		Measurement	Vol Grad Cylinder	VS, SCK
Original	x	Measurement	Precision/Accuracy	SCK
Clone		Measurement	Precision/Accuracy	VS, SCK
Original		Measurement	Density	C
Clone		Measurement	Density	VS, C
Clone		Measurement	Density	VS, C
Clone		Measurement	Density	VS, C
Clone	x	Measurement	Density	VS, C
Clone		Measurement	Density	C
Clone		Measurement	Density	VS, C
Clone	x	Measurement	Density	VS, C
Clone		Measurement	Density	VS, C
Original	x	Measurement	Density	VS, R
Clone		Measurement	Density	SCK, R
Original		Nuclear Symbols	Nucleus	VS, C
Clone		Nuclear Symbols	Nucleus	VS, C
Clone		Nuclear Symbols	Nucleus	VS, C
Clone	x	Nuclear Symbols.	Nucleus	VS, C
Original		Isotopes	Abundance	VS, R
Clone		Isotopes	Abundance	R
Clone		Isotopes	Abundance	R
Clone		Isotopes	Abundance	R
Clone	x	Isotopes	Abundance	VS, R
Clone		Isotopes	Abundance	C

Table X: Cont.

Item	DIF	Content (general)	Content (specific)	Construct
Original	x	Nomenclature	What is the name	SCK
Clone		Nomenclature	What is the name	SCK
Clone		Nomenclature	What is the name	SCK
Clone		Nomenclature	What is the name	SCK
Clone		Nomenclature	What is the name	SCK
Original		Nomenclature	What is the formula	SCK, R
Clone	x	Nomenclature	What is the formula	SCK, R
Original	x	Nomenclature	What is the formula	SCK
Clone	x	Nomenclature	What is the formula	SCK
Clone	x	Nomenclature	What is the formula	SCK
Clone	x	Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	SCK
Clone		Nomenclature	What is the formula	VS, SCK
Clone		Nomenclature	What is the formula	SCK
Original	x	Greatest/Least Number of Atoms		C
Clone		Greatest/Least Number of Atoms		VS, C, R
Clone	x	Greatest/Least Number of Atoms		R, C
Clone		Greatest/Least Number of Atoms		R, C
Clone		Greatest/Least Number of Atoms		C
Clone	x	Greatest/Least Number of Atoms		R, C
Clone		Greatest/Least Number of Atoms		VS, R
Clone		Greatest/Least Number of Atoms		VS, R
Clone		Greatest/Least Number of Atoms		R, C
Clone		Greatest/Least Number of Atoms		C
Original	x	Empirical Formula-% Comp		R, C
Clone		Empirical Formula-% Comp		C
Clone		Empirical Formula-% Comp		R, C
Clone		Empirical Formula-% Comp		R, C
Clone		Empirical Formula-% Comp		C
Clone		Empirical Formula-% Comp		VS, C

Table X: Cont.

Item	DIF	Content (general)	Content (specific)	Construct
Original	x	% Composition		VS, C
Clone		% Composition		VS, C
Clone		% Composition		VS, C
Clone		% Composition		VS, C
Clone		% Composition		VS, SCK
Clone		% Composition		VS, C
Clone		% Composition		VS, C
Clone		% Composition		C
Clone		% Composition		C
Clone		% Composition		C
Original	x	Chemical Reactions	General	VS, R
Clone		Chemical Reactions	General	VS, R
Original	x	Chemical Reactions	Smallest Whole #	VS, C
Clone		Chemical Reactions	Smallest Whole #	VS, C
Clone		Chemical Reactions	Smallest Whole #	VS, C
Clone		Chemical Reactions	Smallest Whole #	VS, C
Original	x	Chemical Reactions	Sum of Simplest Set of Whole #s	VS, C
Clone		Chemical Reactions	Sum of Simplest Set of Whole #s	VS, R, C
Original	x	Limiting Reagents		VS, C
Clone		Limiting Reagents		VS, C
Clone	x	Limiting Reagents		VS, C
Clone		Limiting Reagents		VS, C
Clone		Limiting Reagents		VS, C
Original	x	General Properties of Aq Solns		SCK
Clone		General Properties of Aq Solns		SCK
Clone		General Properties of Aq Solns		SCK
Original	x	General Properties of Aq Solns		VS, SCK
Clone		General Properties of Aq Solns		SCK
Original		Precipitation	Net Ionic	VS, C
Clone	x	Precipitation	Net Ionic	VS, C
Original	x	Precipitation	Solubility	SCK
Clone		Precipitation	Solubility	VS, SCK

Table X: Cont.

Item	DIF	Content (general)	Content (specific)	Construct
Original	x	Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, C
Clone		Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, C
Clone		Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, C
Clone		Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, C
Clone		Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, C
Clone		Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, C
Clone	x	Oxidation-Reduction Reactions	Smallest whole # coefficient	VS, SCK
Original	x	Oxidation-Reduction Reactions		SCK
Clone		Oxidation-Reduction Reactions		SCK, R
Original	x	Oxidation-Reduction Reactions		VS, R
Clone		Oxidation-Reduction Reactions		VS, R
Original	x	Oxidation-Reduction Reactions	Oxidation Number	SCK, C
Clone		Oxidation-Reduction Reactions	Oxidation Number	SCK, C
Original		Concentration of Solutions	What mass	C
Clone		Concentration of Solutions	What mass	C
Clone	x	Concentration of Solutions	What mass	C
Original	x	Gas Laws	What pressure	VS, C
Clone		Gas Laws	What pressure	C
Original	x	Gas Laws	What mass	VS, C
Clone		Gas Laws	What mass	C
Clone		Gas Laws	What mass	C
Original	x	Ideal Gas Equation	Density-relationship	SCK, C
Clone	x	Ideal Gas Equation	Density-relationship	SCK, R
Original	x	Heat	General	SCK, R
Clone		Heat	General	SCK, R
Original	x	Heat	Highest Specific Heat	VS, C
Clone		Heat	Highest Specific Heat	R
Clone		Heat	Highest Specific Heat	VS, C
Original	x	Heat	Final Temp	C
Clone		Heat	Final Temp	C
Clone		Heat	Final Temp	C
Clone		Heat	Final Temp	C
Original	x	Standard Enthalpy of Reaction	Organic Molecule	VS, SCK
Clone		Standard Enthalpy of Reaction	Organic Molecule	VS, SCK
Clone		Standard Enthalpy of Reaction	Organic Molecule	VS, SCK
Clone		Standard Enthalpy of Reaction	Organic Molecule	VS, SCK

Table X: Cont.

Item	DIF	Content (general)	Content (specific)	Construct
Original	x	Standard Enthalpy of Reaction	Hess's Law	VS, C
Clone		Standard Enthalpy of Reaction	Hess's Law	VS, C
Clone		Standard Enthalpy of Reaction	Hess's Law	VS, R, C
Original	x	Standard Enthalpy of Reaction		VS, C
Clone		Standard Enthalpy of Reaction		VS, C
Original	x	Bohr's Theory		R
Clone		Bohr's Theory		R
Clone		Bohr's Theory		R
Clone		Bohr's Theory		R
Clone		Bohr's Theory		VS, R
Clone		Bohr's Theory		VS, R
Clone		Bohr's Theory		SCK, R
Clone		Bohr's Theory		SCK
Clone		Bohr's Theory		VS, R
Original		Quantum Numbers	Which Quantum #	SCK
Clone		Quantum Numbers	Which Quantum #	VS, SCK
Clone		Quantum Numbers	Which Quantum #	VS, SCK
Clone	x	Quantum Numbers	Which Quantum #	SCK
Clone		Quantum Numbers	Which Quantum #	SCK
Clone		Quantum Numbers	Which Quantum #	VS, SCK
Clone		Quantum Numbers	Which Quantum #	VS, SCK
Clone		Quantum Numbers	Which Quantum #	VS, SCK, R
Original	x	Electron Configuration	Paramagnetic	SCK, R
Clone		Electron Configuration	Paramagnetic	SCK, R
Clone		Electron Configuration	Paramagnetic	SCK, R
Original	x	Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C
Clone		Electron Configuration of Ions	Isoelectronic	C

Table X: Cont.

Item	DIF	Content (general)	Content (specific)	Construct
Original	x	Formal Charge		VS, C
Clone		Formal Charge		VS, C
Original		Bond Enthalpy		SCK
Clone	x	Bond Enthalpy		SCK, R
Original		Polarity	Dipole Moment	SCK, R
Clone		Polarity	Dipole Moment	VS, SCK
Clone		Polarity	Dipole Moment	SCK, C
Clone	x	Polarity	Dipole Moment	SCK, C
Original		Polarity		SCK, R
Clone	x	Polarity		SCK, R
Original		Hybridization in Molecules		VS, C
Clone		Hybridization in Molecules		VS, C
Clone		Hybridization in Molecules		VS, C
Clone	x	Hybridization in Molecules		VS, C
Clone		Hybridization in Molecules		VS, C
Clone		Hybridization in Molecules		VS, C
Original	x	Molecular Orbital Theory		SCK
Clone		Molecular Orbital Theory		SCK
Clone	x	Molecular Orbital Theory		SCK
Clone		Molecular Orbital Theory		VS, SCK
Clone		Molecular Orbital Theory		SCK
Original		Intermolecular Forces	Highest Boiling Point	R
Clone		Intermolecular Forces	Highest Boiling Point	R
Clone	x	Intermolecular Forces	Highest Boiling Point	R
Clone		Intermolecular Forces	Highest Boiling Point	R
Original	x	Properties of Liquids	Vapor Pressure	R
Clone		Properties of Liquids	Vapor Pressure	R
Original	x	Crystal Structures	Size of water molecule vs crystal	VS, R
Clone		Crystal Structures	Size of water molecule vs crystal	VS, R
Clone		Crystal Structures	Size of water molecule vs crystal	VS, R
Clone		Crystal Structures	Size of water molecule vs crystal	VS, R
Clone	x	Crystal Structures	Size of water molecule vs crystal	R

Table XI: The effects of common item equating on the examination score.

Semester	Exam	Amount Score Changed By
S10	1	0
S10	2	0
S10	3	0
S10	4	0
F10	1	0
F10	2	0
F10	3	0
F10	4	0
F11	1	1
F11	2	1
F11	3	1
F11	4	2
F12	1	-1
F12	2	-1
F12	3	-1
F12	4	-1
S13	1	0
S13	2	-1
S13	3	-1
S13	4	-1

Table XII: Cloned items in the content area of measurement.

Item	DIF	Identity in Figure 4	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original		A	0.665	0.188	0.693
Clone		B	0.101	2.695	0.093
Clone		C	0.873	0.026	0.654
Clone		D	0.221	1.497	0.195
Clone		E	0.893	0.018	0.979
Clone	x	F	0.012	6.242	
Clone		G	0.567	0.328	0.586
Clone	x	H	0.026	4.976	0.016
Clone		I	0.735	0.115	0.960

Table XIII: Cloned items in the content area of nomenclature.

Item	DIF	Identity in Figure 6	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	A	0.000	24.3	
Clone	x	B	0.000	12.8	
Clone	x	C	0.001	11.0	
Clone		D	0.906	0.0	0.995
Clone		E	0.133	2.3	0.181
Clone	x	F	0.000	12.9	
Clone		G	0.078	3.1	0.057
Clone		H	0.279	1.2	0.281
Clone		I	0.009	6.8	
Clone		J	0.133	2.3	0.181
Clone		K	0.121	2.4	0.123
Clone		L	0.609	0.3	0.746

Table XIV: Cloned items in the content area of greatest/least number of atoms.

