Proof validation in Euclidean geometry: A comparison of novices and experts using eye tracking

by

Paul Michael Flesher

B.S., Fort Hays State University, 2013
M.S., Kansas State University, 2015

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Mathematics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Abstract

This dissertation investigates and compares the methods of proof validation utilized by novice and expert mathematicians within the realm of Euclidean geometry. With the use of eye tracking technology, our study presents empirical evidence supporting claims previously studied only through the use of verbal protocols.

Our investigation settles a series of contentious results surrounding the practical implementation of the generalized validation strategy called *zooming out* (Inglis and Alcock, 2012; Weber, Mejia-Ramos, Inglis, and Alcock, 2013). This strategy analyzes the overall structure of a proof as an application of methods or logical chunks. Settling the debate through use of longer and more complicated proofs devoid of blatant errors, we found that validators do not initially skim-read proofs to gain structural insight. We did however confirm the practical implementation of *zooming out* strategies.

The literature identifies within the proof validation process specific differences between novices and experts. We are interested in a holistic understanding of novice and expert validations. We therefore present the direct comparison of entire validation processes that assess the similarity of novice and expert overall validation attempts. We found that the validation processes of novices and experts share a certain degree of similarity. In fact novices tend to be closer to experts than to other novices. And when validations are clustered, the groups are heterogeneous with regard to mathematical maturity.

Our investigation expands the proof validation literature by including diagrams in the proof validation process. We found that experts tend to spend more time proportionally on the diagram than novices and that novices spend more time on the text. Furthermore, experts tend to draw more connections within the diagram than novices as indicated by a higher proportion of

attentional changes within the diagrams. Experts seem to draw on the power of visualizations within the mathematics itself, spending more time on conceptual understanding and intended connections.

Proof validation in Euclidean geometry: A comparison of novices and experts using eye tracking

by

Paul Michael Flesher

B.S., Fort Hays State University, 2013
M.S., Kansas State University, 2015

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Mathematics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Dr. Andrew Bennett

Abstract

This dissertation investigates and compares the methods of proof validation utilized by novice and expert mathematicians within the realm of Euclidean geometry. With the use of eye tracking technology, our study presents empirical evidence supporting claims previously studied only through the use of verbal protocols.

Our investigation settles a series of contentious results surrounding the practical implementation of the generalized validation strategy called *zooming out* (Inglis and Alcock, 2012; Weber, Mejia-Ramos, Inglis, and Alcock, 2013). This strategy analyzes the overall structure of a proof as an application of methods or logical chunks. Settling the debate through use of longer and more complicated proofs devoid of blatant errors, we found that validators do not initially skim-read proofs to gain structural insight. We did however confirm the practical implementation of *zooming out* strategies.

The literature identifies within the proof validation process specific differences between novices and experts. We are interested in a holistic understanding of novice and expert validations. We therefore present the direct comparison of entire validation processes that assess the similarity of novice and expert overall validation attempts. We found that the validation processes of novices and experts share a certain degree of similarity. In fact novices tend to be closer to experts than to other novices. And when validations are clustered, the groups are heterogeneous with regard to mathematical maturity.

Our investigation expands the proof validation literature by including diagrams in the proof validation process. We found that experts tend to spend more time proportionally on the diagram than novices and that novices spend more time on the text. Furthermore, experts tend to draw more connections within the diagram than novices as indicated by a higher proportion of

attentional changes within the diagrams. Experts seem to draw on the power of visualizations within the mathematics itself, spending more time on conceptual understanding and intended connections.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my advisor Dr. Bennett for all of his support and guidance throughout my graduate studies. He has provided a great number of insights and opportunities that positively shaped not only my research but my life in general. I would like to acknowledge and thank Dr. Hakobyan for his interest and support throughout my exploration of research. Your encouragement, presence, and intellect were instrumental in many ways. I am very grateful for Dr. Loschky introducing me to the study of visual cognition and for the use of his laboratory. I would also like to extend a word of thanks to the other members of my committee: Dr. Natarajan and Dr. Wang.