Item	DIF	Identity in Figure 8	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	A	0.000	24.1	
Clone		B	0.147	2.1	0.153
Clone		C	0.464	0.5	0.364
Clone		D	0.604	0.3	0.790
Clone	x	E	0.000	14.3	
Clone		F	0.924	0.0	0.898
Clone	x	G	0.047	3.9	0.039
Clone		H	0.055	3.7	0.045
Clone		I	0.030	4.7	
Clone		J	0.280	1.2	0.432

Table XV: Cloned items in the content area of limiting reagents.

Item	DIF	Identity in Figure 10	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	A	0.000	15.7	0.000
Clone		B	0.007	7.2	
Clone		C	0.989	0.0	0.918
Clone	x	D	0.006	7.5	
Clone		E	0.850	0.0	0.550

Table XVI: Cloned items in the content area of oxidation-reduction reactions.

Item	DIF	Identity in Figure 12	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	A	0.000	29.5	0.000
Clone		B	0.058	3.6	0.012
Clone		C	0.141	2.2	0.183
Clone		D	0.578	0.3	0.631
Clone		E	0.534	0.4	0.392
Clone		F	0.077	3.1	0.045
Clone	x	G	0.042	4.2	

Table XVII: Cloned items in the content area of the ideal gas equation.

Item	DIF	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	0.026	4.934	0.044
Clone	x	0.000	16.398	

Table XVIII: Cloned items in the content area of the molecular orbital theory.

Item	DIF	Identity in Figure 14	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	A	0.008	7.1	0.023
Clone		B	0.407	0.7	0.665
Clone	x	C	0.032	4.6	
Clone		D	0.168	1.9	0.292
Clone		E	0.576	0.3	0.337

Table XIX: Cloned items in the content area of the crystal structures.

Item	DIF	Identity in Figure 16	One-Stage Significance	Chi-Squared Value	Two-Stage Significance
Original	x	A	0.000	12.3	
Clone		B	0.652	0.2	0.780
Clone		C	0.394	0.7	0.479
Clone		D	0.895	0.0	0.885
Clone	x	E	0.006	7.6	

Table XX: Item performance for semi-structured interviews.

Question	Gender	Mean
Q1	M	40
	F	32
Q2	M	100
	F	88
Q3	M	67
	F	53
Q4	M	60
	F	53
Q5	M	93
	F	79
Q6	M	80
	F	94
Q7	M	73
	F	56
Q8	M	67
	F	71
Q9	M	47
	F	44
Q10	M	53
	F	48
Q11	M	80
	F	86
Q12	M	60
	F	48
Q13	M	33
	F	48
Q14	M	40
	F	79
Q15	M	100
	F	94
Q16	M	53
	F	38
Q17	M	73
	F	66
Q18	M	7
	F	9
Q19	M	53
	F	59
Q20	M	73
	F	75

Table XXI: Overall performance, time on task, and mental effort on the semi-structured interviews.

Gender	Performance	Time on Task (min)	Mental Effort
M	62	2.4	2.5
F	60	2.6	2.6

Table XXII: DIF items on the semi-structured interviews.

Question	DIF	Favored
1	x	M
2		
3	x	M
4		
5	x	F
6	x	F
7	x	M
8	x	M
9	x	M
10		
11		
12		
13		
14	x	F
15	x	F
16		
17	x	F
18	x	F
19	x	M
20		

Table XXIII: The amount of items that were flagged with having an incorrect heuristic compared that those of the flagged items that had a heuristic.

Question	Flagged	Actual	Filter Correct (%)
1	30	25	83
2	-	-	
3	13	13	100
4	10	10	100
5	-	-	
6	2	2	100
7	11	10	91
8	6	3	50
9	12	7	58
10	13	5	38
11	2	0	0
12	5	5	100
13	9	9	100
14	8	2	25
15	1	1	100
16	18	16	89
17	4	1	25
18	37	32	86
19	7	5	71
20	8	4	50

Table XXIV: The number of heuristics that would have been missed if the fourth stage of the filter would have included the pupil diameter.

Question	Missed
1	3
2	0
3	3
4	2
5	0
6	1
7	2
8	2
9	1
10	0
11	0
12	0
13	4
14	1
15	1
16	1
17	1
18	3
19	2
20	0

Table XXV: The percentage of each heuristic used per question.

Question	Recognition	Representativeness	One-Reason Decision Making	Arbitrary Trends
1		4	96	
2				
3	69		31	
4	91		9	
5				
6		100		
7		8	75	17
8		60	40	
9		10	90	
10		100		
11				
12	38	38	25	
13			100	
14	100			
15	50		50	
16		50	50	
17			100	
18		100		
19	22	22	44	11
20			100	

Table XXVI: The number of times each heuristic was used per item separated by gender. Item nine has five heuristics that could not be classified as either recognition or representativeness without more detail.

Item	Gender	Recognition	Representativeness	One-Reason Decision Making	Arbitrary Trends
Q1	M			7	
	F		1	18	
Q2	M				
	F				
Q3	M	4			
	F	5		4	
Q4	M	3			
	F	7		1	
Q5	M				
	F				
Q6	M		1		
	F		1		
Q7	M			2	
	F		1	7	2
Q8	M				
	F		3	2	
Q9*	M			4	
	F		1	5	
Q10	M		2		
	F		3		
Q11	M				
	F				
Q12	M			2	
	F	3	3		
Q13	M			4	
	F			5	
Q14	M	2			
	F				
Q15	M				
	F	1		1	
Q16	M		2	2	
	F		13	13	
Q17	M				
	F			1	
Q18	M		9		
	F		23		
Q19	M		1	1	1
	F	2	1	3	
Q20	M			3	
	F			1	

Table XXVII: Percentage of heuristics per item separated by gender.

Item	DIF	Heuristics Used by Male Students	Heuristics Used by Male Students (%)	Heuristics Used by Female Students	Heuristics Used by Female Students (%)	Total Number of Heuristics Used	Support Hypothesis
1	X (M)	7	27	19	73	26	*
2		-		-			
3	X (M)	4	31	9	69	13	*
4		3	27	8	73	11	
5	X (F)	-		-			
6	X (F)	1	50	1	50	2	
7	X (M)	2	17	10	83	12	*
8	X (M)	0	0	5	100	5	*
9	X (M)	5	33	10	67	15	*
10		2	40	3	60	5	
11							
12		2	25	6	75	8	
13		4	44	5	56	9	
14	X (F)	2	100	0	0	2	*
15	X (F)	0	0	2	100	2	
16		4	13	26	87	30	
17	X (F)	0	0	1	100	1	
18	X (F)	9	28	23	72	32	
19	X (M)	3	33	6	67	9	*
20		3	75	1	25	4	

Figure 1: Items that were flagged with DIF from the two trial tests from the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS DivCHED EI) separated by gender and the format of the item.

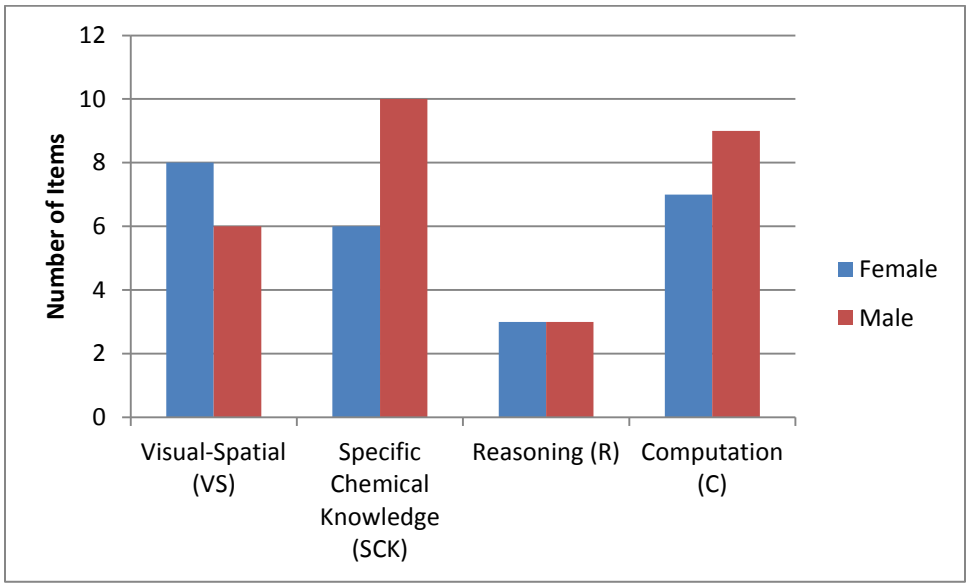


Figure 2: Items that were flagged with DIF from 24 hourly examinations separated by gender and the format of the item.

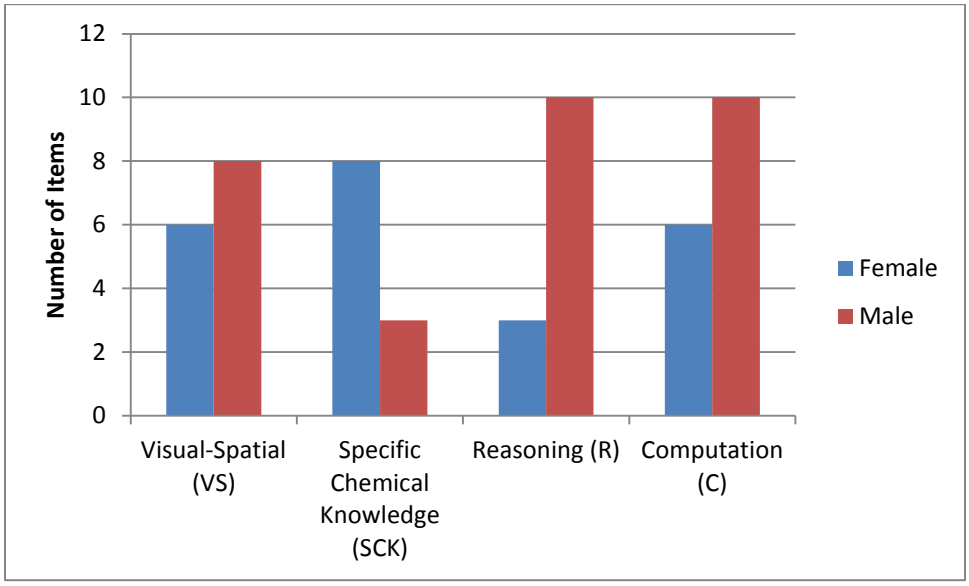


Figure 3: Items that were flagged with DIF from three standardized examinations separated by gender and the format of the item.

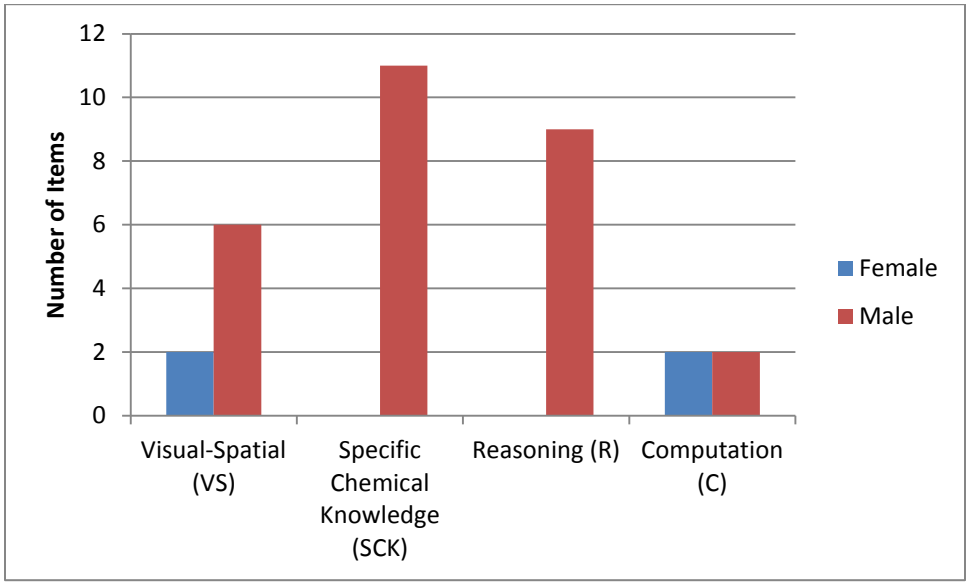


Figure 4: Questions that were cloned items from the content area of measurement.

A. [redacted]

Content clone of item

B. [redacted]

Content and format clones of item

C. When 29.30 g of a metal is placed in 24.0 mL of water, the volume increases as shown in the figure. What is the density of the metal?

A. 0.15 $g \cdot mL^{-1}$

B. 1.0 $g \cdot mL^{-1}$

C. 1.2 $g \cdot mL^{-1}$

D. 6.5 $g \cdot mL^{-1}$

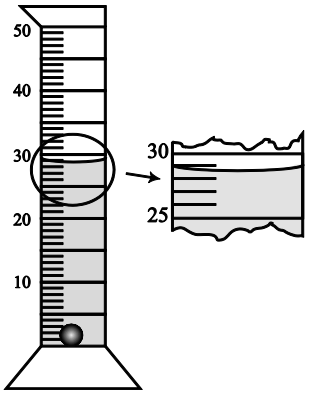
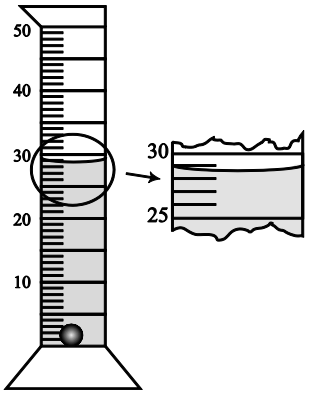


Figure 4: Cont.

D. When 42.25 g of a metal is placed in 22.0 mL of water, the volume increases as shown in the figure. What is the density of the metal?



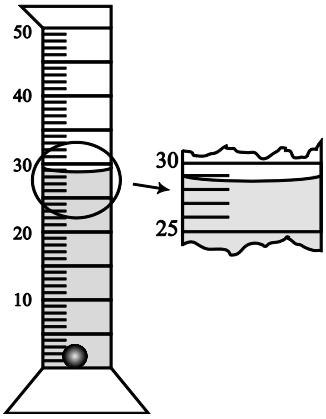
A. 0.15 g·mL^{-1}

B. 1.0 g·mL^{-1}

C. 1.2 g·mL^{-1}

D. 6.5 g·mL^{-1}

E. If the original volume of the liquid in the graduated cylinder was 25.00 mL and the metal has a mass of 31.4 g, what is the density of the metal (in g·cm^{-3})?



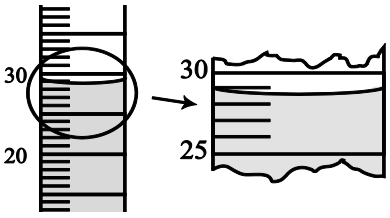
A. 1.1 g·cm^{-3}

B. 1.3 g·cm^{-3}

C. 7.9 g·cm^{-3}

D. 9.0 g·cm^{-3}

F. A 44.724 g sample of lead is carefully placed in the graduated cylinder shown. What is the final volume (in mL) of the water? The density of lead is 11.36 g·cm^{-3}



(A) 0.254 mL **(B)** 3.94 mL

(C) 32.4 mL **(D)** 508 mL

Figure 4: Cont.

G. A sample of iron is carefully placed in a graduated cylinder. The experimental data and the density of iron are listed in the table. What is the final volume (in mL) of the water?

Sample size of iron	38.455 g
Initial volume of water	26.50 mL
Density of iron	7.87 $g \cdot cm^{-3}$

(A) 0.205 mL **(B)** 4.89 mL

(C) 31.4 mL **(D)** 303 mL

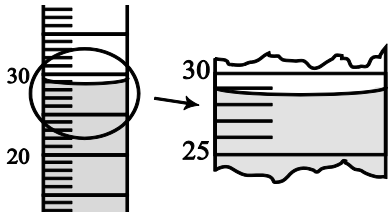
H. Which sample will raise the volume of water by displacement in a graduated cylinder the most?

Density / $g \cdot cm^{-3}$	
Aluminum	2.70
Tin	7.29

(A) 1 g of aluminum **(B)** 10 g of aluminum

(C) 1 g of tin **(D)** 10 g of tin

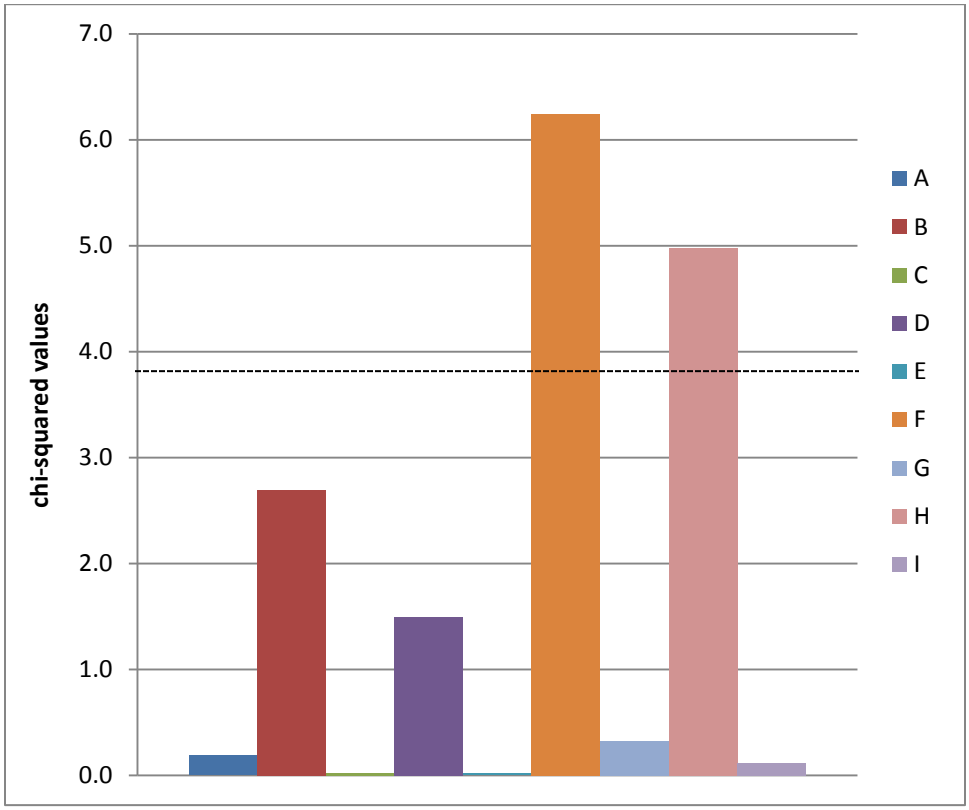
I. A 26.756 g sample of lead is carefully placed in the graduated cylinder with the initial volume of water shown. What is the final volume (in mL) of the water? The density of lead is 11.36 $g \cdot cm^{-3}$.



A. 0.425 mL **B.** 2.36 mL

C. 30.86 mL **D.** 304 mL

Figure 5: A comparison of the chi-squared values for the cloned items (Figure 4) from the content area of measurement.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 6: Questions that were cloned items from the content area of nomenclature.

A. What is the formula of copper (II) phosphate?

(A) Cu_2PO_4 **(B)** $Cu(PO_4)_2$

(C) $Cu_2(PO_4)_3$ **(D)** $Cu_3(PO_4)_2$

Retest of the original item ↓

B. What is the formula of copper(II) phosphate?

A. Cu_2PO_4 **B.** $Cu(PO_4)_2$ **C.** $Cu_2(PO_4)_3$ **D.** $Cu_3(PO_4)_2$

Content clone of the original item ↓

C. What is the formula of iron(II) phosphite?

(A) Fe_2PO_3 **(B)** $Fe(PO_3)_2$

(C) $Fe_2(PO_3)_3$ **(D)** $Fe_3(PO_3)_2$

Retest of the above content clone ↓

D. What is the formula of iron(II) phosphite?

(A) Fe_2PO_3 **(B)** $Fe(PO_3)_2$

(C) $Fe_2(PO_3)_3$ **(D)** $Fe_3(PO_3)_2$

Content clones of the original item with a multivalent cation of +2 or +3 charge

↓

Figure 6: Cont.

E. What is the formula of iron(II) phosphate?

(A) Fe_2PO_4 **(B)** $Fe(PO_4)_2$

(C) $Fe_2(PO_4)_3$ **(D)** $Fe_3(PO_4)_2$

F. What is the formula of chromium(III) carbonate?

A. Cr_3CO_3 **B.** $Cr(CO_3)_3$ **C.** $Cr_2(CO_3)_3$ **D.** $Cr_3(CO_3)_2$

G. What is the formula for chromium(III) sulfite?

A. $Cr_2(SO_3)_3$ **B.** $Cr_2(SO_4)_3$ **C.** $Cr_3(SO_3)_2$ **D.** $Cr_3(SO_4)_2$

Content clones of the original item with a multivalent cation of +4 charge

H. What is the formula of lead(IV) sulfate?

A. $PbSO_4$ **B.** Pb_4SO_4 **C.** $Pb(SO_4)_2$ **D.** $Pb(SO_4)_4$

Content clones of the original item without multivalent cation

I. What is the formula for barium phosphate?

A. $Ba_2(PO_3)_3$ **B.** $Ba_2(PO_4)_3$ **C.** $Ba_3(PO_3)_2$ **D.** $Ba_3(PO_4)_2$

Figure 6: Cont.