John Hutson deserves a big thank you for providing not only his expertise but vast amounts of his time for the research at hand. He played an instrumental role in planning, implementing, and running the experiment, not to mention helping in the data analysis as well. I greatly appreciate your contributions. The bulk of my research was tempered through the meetings of the visual cueing group. These meetings pushed me cognitively and continually led to important developments in this research endeavor. I extend a special thank you for all of the challenges and feedback from the visual cueing group: Dr. Loschky, Dr. Bennett, Dr. Sanjay Rebello, Dr. DePaola, John Hutson, Tianlong Zu, and Xian Wu.

I would like to thank my trombone professor, Dr. Hunt. You have been a great mentor and friend. I am extremely grateful for your example of care and tutelage. Finally, I would like to thank my family. Without their continual encouragement and presence, I would be naught.

Chapter 1 - Introduction

Proof and its Reading

The fundamental purpose of mathematics education is to foster mathematical thought. Mathematical concepts and processes must be presented to and internalized by the students. In essence, mathematical content is conveyed via statement and evidence. Variance within can emphasize certain aspects of the mathematical process and thus has educational utility. Throughout the entirety of our mathematical exposure, examples have the capability to elicit novel insight and to promote keen understanding of concepts and processes. With a sufficient framework, individual intuition often formulates and completes logical connections. However, the most rigorous delivery is establishing a theorem via proof. And, here lies our academic and educational interest.

For various reasons the mathematical community does not have an overwhelming consensus on what constitutes proof (Weber, 2008), so a general conceptualization of proof structure serves as an entry point to the topic. A purported theorem asserts that, under certain suppositions, certain claims are true. A proof is an argument, composed as a sequence of statements, intended to establish such claims. These statements form a line of reasoning that provides suppositions, constructions, definitions, and claims which guide the reader through the meaning and content of the argument (Rav, 1999). In general most of these statements require justification. This justification, or warrant, invokes rules of logic and commonly known mathematical principles to draw the intended conclusions. In practice not all warrants are expressed explicitly (Weber and Alcock, 2005). Warrants may be left implicit for various reasons: perceived ease of inference (Davis, 1972; Inglis and Alcock, 2012), commonality of argument technique within the field (Thurston, 1994), or merely for brevity. Given readers' non-

standard exposure to mathematical content across all fields of study, the justification of implicit warrants is not universally accessible. Furthermore, in many educational settings, particularly proof courses, implicit warrants are not accepted as they do not explicitly demonstrate student understanding. Herein lays a portion of the aforementioned dissention.

The process one goes through to establish the correctness of a proof is proof validation. Requiring the evaluation of statements, the understanding of explicit warrants, and the necessary constructions in establishing implicit warrants, this process is quite complex (Selden and Selden, 2003). The precise methodology of individual validators is unclear. However, with verbal protocol and retrospective studies, particular actions within have been identified (Selden and Selden, 2003; Alcock and Weber, 2005; Weber, 2008; Ko and Knuth, 2013). As proposed by Weber and Mejia-Ramos (2011), three distinct interpretations of proof, as a whole, affect the validation process. When proof is taken as a cultural artifact, a validator defers authority beyond the mathematical content to author or publisher. When viewing proof as a sequence of inferences, a validator focuses on the sequential inferences generally resulting in a line-by-line validation process. A proof may also be seen as an application of methods. Operating from this standpoint, a validator evaluates logical chunks and overarching methods, and how they flow. These validator perspectives are not exclusive and may be switched between throughout a single reading. Furthermore, a separate characterization of validity judgments dichotomizes the process between a search for errors (negative characterization) and a reconstruction and subsequent understanding of the inference flow (positive characterization) (Inglis, Mejia-Ramos, Weber, Alcock, 2013).

Given the lack of a formal standard, one may expect there to be discrepancies among individual mathematicians regarding proof validity. This contends with the claims of uniformity

among mathematicians (Selden and Selden, 2003; Azzouni, 2004). Recent empirical studies, however, bolster the former idea that mathematicians do not exhibit uniform classifications when validating proofs (Weber, 2008; Inglis and Alcock, 2012; Inglis, Mejia-Ramos, Weber, Alcock, 2013). Despite the lack of uniformity, the reliability and validity of mathematics remain sound. Selden and Selden (2003) discuss mathematicians' keen ability to hash out differences to reach uniform conclusions. Furthermore, it is important to note that proof validation occurs not only individually but as a mathematical community. The communal confidence in a theorem develops through the occurrence of many individual validations and through the construction of varying proofs for the same theorem (Inglis, Mejia-Ramos, Weber, Alcock, 2013).