J. What is the formula of cadmium phosphate?

(A) Cd_2PO_4 **(B)** $Cd(PO_4)_2$

(C) $Cd_2(PO_4)_3$ **(D)** $Cd_3(PO_4)_2$

14. What is the chemical formula of zinc phosphite?

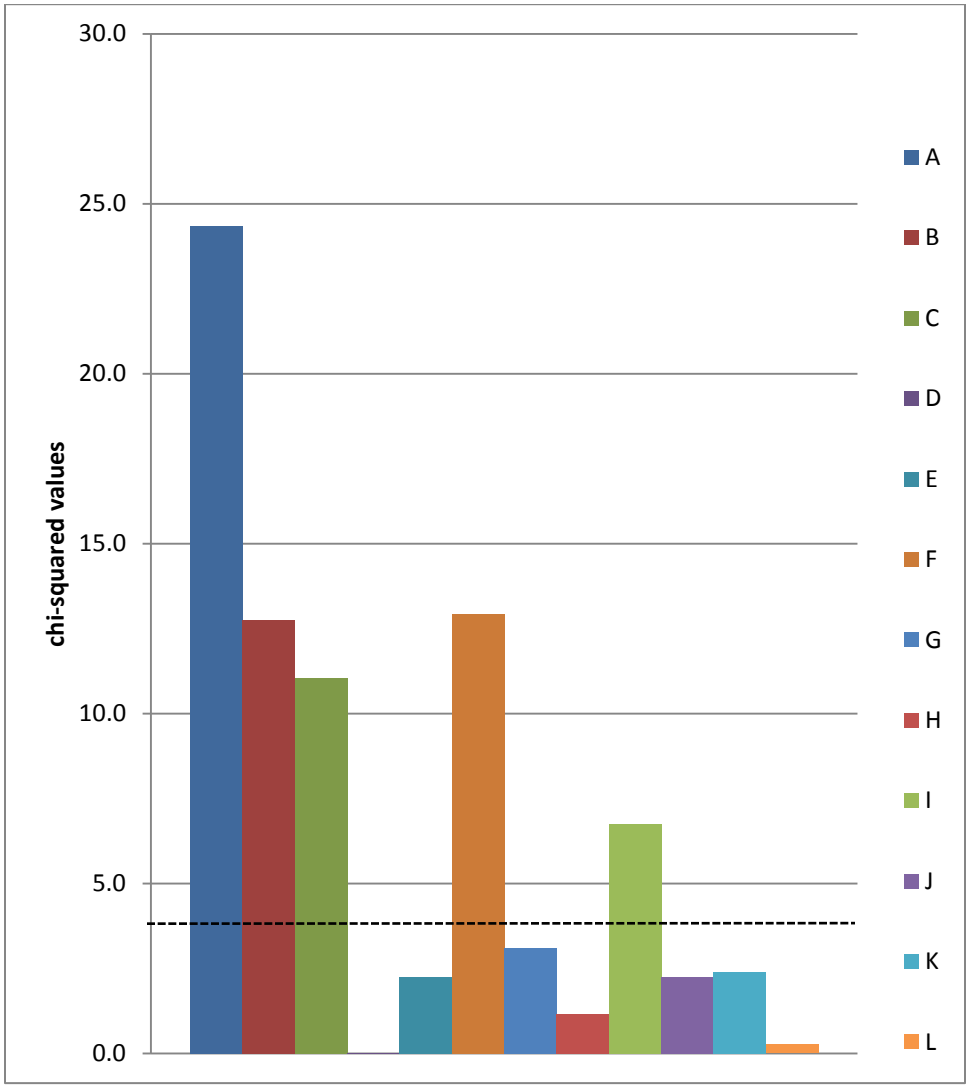
A. $ZnPO_3$ **B.** Zn_2PO_3 **C.** $Zn_3(PO_3)_2$ **D.** $Zn_2(PO_3)_3$

Format clone of the original item 

L. What subscripts would make the correct formula for copper(II) phosphate, $Cu_x(PO_4)_y$?

	x	y
(A)	2	1
(B)	1	2
(C)	2	3
(D)	3	2

Figure 7: A comparison of the chi-squared values for the cloned items (Figure 6) from the content area of nomenclature.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 8: Questions that were cloned items from the content area of greatest/least number of atoms.

A. Which sample has the fewest atoms?

(A) 1.00 g of Pb **(B)** 1.00 g of He

(C) 1.00 g of Fe **(D)** 0.500 g of Na

Content clones of the original item ↓

B. Which sample contains the greatest number of atoms?

A. 2.00 g Li **B.** 1.00 g S **C.** 1.00 g of Al **D.** 2.00 g Au

C. Which sample contains the largest number of atoms?

A. 2.0 g Li **B.** 2.0 g Be **C.** 2.0 g B **D.** 2.0 g C

Content and format clones of the original item ↓

D. Assign which sample has the greatest mass and the greatest number of atoms.

I	1.0 mol of hydrogen molecules
II	0.5 mol of water molecules

	Greatest mass	Greatest number of atoms
A.	**I**	**I**
B.	**I**	**II**
C.	**II**	**I**
D.	**II**	**II**

Figure 8: Cont.

E. Which sample contains the greatest number of total atoms?

A. 1.00 g of chlorine atoms
B. 1.00 g of chlorine molecules
C. 1.00 mol of chlorine atoms
D. 1.00 mol of chlorine molecules

F. Which compound contains the greatest number of atoms?

(A) 1.0 mol of hydrogen molecules
(B) 0.75 mol of water molecules
(C) 1.5 g of hydrogen atoms
(D) 1.5 g of oxygen atoms

G. Which sample contains the greatest number of atoms?

(A) 1.00 mol of nitrogen atoms
(B) 1.00 mol of nitrogen molecules
(C) 1.00 g of nitrogen atoms
(D) 1.00 g of nitrogen molecules

H. Which sample contains the greatest number of atoms?

A. 1.0 mol of water molecules
B. 1.0 mol of oxygen molecules
C. 1.0 mol of hydrogen molecules
D. 1.0 mol of hydrogen atoms

I. Which sample has a mass greater than 25 g?

A. Neither **(A)** nor **(B)**
B. Only **(A)**
C. Only **(B)**
D. Both **(A)** and **(B)**

(A) 0.5 mol chlorine atoms
(B) 0.5 mol chlorine molecules

Figure 8: Cont.

J. Which sample has a mass less than 15 g?

A. Neither **(A)** nor **(B)**

B. Only **(A)**

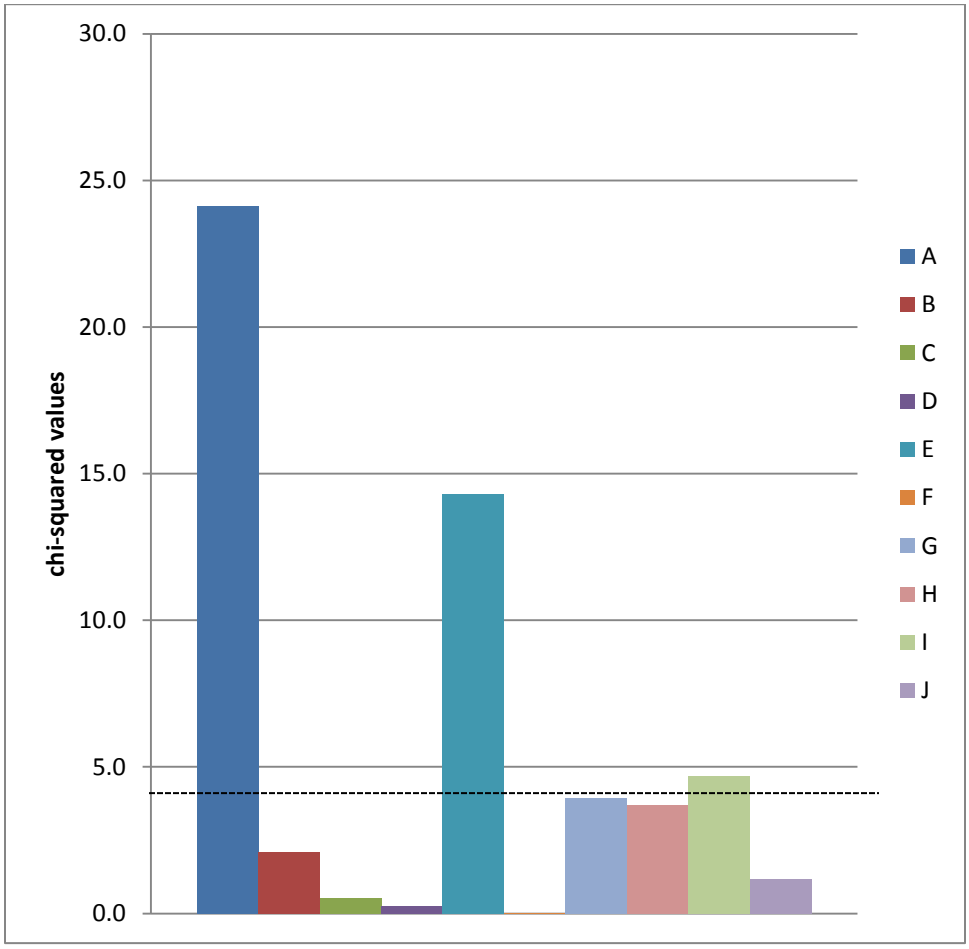
C. Only **(B)**

D. Both **(A)** and **(B)**

(A) 0.5 mol oxygen atoms

(B) 0.5 mol oxygen molecules

Figure 9: A comparison of the chi-squared values for the cloned items (Figure 8) from the content area of greatest/least number of atoms.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 10: Questions that were cloned items from the content area of limiting reagents.

A. If 12 moles of fluorine gas are mixed with 5 moles of sulfur, how many moles of sulfur remain after 9 moles of fluorine react?

$$S + 3F_2 \rightarrow SF_6$$

(A) 1 **(B)** 2 **(C)** 3 **(D)** 4

Retest of the original item ↓

B. If 12 moles of fluorine gas are mixed with 5 moles of sulfur, how many moles of sulfur remain after 9 moles of fluorine react?

$$S + 3F_2 \rightarrow SF_6$$

A. 1 **B.** 2 **C.** 3 **D.** 4

Content clones of the original item ↓

C. How many **total** moles are present when 4 mol of sulfur reacts completely 6 mol of fluorine to produce sulfur hexafluoride?

$$S + 3F_2 \rightarrow SF_6$$

(A) 2 mol **(B)** 4 mol **(C)** 6 mol **(D)** 12 mol

D. If 15 mol of hydrogen is allowed to react with 6 mol of nitrogen, how many moles of nitrogen remain after
12 mol of hydrogen reacts?

$$N_2 + 3H_2 \rightarrow 2NH_3$$

(A) 1 **(B)** 2 **(C)** 3 **(D)** 4

Figure 10: Cont.

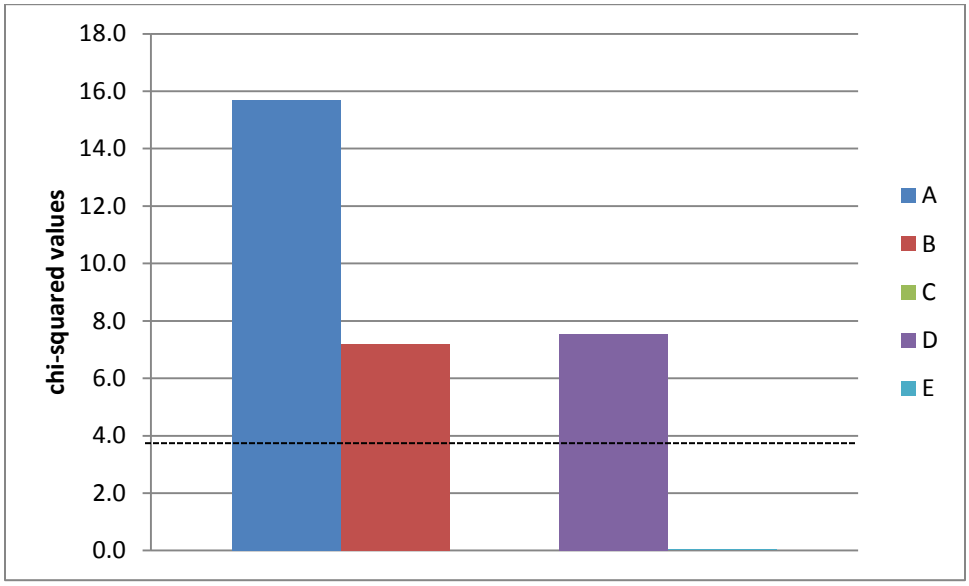
E. What amount (in mol) of excess reactant remains when 6.0 mol of hydrogen reacts with 4.0 mol of nitrogen to produce 4.0 mol of ammonia?

$$3H_2(g) + N_2(g) \rightarrow 2NH_3(g)$$

(A) 2.0 mol H_2 **(B)** 4.0 mol H_2

(C) 1.0 mol N_2 **(D)** 2.0 mol N_2

Figure 11: A comparison of the chi-squared values for the cloned items (Figure 10) from the content area of limiting reagents.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 12: Questions that were cloned items from the content area of oxidation-reduction reactions.