The construction of multiple distinct proofs for a single theorem not only serves to augment confidence in the theorem itself, but to provide novel insight regarding the particular structure within. Some mathematicians even question the value of proofs not containing new ideas or ingenious methods (Mejia-Ramos and Weber, 2013). Reading proofs in this manner, that is in order to gain comprehension of the ideas and underlying structure of the mathematical content, differs from the evaluative reading undertaken in the process of proof validation. None the less, both forms of reading are essential and prevalent in the mathematical community (Weber and Mejia-Ramos, 2011).

Given the utility the reading of proof has within the research-active mathematical community, its utility in mathematics education is manifest. The reading of proof ought to have a dual effect: an espousal of validity and an understanding as to why (de Villiers, 1990). For this reason when educators present a proof, they ought to ensure it is not "a formal demonstration that is devoid of insight" (Weber and Mejia-Ramos, 2014, p.3). The reading of proof thus enables students not only to learn the truth of mathematics, but to also understand why. This in

turn enables the systematization, discovery, and communication of mathematical theory (de Villiers, 1990). Through continual exposure to proof in the classroom and the textbooks, students learn to write, evaluate, and edit their own proofs. It is therefore no surprise that the reading of proofs constitutes a significant portion of the expected study regimens at the university level (Weber, 2004). Furthermore, within the realm of geometry, dynamic approaches to proof, which allow students to actively construct and manipulate figures, have been shown to aid in the realization of the various important roles proof plays in mathematics (de Villiers, 2004).

As discussed above, the reading of proof is integral in the formation of research-active mathematicians and remains as such throughout the entirety of their careers. In a culminating study across a variety of fields, the National Research Council (2000) analyzed and characterized the differences between novices and experts. Of the noted differences, four are of particular interest to proof validation.

Experts display pattern and pertinent information recognition to degrees beyond that of novices. This key difference speaks to the ability to identify relevant information effectively and efficiently. Furthermore, patterns, classes, or structures within the subset of relevant data are quickly recognized. From here the second difference may be fully utilized. That is, the experts' organization of knowledge reflects deep understanding. With a well-formed pattern in mind, experts utilize the big principles. Coupling patterns and principles, experts readily develop the context of the situation. Conditionalized knowledge is knowledge readily classified as useful or not, given the context. Since expert knowledge is conditionalized, quick application of useful knowledge follows. Finally, knowledge is accessible with minimal attentional effort. These four

differences have a synergistic effect making the difference between novices and experts all the more significant. Bridging this gap quickly and exhaustively is a primary pedagogical goal.

While mathematics helped inform the development of these general differences, the specifics and instantiations of these differences have not been thoroughly developed throughout the whole field. Significant progress in cataloging these particular differences in proof validation has been made. Undergraduate mathematics students perform at the level of chance when asked to validate proofs. Sadly, students who have just completed a transition-to-proof course focusing on validation or even an advanced mathematics course, continue to struggle as validators (Selden and Selden, 2003, 2015; Ko and Knuth, 2013). Research-active mathematicians display a more reliable, although not perfect, ability as validators. Students show an inclination toward equations and formula, and, thus, tend to look at the surface features of proofs (Selden and Selden, 2003; Inglis and Alcock, 2012). The attentional changes of expert mathematicians also differed from those of novices. Significantly more attentional changes are made between consecutive lines by experts than novices. This difference can be attributed to experts seeking out warrants (Inglis and Alcock, 2012), unlike novices (Alcock and Weber, 2005). An additional difference lies in the attentional change between different lines of the proof. Experts have been shown to make significantly more attentional changes between different lines of the proofs (Inglis and Alcock, 2012; Alcock, Hodds, Roy, and Inglis, 2015). Drawing connections throughout proofs rather than absorbing isolated surface features develops with mathematical maturity.