A. What is the smallest whole number coefficient of Cu^{2+} when the equation is balanced?

$Cu^{2+}(aq) + Al(s)$ r $Cu(s) + Al^{3+}(aq)$

(A) 1 **(B)** 2 **(C)** 3 **(D)** 6

Retest of the original item ↓

B. What is the smallest whole number coefficent of Cu^{2+} when the equation is balanced?

$$Cu^{2+}(aq) + Al(s) \rightarrow Cu(s) + Al^{3+}(aq)$$

A. 1 **B.** 2 **C.** 3 **D.** 6

Content clones of the original item ↓

C. What is the smallest whole number coefficent of Cd^{2+} when the equation is balanced?

$$Cd^{2+}(aq) + Cr(s) \rightarrow Cd(s) + Cr^{3+}(aq)$$

(A) 1 **(B)** 2 **(C)** 3 **(D)** 6

D. What is the smallest whole number coefficent of Cd^{2+} when the equation is balanced?

$$Cd^{2+}(aq) + Cr(s) \rightarrow Cd(s) + Cr^{3+}(aq)$$

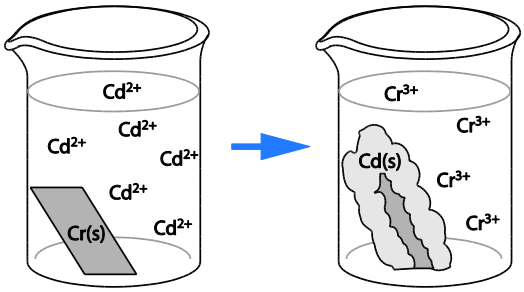
(A) 1 **(B)** 2 **(C)** 3 **(D)** 6

Figure 12: Cont.

E. What is the smallest whole number coefficient of Cd^{2+} when the equation is balanced?

$$Cd^{2+}(aq) + Cr(s) \rightarrow Cd(s) + Cr^{3+}(aq)$$

A. 1 **B.** 2 **C.** 3 **D.** 6



F. What is the smallest whole number coefficient of Cd^{2+} when the equation is balanced?

(A) 1 **(B)** 2 **(C)** 3 **(D)** 6

Content and format clone of the original item

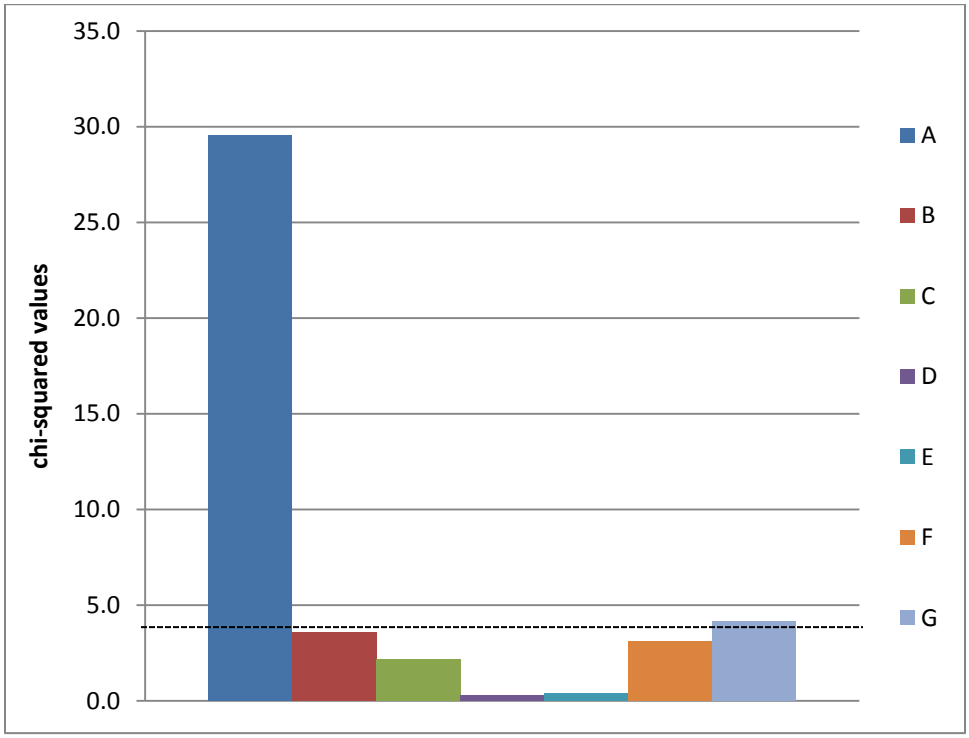
G. What must balance to determine the smallest whole number coefficient of Cd^{2+}?

$$Cd^{2+}(aq) + Cr(s) \rightarrow Cd(s) + Cr^{3+}(aq)$$

A. mass **B.** charge

C. both mass and charge **D.** neither mass nor charge

Figure 13: A comparison of the chi-squared values for the cloned items (Figure 12) from the content area of oxidation-reduction reactions.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 14: Questions that were cloned items from the content area of molecular orbital theory.

A. According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?

(A) A bonding σ molecular orbital and an antibonding σ molecular orbital

(B) A bonding σ molecular orbital and an antibonding π molecular orbital

(C) A bonding π molecular orbital and an antibonding σ molecular orbital

(D) A bonding π molecular orbital and an antibonding π molecular orbital

Retests of the original item 

B. According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?

(A) A bonding σ molecular orbital and an antibonding σ molecular orbital

(B) A bonding σ molecular orbital and an antibonding π molecular orbital

(C) A bonding π molecular orbital and an antibonding σ molecular orbital

(D) A bonding π molecular orbital and an antibonding π molecular orbital

Figure 14: Cont.

C. According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?

(A) A bonding σ molecular orbital and an antibonding σ molecular orbital

(B) A bonding σ molecular orbital and an antibonding π molecular orbital

(C) A bonding π molecular orbital and an antibonding σ molecular orbital

(D) A bonding π molecular orbital and an antibonding π molecular orbital

Content clone of the original item 

D. According to molecular orbital theory, if the two 1s atomic orbitals of hydrogen are mixed, what are the resulting lowest two molecular orbitals?

(A) 2 electrons in the sigma bonding and 0 electrons in the sigma antibonding

(B) 2 electrons in the pi bonding and 0 electrons in the sigma antibonding

(C) 2 electrons in the sigma bonding and 0 electrons in the pi antibonding

(D) 2 electrons in the pi bonding and 0 electrons in the pi antibonding

Format clone of the original item 

Figure 14: Cont.

E. According to molecular orbital theory, if two 1s atomic orbitals are mixed, what are the resulting molecular orbitals that are formed?

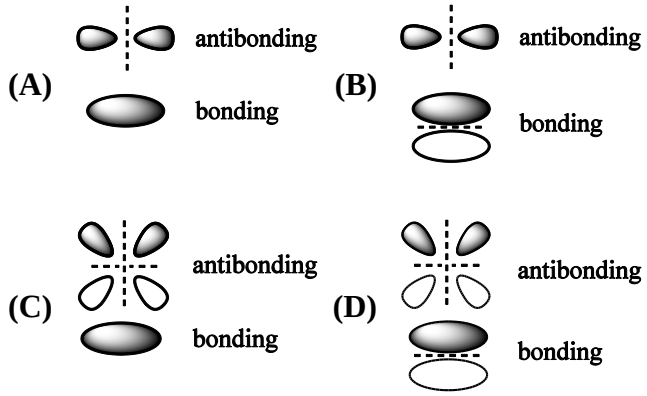
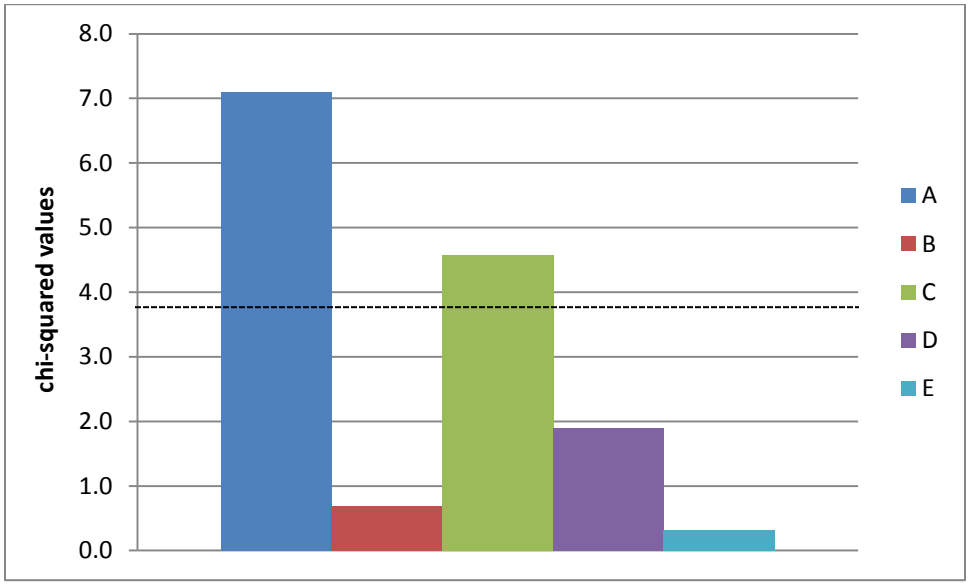


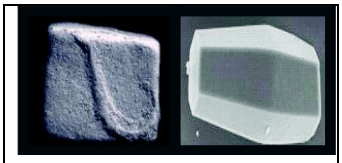
Figure 15: A comparison of the chi-squared values for the cloned items (Figure 14) from the content area of molecular orbital theory.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 16: Questions that were cloned items from the content area of crystal structures.

A. Two crystals are shown to the right.



I **II**

Salt crystal **SiC crystal**

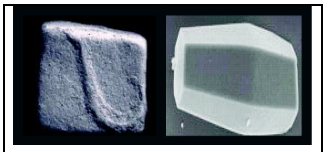
Which is/are smaller than a water molecule?

A. Both **I** and **II** **B.** Only **I**

C. Only **II** **D.** Neither **I** nor **II**

Retests of the original item

B. Two crystals are shown to the right.



I **II**

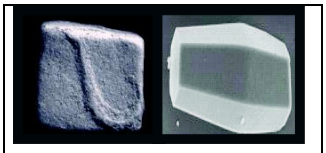
Salt crystal **SiC crystal**

Which is/are smaller than a water molecule?

(A) both **I** and **II** **(B)** only **I**

(C) only **II** **(D)** neither **I** nor **II**

C. Two crystals are shown to the right.



I **II**

Salt crystal **SiC crystal**

Which is/are smaller than a water molecule?

(A) both **I** and **II** **(B)** only **I**

(C) only **II** **(D)** neither **I** nor **II**

Figure 16: Cont.