This dissertation advances both the understanding of the proof validation process and the differences displayed by novice and expert validators. It also addresses the disadvantages of the common methodologies of the validation literature by incorporating eye tracking, which has

been underutilized. The results of this research endeavor settle the contested results of the first direct comparison of novice and expert validators that used eye tracking (Inglis and Alcock, 2012). This dissertation continues testing the conclusions made from verbal protocols by means of eye tracking technology. Furthermore, it incorporates diagrams in the validation process, a feature that has gained minimal attention in the validation literature.

Cognitive Processing Behind Proof Validation

Advances in technology have enabled the advance of our understanding of both proof validation and the differences therein between novices and experts. Initially, studies relied solely upon self-reporting through either retrospective interviews or think-aloud verbal protocols. Now, however, the foundations of our understanding of proof validation benefit from the empirical approach of eye movement analysis. Since proof validation might not be an entirely conscious process (Selden and Selden, 2003), utilizing an approach that detects such subtleties has tremendous value. Furthermore the risk of behavior alteration in verbal protocols raises doubts about the results and conclusions. In cognitively demanding situations, especially when nonconscious activity plays a role, participant efficacy may drop. The notion is that the verbalization could reduce the cognitive resources available for the given task (Schooler, Ohlsonn, and Brooks, 1993). On the other hand, the act of verbalization may increase the development of satisfactory explanations, and thus increase efficacy (Chi, Bassok, Lewis, Reimann, and Glaser, 1989). A study in proof comprehension has shown the benefits of utilizing self-explanations given training (Alcock, Hodds, Roy, and Inglis, 2015). In the broader category of self-reporting, there is a two-fold risk. First participants may fail, intentionally or unintentionally, to report the essentials of their process. Secondly, they may report that which did not actually occur. The latter may occur without any malicious intent. Evidence legitimizing the

existence and occurrence of the above risks has been gathered and shown (Russo, Johnson, and Stephen, 1989). Eye tracking addresses these concerns.

Recording eye movements is a recent method utilized in various fields to gain insight into cognitive processes (Deubel and Schneider, 1996; Ball, Lucas, Miles, and Gale, 2003; Thomas and Lleras, 2007, 2009; Rayner 1998, 2009; Madsen, Larson, Loschky, and Rebello, 2012). Eye trackers record the movement of the eye(s) as a series of saccades (when the eyes are in motion) and fixations (when the eyes are stationary at a static location). The eye tracker records a plethora of measurements; most commonly used data include location, duration and order of the fixations and saccades. Other measurements include pupil dilation and blink occurrence. These measurements are then analyzed in order to understand participant cognitive processing.

Given our ability to think covertly, this method of insight requires justification. The approach is justified theoretically through the eye-mind hypothesis (Just and Carpenter, 1980; Rayner, 1998, 2009). Originally a model for reading, longer fixation durations were predicted to result from heavier processing loads. This hypothesis fit particularly well with actual reading patterns. So they theorized that a word must be fixated upon in order to process it and the fixation duration relates to the cognitive demands (Just and Carpenter, 1980). A deeper understanding of the eye-mind hypothesis has developed from the framework of overt and covert visual attention (Hoffman and Subramaniam, 1995; Kowler, Anderson, Dosher, and Blaser, 1995; Deubel and Schneider, 1996; Zhao, Gersch, Schnitzer, Dosher, and Kowler, 2012). The existence of a pre-saccadic covert attentional shift was a significant development. The significance of this result is in the implication that for each fixation there is a portion, albeit a small portion, during which covert attention and its associated processing do not align with fixation location. Attention allocation, however, corresponds to the point of fixation from the

onset of a fixation. Given the pre-saccadic covert shift occurs for each fixation and the relatively short duration of such, eye movement still gives an adequate window into the cognitive processes.