Format clones of the original item ↓

D. Which is/are smaller than a water molecule?

I	Salt crystal
II	SiC crystal

(A) both **I** and **II** **(B)** only **I**

(C) only **II** **(D)** neither **I** nor **II**

E. Which is smaller, a water molecule or a NaCl crystal and why?

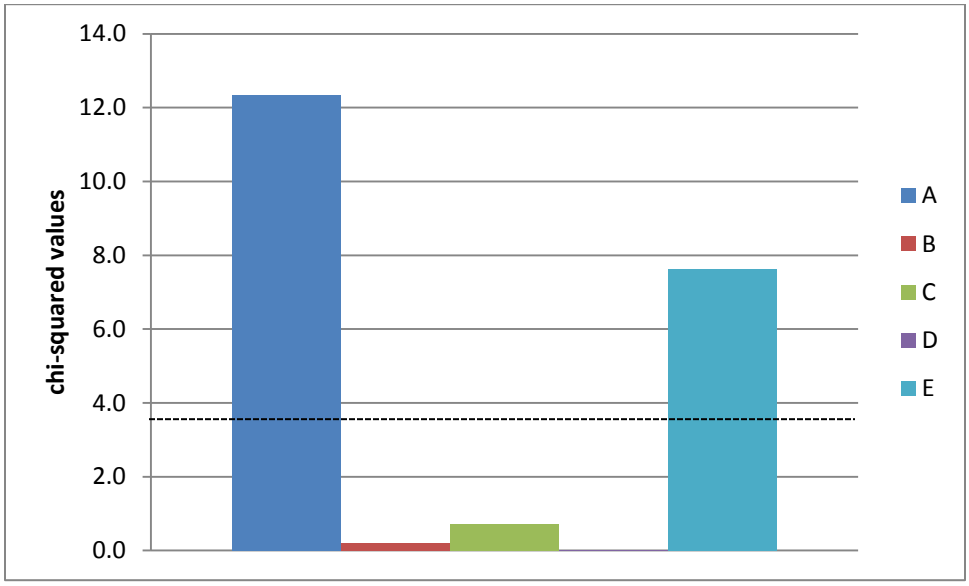
(A) A NaCl crystal because it is made up of 2 atoms

(B) A NaCl crystal because this crystal is smaller than a molecule

(C) A water molecule because atoms are smaller than ions

(D) A water molecule because this molecule is smaller than a crystal

Figure 17: A comparison of the chi-squared values for the cloned items (Figure 16) from the content area of crystal structures.



Any item with a chi-squared above the dashed line has a significant value for the one-stage DIF analysis using the Mantel-Haenszel statistic.

Figure 18: The performance of the graduate student and undergraduate student participants on the assessment-like interviews conducted on the eye tracking instrument.

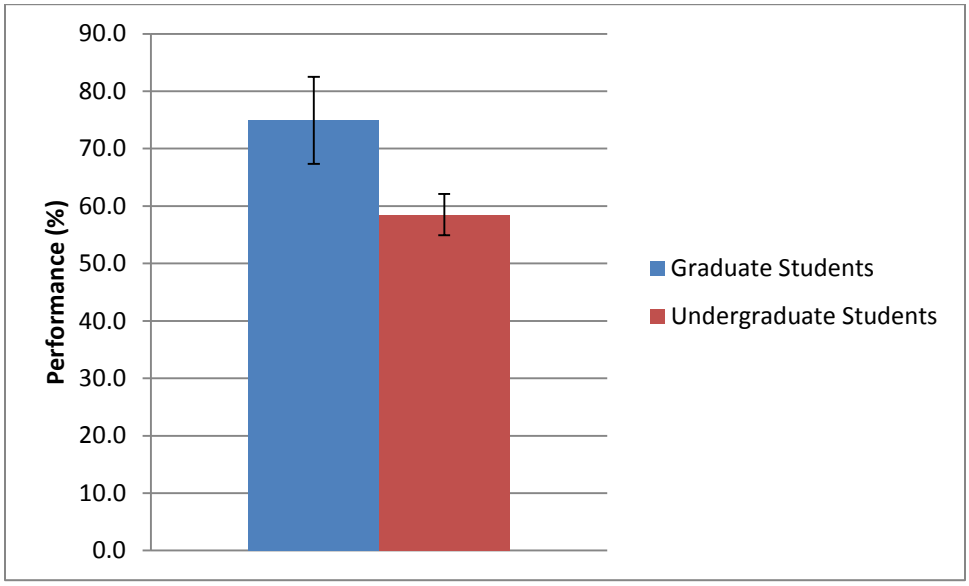


Figure 19: The performance of the participants separated by gender and expertise on the assessment-like interviews conducted on the eye tracking instrument.

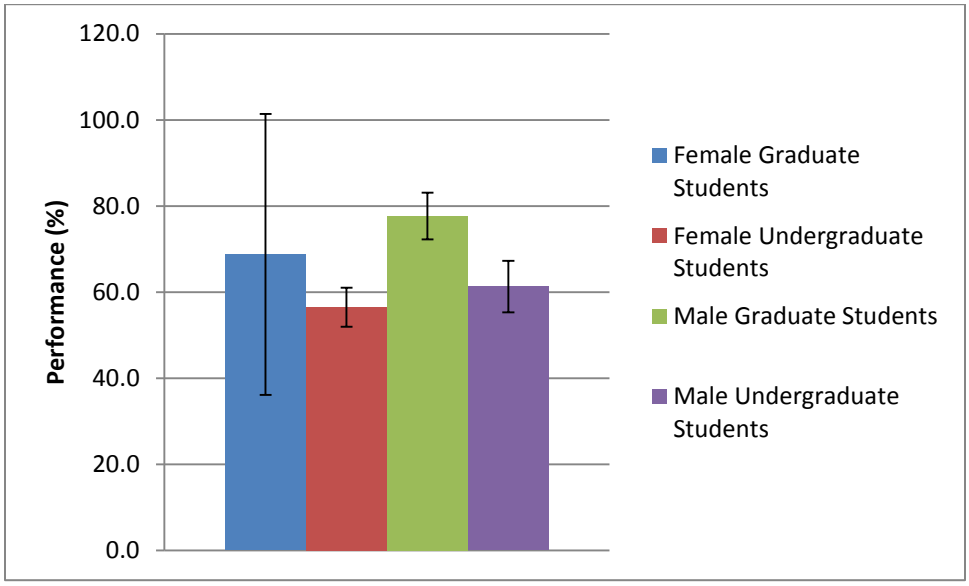


Figure 20: The performance of the participants separated by gender on the semi-structured interviews conducted on the eye tracking instrument.

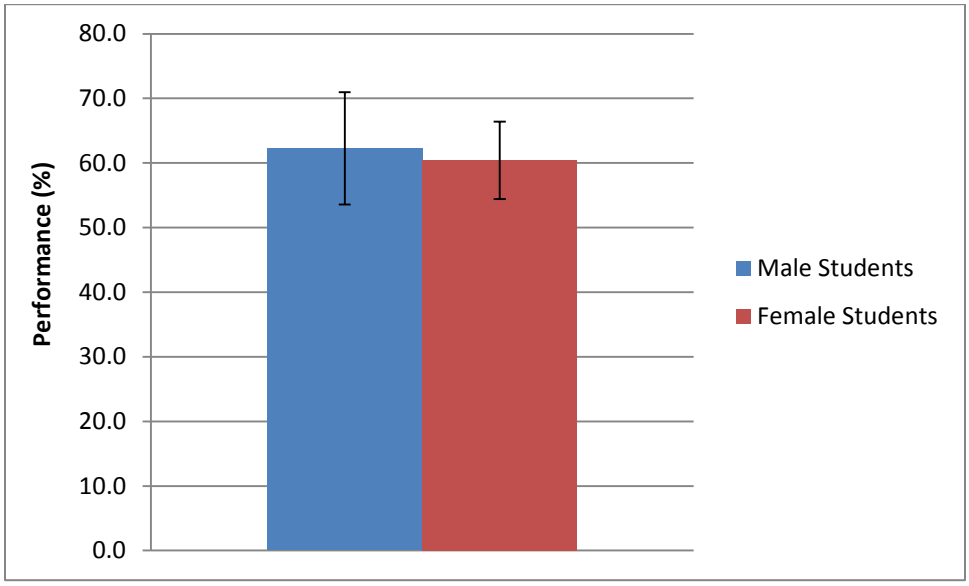


Figure 21: The average time on task of the participants separated by gender on the semi-structured interviews conducted on the eye tracking instrument.

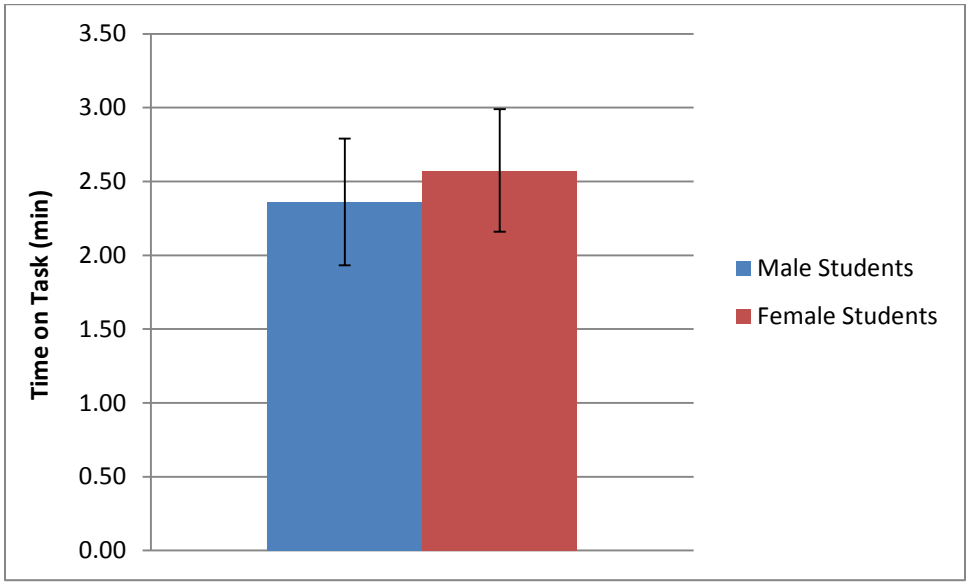
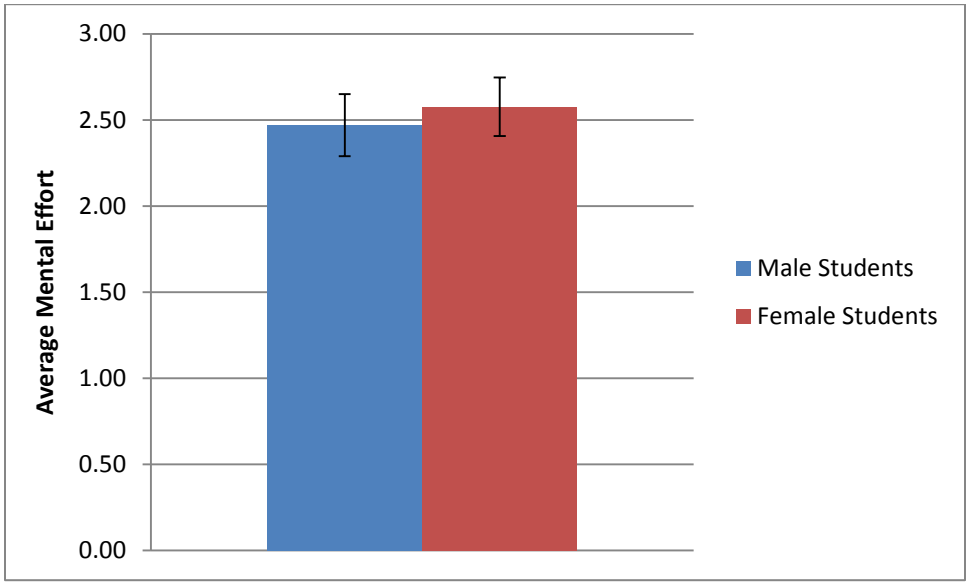


Figure 22: The average mental effort of the participants separated by gender on the semi-structured interviews conducted on the eye tracking instrument.



Reference

1. Maccoby, E. E.; Jacklin, C. N., *The Psychology of Sex Differences*. Stanford University Press: Stanford, California, 1974.

2. Zenisky, A. L.; Hambleton, R. K.; Robin, F. *DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices*; University of Massachusetts Amherst: Amherst, MA, 2003b; pp 1-22.

3. Hamilton, L. S.; Snow, R. E. *Exploring Differential Item Functioning on Science Achievement Tests*; 483; National Center for Research on Evaluation, Standards, and Student Testing. : Los Angeles, CA, 1998; pp 1-43.

4. Kendhammer, L.; Holme, T.; Murphy, K., Identifying Differential Performance in general Chemistry: Differential Item Functioning Analysis of ACS General Chemistry Trial Tests. *Journal of Chemical Education* **2013,** *90* (7), 846-853.

5. Toledo Chemistry Placement Examination. ACS DivCHED, Examinations Institute: 1992.

6. First Term General Chemistry Paired Questions ACS DivCHED, Examinations Institute: 2005.

7. Conceptual Exam ACS DivCHED, Examinations Institute: 2008.

8. Holme, T., Assessment and Quality Control in Chemistry Education. *Journal of Chemical Education* **2003,** *80* (6), 594.

9. Schroeder, J.; Murphy, K. L.; Holme, T. A., Investigating Factors That Influence Item Performance on ACS Exams. *Journal of Chemical Education* **2012,** *89*, 346-350.

10. St. B. T. Evans, J.; Frankish, K., *In Two Minds Dual Processes and Beyond*. Oxford University Press: Oxford, 2009.

11. Maeyer, J.; Talanquer, V., The Role of Intuitive Heuristics in Students' Thinking: Ranking Chemical Substances. *Science Education* **2010**, 1-22.

12. Barrett, L. F.; Tugade, M. M.; Engle, R. W., Individual Difference in Working Memory Capacity and Dual-Process Theories of the Mind. *Psychological Bulletin* **2004,** *130* (4), 553-573.

13. Beatty, J., Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin* **1982,** *91* (2), 276-292.

Chapter 5: Conclusions and Future Work

5.1 Introduction

This chapter will be split into three sections. The first section will discuss the conclusions reached for all four parts of the study: the results of the trial tests, the persistence study, the clone analysis, and finally the semi-structured interviews using an eye-tracking instrument. The next section will focus on the implications for the classroom and how the results of this study can be incorporated into teaching, assessment writing, and how the students are learning. The last section will focus on future directions for the study.

5.2 Conclusions

5.2.1 Determining Items that Exhibited DIF

A two-stage DIF analysis[1] was performed on the two trial tests from the American Chemical Society, Division of Chemical Education, Examinations Institute (ACS-EI). Out of the 140 unique items, 33 items exhibited DIF. On Form A there were 14 items that exhibited DIF, seven that favored male students and seven that favored female students. On Form B there were 19 items that exhibited DIF, 11 that favored female students and eight that favored male students. This analysis satisfied the goal of initially

determining items that exhibited DIF. By using the trial tests, a large nationwide data set (n = 2518) ensured that a widely diverse and large sample set was used for this analysis.

5.2.2 Determining Persistent DIF

Once the initial analysis for determining items that exhibited possible DIF was completed, the next analysis was to determine if DIF was real for these items. This was done by a persistence study using different relevant measures of proficiency. On the 24 hourly examinations that were used to study this persistence, there were 687 items total; 33 (5%) items had a significant value using the Mantel-Haenszel statistic exhibiting persistent DIF. Of those 33 items, 15 of the items that were flagged with persistent DIF favored female students and 18 of the items that were flagged with persistent DIF favored male students. On the three different standardized examinations there were 140 items total; 19 (14%) of them had a significant value using the Mantel-Haenszel statistic exhibiting persistent DIF. Of those 19 items, two of the items that were flagged with persistent DIF favored female students and 17 of the items that were flagged with persistent DIF favored male students. This satisfied the second goal of determining if DIF was persistent. The persistence of DIF for these items then warranted further study as to what component of the item (the content, the format or both) was contributing to the observed DIF as well as the reason why this DIF was occurring.

5.2.3 Possible Causes of DIF

As a way to determine the possible causes of DIF, the items were first classified by content and format. When a grouping of items that test the same specific content area using the same format of the item favor only one subgroup, these items are worthy of discussion. There were many items presented that had the same content and format areas that favored only one gender, such as the nomenclature items that favored female students, or the measurement (density) items that favored male students. Most of the items that exhibited persistent DIF were cloned based on the content and/or format of the item. For example, through cloning studies it was found that there were certain content areas such as nomenclature that seemed to favor female students when the item contained a multivalent cation with a +2 or +3 charge combining with a polyatomic anion with a –2 or –3 charge. It was also found that for other types of nomenclature items, there was no differential performance on the item. When considering the format of the items, there were certain formats that tended to favor one subgroup. For example, when clones in the content area of oxidation-reduction reactions were tested, the data suggests that the cause of DIF for those items was both the format combined with that specific content area. Over six semesters of testing, the specific content areas that consistently show DIF favoring male students were measurement (density), greatest/least number of atoms, limiting reagents, ideal gas equation, and crystal structures and favoring female students were nomenclature and molecular orbital theory. The formats that tend to favor male students were visual-spatial, reasoning, and computation and to favor female students was specific chemical knowledge. While it would be irresponsible to say the sole cause of DIF on items was the format or the content, or even that certain content or formats

would always favor one gender, the results do give more insight into the possible causes of DIF.

5.2.4 Why DIF Happens

Finally to determine the possible reasons why DIF was happening, semi-structured interviews using an eye-tracking instrument were conducted to gain information about the student's problem-solving process. It was found that in seven cases the possible reasons why DIF was happening were a result of one subgroup preferentially using an incorrect heuristic. These items were in the specific content areas of measurement (density), greatest/least number of atoms, stoichiometry-general, and crystal structures. Additionally, the format inclusions of visual-spatial, reasoning, and computation for these items could also be a contributing factor to the observed results. Through investigating objective measures, such as the student's performance and time on task as well as subjective measures such as the student's self-reported mental effort, it was evident that students were using incorrect heuristics to solve some of the items. These measures were low as one would expect if the student were using a heuristic suggesting that they were bypassing their working memory and instead using system 1 processing from the Dual Processing Theory. The results from this study helps to provide greater understanding about DIF and some of the reasons it occurs. Ultimately, a test most likely cannot be designed that entirely avoids the existence of possible DIF via random fluctuation or otherwise. Therefore, knowing more about methods to identify DIF

and how to that minimize DIF, will improve the quality of the measurement and associated judgment from the assessment.

5.3 Implications for the Classroom

By knowing more information about the types of questions that exhibit persistent DIF, one way to avoid assessment bias is to not put those types of items on high-stakes assessments. However, that is not an ideal situation. Avoiding certain items/topics is tantamount to narrowing the domain which is a pretty significant problem when talking about score interpretation. So instead of avoiding putting these items on assessments, the more practical application would be to look at the content areas and formats that seem to favor one subgroup versus another. When writing an assessment take these areas under consideration so if asking a question in a content area such as molecular orbital theory that tends to favor female students, include a visual-spatial item, computation, or reasoning format that tend to favor male students. This way the direction that the content favors is essentially "canceled" out by the format favoring other direction. Additionally, being mindful of the combination of formats and content areas that may amplify DIF can also assist in minimizing the effects of incorporating items with possible favor. For example, knowing that the specific content area of formula calculations focusing on assessing the number of atoms in a sample tends to favor male students, asking this question on an assessment with a visual-spatial component (a format that also tends to favor male students) can increase the probability that the item will show favor for male

students. This approach may be one that is used by a larger number of instructors (as it is easier to simply avoid asking certain questions or certain ways).

However, a proactive approach could be to consider how the material is presented during instruction. This would require examining the curriculum and finding the content areas that tend to favor one gender versus another and reformat how that section is taught making sure to include examples with formats that favor both genders. For example, if discussing a crystal structure, make sure when showing a figure to thoroughly explain everything with that figure instead of just pointing out that information can be gained from it. Actually go through verbally, and explain how to use the figure to extract the needed information. Therefore, possible favor that has been found on assessment items to favor one gender of students may now not occur because the initial instruction of the material assisted in the learning of both subgroups as the images that are presented to students may be interpreted differently without specific guidance.

Another useful application of the results is to focus on incorrect heuristics the students exhibit having when solving certain items. Again, going through the curriculum and highlighting those areas where the heuristics are occurring and making sure to point out the correct way to solve the items in the area using analytical reasoning, instead of short cuts, can assist in minimizing possible DIF. Another implication to motivate students to engage in analytical processing for solving any problem could be that when working through problems as a class to ask students what they would expect as an answer, then when the answer is determined instead of asking "does the answer make sense", asking "why does this answer make sense?" If the students are forced to go back

and analytically reason their way through the problem they are less likely to use incorrect heuristics.

5.4 Future Work

While this study did identify items that exhibited persistent DIF, some of the causes of DIF, and possible reasons why DIF occurs, questions still remain and as with any research project, new ones have surfaced. Considering another cycle of clones of each of the items could be conducted to focus on specific contents and/or formats that are causing the items to exhibit DIF. More interviews could be conducted, gaining a more representative sample, as well as strengthening some of the conclusions based on the heuristics that were observed. A new area to explore is a two-stage interview process. First, the students would sit for semi-structured interviews using the eye-tracking interview, then after using the filter and checking for heuristics the students would come back and redo those items, but this time the interviewer would ask more probing questions as to why they are doing each step. Because the initial data is already collected one wouldn't have to worry about leading students to a correct answer, but instead if they got an answer wrong to have them walk through how they would get to the correct answer, seeing if they rely on a different heuristic or if they will resort to analytical reasoning. Similarly, one could conducting the semi-structured interviews on experts, not only to obtain their problem-solving processes, but also to determine if the filter design is applicable for different populations of students. Using the same methods from this work, other subgroups could be examined for differential performance as well. Finally, along

with the application for the filter design, more analysis should be conducted on the use of pupil diameter. The possibility of using the rich dataset of pupil diameters as a window into objective load on working memory warrants determining how to properly use it and in what applications it would be most useful.

Perhaps the greatest use of this research lies outside of future research endeavors and, instead, resides in practice. Studying how to integrate the findings of this research into practice would require additional research, however the applications of that work have the potential to benefit the greatest number of future students. For example, studying different methods of content delivery or assessment techniques that attempt to minimize the observed differential performance could be widely disseminated and integrated into similar courses at other institutions. Extending these studies into examining other subgroups using the same methodologies can also reach into classrooms through changes to improve differential performance overall. Ultimately, people are regularly judged based on how they perform on assessments. Those who teach, those who assess and those who make judgments based on these assessments can work towards a common goal of reducing favor on these assessments. It is imperative that assessments are fair and this study focused on one avenue that provides information to come closer to achieving that. By implementing the results of this study, assessment writers are one step further in achieving the goal of reducing favor on assessments.

References

1. Zenisky, A. L.; Hambleton, R. K., Detection of Differential Item Functioning in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach. *Educational and Psychological Measurement* **2003a,** *63* (1), 51-64.