Research Questions

Given the level of incidence and importance of proof and its validation, mathematics educators have invested interest in fostering its mastery in their students. The key to bridging the gap between novices and experts lies in recognizing and responding to the differences in their respective processes. Droves of questions surround proof validation. As discussed above, interest is not novel so many answers have been attempted. Utilizing the analysis of eye movements, this dissertation addresses conflicting conclusions rising from the first direct comparison of expert-novice reading behavior during proof validation (Inglis and Alcock, 2012). This study also enhances the breadth of the proof validation literature by including elements of visualization. Our research differs in both method and intent from the previous studies that incorporated diagram usage (Ko and Knuth, 2013; Komatsu, Jones, Ikeda, and Narazaki, 2017). Visualization plays a significant role in the understanding and development of mathematics. Through the use of Euclidean geometry, a greater understanding of diagram utilization and expert-novice differences in proof validation is presented. How novices and experts allocate attention during proof validation is the focus of the research endeavor; this interest is specified through the following research questions. These research questions are revisited in depth at the end of chapter 2.

1. Does *zooming out* occur as a legitimate practice in proof validation?
2. Does overall validation performance vary by mathematical expertise?
3. Does the utilization of diagrams vary by mathematical expertise?

Content and Layout of Dissertation

This dissertation is structured into 5 chapters. The following chapter provides a thorough literature review. Certain topics previously given only a cursory introduction will be revisited at a more suitable length. Diagram usage in argument and problem solving is also addressed. The chapter concludes with a further exposition of the research questions. Chapter 3 explains the two part development and implementation of this research endeavor: a pilot study and an eye tracking experiment. The impact of the pilot study on the development of the eye tracking experiment is discussed. In chapter 4, analyses conducted on response and eye movement data are provided and related to the research questions. The final chapter presents the implications and conclusions. The limitations of the study follow; the dissertation concludes with possible directions for future research.

Chapter 2 - Relevant Literature

The literature review is presented in three segments. The nontrivial justification for the use of eye movement as a measurement of cognition is first presented. It is accompanied by an exposition of related analyses. Once the measures have been established we explore the relevant mathematical literature. Since we are using Euclidean geometry as a means to study both proof validation and diagram utilization, the remaining two sections discuss these two content areas.

Eye Movements and Visual Attention

The ultimate methodology in research attains the valid and reliable measurement of the unaltered processes of interest. Given the educational interest, these processes occur within living, learning participants. Memories are prone to alteration, so a natural emphasis is made on real-time measurement of the validation process. Recall two of the associated risks of self-report, especially for retrospective interviews: failure to report significant aspects of the process and recounting non-occurring processes (Russo, Johnson, and Stephen, 1989). Real-time verbal protocols remedy the temporal aspect of the risks, but introduce further confounding possibilities. The efficacy of the participant in the prescribed procedure may benefit from non-customary self-explanation (Chi, Bassok, Lewis, Reimann, and Glaser, 1989) or suffer from the limiting of cognitive resources (Schooler, Ohlsonn, and Brooks, 1993). Interviewers must additionally be cognizant of the active role they play and the risks therein. While methodology presents its own advantages and disadvantages, their weight varies tremendously based on a plethora of factors. Eye tracking is an alternative methodology which addresses many of the aforementioned issues, but, as in all methodologies it presents its own unique disadvantages. Used widely across various disciplines, the recorded eye movements grant access to the real time

processes of cognition (Deubel and Schneider, 1996; Ball, Lucas, Miles, and Gale, 2003; Thomas and Lleras, 2007, 2009; Rayner 1998, 2009).

The initial connection between eye movement and cognitive processes, proposed as the eye-mind hypothesis, assumes that in order to process a word one must first fixate upon the word and the cognitive processing demands are related to the duration of the fixation. This model fit well with actual reading patterns (Just and Carpenter, 1980). A greater depth of understanding regarding the model was attained through the study of attention. Overt visual attention refers to attending to the stimuli landing on the fovea, the portion of your retina responsible for high clarity eye sight. Aptly named, this sort of attention is prevalent and its recipient is readily recognized. However, overt visual attention is not the only type of attention; covert visual attention refers to the ability to attend elsewhere. Overt and covert attention are not unrelated, but the exact manner and extent of the relation is elusive (Kaspar, 2013). Significant work delving into this relationship provides the proper framework in which to view the eye-mind hypothesis.