Appendix A

Figure 1: The PARSCALE command file that was used for the 2-parameter item response theory fit.

```
F09 FinalI

>FILE     DFNAME='C:\F09_FinalI_NEW\F09_FinalI_New.DAT',SAVE;
>SAVE     PARM='C:\F09_FinalI_New\TEST.PAR',SCORE=
          'C:\F09_FinalI_New\TEST.SCO';
>INPUT    NIDW=10, NTOTAL=40, LENGTH=40, NTEST=1, MAXCAT=2, MGROUP=2;
          (10A1,1X,1A1,1X,40A1)
>TEST     TNAME='TRIAL', NBLOCK=40;
>BLOCKS   NITEMS=1, NCAT=2, ORIGINAL=(0,1), MODIFIED=(1,2), REPEAT=40;
>MGROUP   GNAME=(FEMALES,MALES), GCODE=(F,M), DIF=(0,1,0,0),
          REFERENCE=1;
>CALIB    LOGISTIC, NQPT=31, CYCLES=200, CRIT=0.001, NEWTON=2,
          PRIORREAD, SPRIOR;
>PRIORS   SMU=(1.0(0)40), SSIGMA=(0.2(0)40);
>SCORE    ;
```

This command file was for the Fall 2009 ACS DivCHED EI First Term General Chemistry Paired Questions final examination. The following information about the command file was obtained from the IRT manual from Scientific Software International (SSI).[1] The FILE command line tells where the data is and the SAVE command line specifics where the output files will be saved. The INPUT command line describes the data and other criteria about the data for the analysis. If doing a DIF analysis it is important to include the MGROUP which specifies the number of groups. The TEST command line is used to indicate the test. The BLOCKS command line is to give the blocks a title and to select which items belong in which blocks. The MGROUP command line is used for the DIF model and includes information about the subgroups. The CALIB command line used to determine the type of fit that is being used and any

adjustments to the items or category parameters. The use of PRIORREAD and SPRIORS was used for this study to help calibration the tails of the distribution that contained some odd outliers. This did not fix the slope but clean up the ends of the distribution. The PRIORS command line indicate that amount of contrain put on the mean and standard deviation of the item slopes. If a DIF analysis is conducted the SCORE command line is not used.

A lot of information can be obtained for the output files, but for the purpose of this study the location values (b) were of interest to use for common item equating. There are different ways to do common item equating; because for each semester there is a different sample of students who take different hourly exams, in addition to a common test that is the same every semester (the anchor test), it was decided to do common item equating with Item Response Theory using an anchor test.[7] This was done with the PARSCALE program by Scientific Software International.[12] For information on the PARSCALE command file see Appendix A. The anchor test used was the ACS DivCHED EI First Term General Chemistry Paired Questions final examination and every test and semester was equated to the Fall 2009 semester. In order to obtain the most valid measure of equating students, the final examination for the course was selected as the anchor test because it was given after a semester of the same instruction to all students under the same conditions of testing. A 2-parameter fit was chosen because of the sample size of students. Also the use of priors was incorporated to help calibration for they give a little extra stability to estimation but doesn't bias or distort the estimates. To use common item equating with a 2-parameter fit in item response theory, the location values (*b*) were calculated for all the anchor items for each semester. Then the difference

in *b* values for each item from the anchor test was calculated comparing each semester back to Fall 2009. For example, the location value for question 1 on the Spring 2009 ACS DivCHED EI First Term General Chemistry Paired Questions final examination was subtracted from the location value for question 1 on the Fall 2009 ACS-EI First Term General Chemistry Paired Questions final examination, and so on for each question. Then the average of the differences (*m*) was calculated. This average of the differences was then added to each students' latent trait score (Z-score) for each hourly examination given.

References

1. IRT from SSI. du Toit, M., Ed. Scientific Software International, Inc.: 2003.

Curriculum Vitae for Lisa Kendhammer

Education

University of Wisconsin – Milwaukee, Milwaukee, WI
Ph.D., Chemical Education and Analytical Chemistry; GPA 3.9/4.0
August 2007-August 2013

Thesis: Studies in Analytical Chemistry and Chemical Education.

Part 1: Characterization of Complex Organics by Raman Spectroscopy and Gas Chromatography.

Part 2: Differential Item Functioning on Multiple-Choice General Chemistry Assessments.

Cardinal Stritch University, Milwaukee, WI
Degree: B.S. in Chemistry; *magna cum laude;* GPA 3.8/4.0
August 2003 – May 2007

Work Experience

August 2007–Present: Research Assistant, University of Wisconsin – Milwaukee, Department of Chemistry & Biochemistry, Milwaukee, WI.

Chemical Education

- Writing exam and homework items
- Conducting semi-structured interviews
- Working with an eye tracking instrument
- Applying Dual-Processing Theory
- Using Statistical Analysis
    - Mantel-Haenszel Statistic
    - Logistic Regression
    - Correlations
    - Item Response Theory
    - Common Item Equating
    - Cross-tab Analysis
    - T-Tests
- Proficient in SPSS, Endnote, Experiment Center, BeGaze, Microsoft Word, Microsoft Excel, and Microsoft Powerpoint

Analytical

- Identification of functional groups using Raman Spectroscopy to build a spectral database for varnish identification

- Using headspace SPME with GC-PFPD for varnish identification
- Aging of varnishes to determine a "naturally aged" sample
- Built an aging chamber to create varnish standard reference materials
- Instrumentation
    - GC (TCD)
    - GC-MS (ion trap)
    - GC-PFPD
    - UV-Vis Spectrometer
    - FT-Raman Spectrometer
- Proficient in Endnote, OMNIC, ChemStation, Saturn/Star, Microsoft Word, Microsoft Excel, and Microsoft Powerpoint

August 2007–May 2012: Teaching Assistant, University of Wisconsin – Milwaukee, Department of Chemistry & Biochemistry, Milwaukee, WI.

Skills

- Wrote weekly quizzes and worksheets
- Mini lectures and reviews on lecture topics
- Taught problem solving strategies
- Graded (quizzes, exams, laboratory write-ups)
- Pre-laboratory lectures on concepts and techniques required for the laboratory
- Techniques used in teaching laboratories:
    - UV-Vis Spectroscopy
    - Titrations (pH, complexiometric, colorimetric, iodometric)
    - Chromatography (paper, ion-exchange, gas)
    - Flame Photometry
    - Fluorimetry
    - Cyclic Voltammetry

Courses Taught

- Chemistry 102, General Chemistry I
- Chemistry 104 General Chemistry II
- Chemistry 105, General Chemistry II (Engineer Majors)
- Chemistry 221, Elemental Quantitative Analysis

Guest Lecturer

- Chemistry 104, General Chemistry II-multiple times
- Chemistry 221, Elemental Quantitative Analysis-multiple times
- Chemistry 524, Intermediate Analytical Chemistry (Instrumental Analysis)-multiple times

Mentor for new Teaching Assistants: Fall 2011

May 2007-August 2007: BOD Analyst Intern, Milwaukee Metropolitan Sewage District, Milwaukee, WI.

May 2006-September 2007: BOD Analyst Intern, Milwaukee Metropolitan Sewage District, Milwaukee, WI.

August 2003-May 2007: Chemistry Prep Room Attendant, Cardinal Stritch University, Milwaukee, WI.

Publications

First Author

1. L. Kendhammer, T. Holme, K. Murphy. Identifying differential performance in general chemistry: Differential Item Functioning analysis of ACS General Chemistry trial tests. *J. Chem. Educ.*, **2013**, *90* (7), pp 846–853.

Presentations

Invited Presentations-Presenting Author

1. L. Kendhammer, T. Holme, K. Murphy. "Differential item functioning on multiple choice general chemistry assessments" Carroll College Seminar Series, Waukesha, WI (November 2012).
2. L. Kendhammer , T. Holme K. Murphy. "Examining the persistence of potential DIF on general chemistry assessments" 22st Biennial Conference on Chemical Education, University Park, PA (August 2012).
3. L. Kendhammer, K. Murphy. "Differential Item Functioning (DIF) on multiple choice general chemistry assessments" American Chemical Society's 42nd Meeting of the ACS Central Region, Indianapolis, IN (June 2011).

Peer-Reviewed Presentations-Presenting Author

1. L. Kendhammer , T. Holme K. Murphy. "Differential item functioning on multiple choice general chemistry assessments" 22st Biennial Conference on Chemical Education, University Park, PA (July 2012).

Contributed Presentations—Presenting Author

1. L. Kendhammer, K. Murphy. "Examining students' problem solving pathways on items that exhibited potential Differential Item Functioning." CER Graduate Student Conference, Oxford, OH (July 2013). {poster}
2. L. Kendhammer, K. Murphy. "Examining the effect of format and content on the persistence of potential DIF on general chemistry assessments." "Chemistry Education Research & Practice" Gordon Research Conference, Newport, RI (June 2013). {poster}
3. L. Kendhammer, K. Murphy. "Overview of persistence of potential DIF on general chemistry assessments." American Chemical Society's Great Lakes Regional Meeting, LaCrosse, WI (June 2013).
4. L. Kendhammer, J. H. Aldstadt. "Characterization of Carboxylic Acids by Raman Spectroscopy and Gas Chromatography." American Chemical Society's Great Lakes Regional Meeting, LaCrosse, WI (June 2013).
5. L. K. Kendhammer, K. Murphy. "Differential item functioning on multiple choice general chemistry assessments" American Chemical Society's Fall 2011 National Meeting & Exposition, Denver, CO (August 2011).

6. L. K. Kendhammer, Joseph H. Aldstadt III. "Characterization of complex organics by FT-Raman: Building a spectral database to provide insight into the reactivity of carboxylic acids found in soils and varnishes" American Chemical Society's Fall 2011 National Meeting & Exposition, Denver, CO (August 2011). {poster}
7. L. K. Kendhammer, T. Holme, K. L. Murphy. "Differential item functioning on multiple choice general chemistry assessments" 21st Biennial Conference on
Chemical Education, Denton, TX (August 2010).
8. L. K. Kendhammer, T. Holme, K. L. Murphy. "Differential item functioning on multiple choice general chemistry assessments" American Chemical Society's Spring 2010 National Meeting & Exposition, San Francisco, CA (March 2010). {poster}

Contributed Presentations—Co-Author

1. Kendhammer, Lisa, Holme, Thomas, Murphy, Kristen. "Differential item functioning on multiple choice general chemistry assessments" 22nd International Conference on Chemistry Education and 11th European Conference on Research in Chemical Education, Rome, Italy (July 2012).
2. D.T. Qadah, L.K. Kendhammer, F. El-Sheikh, and J.H. Aldstadt. "Modification of a purge & trap gas chromatograph for pyrolysis studies: Application to humic substances in soils and antique varnishes on museum objects", 59th Pittsburgh Conference on Analytical Chemistry & Applied Spectroscopy, Chicago, IL (March 2009). {poster}

Honors & Awards

- Certificate of Recognition for notable impact on a program (2012)
- Student Organization Service Award (2012)
- DES Scholastic Honors Society (2006-2007)
- Science Department Scholarship Ambassador (2003-2007)
- Cardinal Stritch University's Dean's List (2003-2007)
- National Dean's List (2003-2007)

Volunteer and Service:

- American Chemical Society's Younger Chemists Committee (YCC), chair, organized a Career Seminar Development Day with workshops and resume reviews for over 50 participants. Fall 2012
- American Chemical Society's Younger Chemists Committee (YCC), co-chair, organized a poster session. Fall 2011
- Mentor for new teaching assistants in the Department of Chemistry & Biochemistry. Fall 2011
- Society for Applied Spectroscopy (SAS), community education outreach, creating and implementing an original forensic UV-Vis Spectroscopy lab designed for high-school students. This was performed with 4 area high schools (approx. 500 students). May 2011 & May 2012.

Professional Affiliations:

- American Chemical Society's Younger Chemists Committee (YCC) 2011-current

- Society for Applied Spectroscopy (SAS), member of the student section at UWM, 2010–current
- American Chemical Society member (ACS) 2010-current