With roots as far back as 1980 from Remington (Posner and Peterson, 1990), the pre-saccadic shift of covert attention acts simultaneously as a blessing and a curse for the eye-mind hypothesis. Given that readers overtly attend to the point of fixation at the onset of fixations (Glaholt, Rayner, and Reingold, 2012), a sequence of fixations provides an accurate attentional skeleton. Due to the pre-saccadic shift, however, the time spent overtly attending to the fixation is not precisely measured by the fixation time (Hoffman and Subramaniam, 1995; Kowler, Anderson, Dosher, and Blaser, 1995; Deubel and Schneider, 1996; Zhao, Gersch, Schnitzer, Dosher, and Kowler, 2012). The systematic occurrence of this shift and an allowance for a small variance in its duration provide an approximation for participant overt attention. This

approximation is thought to be particularly strong in the effortful tasks of reading and reasoning (Rayner, 1998, 2009; Deubel and Schneider, 1996; Ball, Lucas, Miles, and Gale, 2003). As with the disadvantages of real-time verbal protocols and retrospective interviews, the process of eye tracking inserts its own disadvantages. Participant motion is restricted throughout each recording session as head stabilization is often required. Additionally, knowledge of the recording in progress may alter the participant's processing.

Eye Movement Analyses

Having an understanding of the relationship between eye movement and cognition, we now look at the analyses eye tracking enables. Recall eye movements are recorded as fixations and saccades. As measured and reported by eye tracking technology, individual fixations are coordinate pairs indicating location. Similarly saccades are referenced by the source and target locations. By partitioning the images displayed on screen, areas of interest (AOIs) can be formed to serve in classifying individual fixations. In creating this partition, the researcher identifies the significant areas of each image. There is great flexibility in the shape and placement of each individual AOI. These AOIs enable various analyses based upon the significance of the areas. Saccades recognized as attentional changes between AOIs provide deeper understanding to how a participant digests the information in each area. Furthermore, fixation duration proportions are used to indicate which AOIs were significant to each participant. There is a large variety of analyses available depending on your research goals and interests.

Utilizing sequences of consecutive fixation locations and associated durations, eye movement can be compared for similarity. ScanMatch (Cristino, Mathôt, Theeuwes, and GilChrist, 2010) is an algorithm that given two these scan paths returns a normalized similarity score. The eye movement data inputted into this algorithm requires no alteration. The algorithm

creates a mesh over the screen from which the data originated and via a unique string of letters classifies each fixation accordingly. From here the temporal aspect of the data is incorporated by relating fixation duration with repetitions of the classification. Each scan path is now a sequence of these strings of letters. Similar to the method for classifying DNA sequences, the algorithm is based upon the Needleman-Wunsch algorithm. As to be expected similarity scores are commutative, and a similarity matrix can be formed running each unique pair through the algorithm. Individual scores or similarity matrices can be analyzed accordingly.

Research on Proof Validation

Interest in proof is not novel; there is an abundance of research relating to its construction, validation, and comprehension. However, our interest lies in the proof validation process. A general exposition of the role of validation is first provided. Research on the general strategy of validation follows. This section will end by reporting the series of articles which served as a primary motivation for this research.

The role and standard of proof validation

Considering a proof as a text intended to establish the veracity of a purported theorem, the evaluative reading, pondering, and processing of the text is called proof validation. There is an onus on the reader, to a certain degree, to elicit meaning from the text itself. With this in mind, validators may participate in the tasks of recalling or researching pertinent mathematical knowledge, understanding claims and justifications, the construction of subproofs, and surveying the overall argument. The educational background of the validator influences the facilitation of many of these tasks. Given different readers or even different readings, proof validators may therefore glean insight in various ways from a single proof (Selden and Selden, 2003). These

confounding factors contribute to the absence of a universal consensus on what constitutes proof (Weber, 2008).

Despite this lack, however, the claim that mathematicians display high degrees of unanimity in validation assertions has a presence in the literature (Selden and Selden, 2003; Azzouni, 2004). There are doubts to this claim which have been tested and supported using empirical methods (Weber, 2008; Inglis and Alcock, 2012; Inglis, Mejia-Ramos, Weber, Alcock, 2013).

The source of this discrepancy could lie solely in the individuals or could also stem from differences in expertise. For example, a study comparing validation assertions on a single purported proof from undergraduate calculus found that applied mathematicians were more inclined to deem the proof valid than pure mathematicians (Inglis, Mejia-Ramos, Weber, Alcock, 2013). Selden and Selden (2003) discussed the role of differing opinions in validation assertions. In their experience, partial or complete expert joint validations generally resulted in agreeable conclusions. In certain cases, minor flaws that had no weight on the overall argument were identified and resolved the discrepancy. This proclivity to resolution was in part tested (Inglis, Mejia-Ramos, Weber, Alcock, 2013) by presenting specific objections to validators regarding their assertion. These objections were followed by two questions asking about the reasonableness of the objection and if they rendered the proof invalid. The study found that validators who first asserted the validity of the proof usually retained that assessment even after reading the provided objection, and they concluded that these findings suggest more than one standard of validity among mathematicians (Inglis, Mejia-Ramos, Weber, Alcock, 2013). This testing procedure does not replicate the genuine interaction between mathematicians or approximate the discourse of joint validations as described by Selden and Selden (2003). Nonetheless, the uniformity of

individual mathematicians in their validity assertions is not to the degree of unanimity. In a conjectural work, Dawkins and Weber (2017) construe proof in terms of mathematical values and associated norms. They argue that differences in individual mathematical values provide a source for the discrepancies in practice. Validators without a sound understanding or acceptance of a given mathematical value may not fully adhere to its associated norm which leads to discrepancy. Additionally it is possible that the set of values and hence norms differ between applied and pure mathematics.

Within the mathematical community, proofs are studied widely and undergo many validations, individually and communally. These joint validations expand, identify, and resolve issues more precisely (Selden and Selden, 2003). Furthermore, the mathematical community works tirelessly to whittle arguments down to their elementary aspects until they become mere trivialities (Rota, 1991). These procedures bolster confidence in the veracity of both theorem and proof. Independent proofs are generated throughout the study of theorem and proof alike. While not only securing confidence in the theorem (Inglis, Mejia-Ramos, Weber, Alcock, 2013), these independent proofs provide further insight about the inner workings of and connections between mathematical structures.

Methods and characterization of proof validation

Recall the conceptualization of proof as a sequence of statements establishing the claims of a purported theorem (Rav, 1999). These statements in general require a warrant, a form of justification that enables the reader to conclude that the claims within the statement logically follow from the previous statements, assumptions, and other relevant mathematical knowledge. As the justification used in implicit warrants is not expressed, they may be seen as gaps in the proof. Readers are left with the formulation of the argument, graders question the presence of the

gap, and validators are left with a combination of both. Explicit warrants vary in thoroughness and may lead to further formulation on the part of the reader. The general conceptualization of proof does not compel particular methods of proof validation. These methods are not fully understood or even entirely known. Nevertheless, studies explore these mental processes and provide guidance for their continued study.

Through a series of interviews (Weber and Mejia-Ramos, 2011) and subsequent surveys (Mejia-Ramos and Weber, 2013) with research-active mathematicians, three general strategies of proof validation were formulated: proof as a cultural artifact, proof as a sequence of inferences, and proof as an application of methods.

This first strategy primarily considers the implications of published proof. A published proof contains a historical context beyond the text itself. This context includes the associated requirements for publishing and the entities associated with publication. The peer-review process ensures the occurrence of validations by mathematicians, presumably, with significant expertise in that particular field. Furthermore, the clout within the community, of both the journals and the mathematicians authoring the proofs, may produce enough confidence to conclude the theorem and proof are true.

Viewing proof as a sequence of inferences, the second proposed strategy, aligns well with the aforementioned conceptualization of proof. Viewing each warranted statement as an inference, the proof's validity is verified by confidence of each inference or nullified by a flawed inference. Confidence is not necessarily built through thorough argument. Validators may target the problematic inferences and analyze them in depth in order to build confidence in the proof. They then spend less time investigating the remaining inferences. Checking examples, failing to find a counterexample, and noticing a pattern could bolster confidence in either of these

situations (Weber and Mejia-Ramos, 2011). This method, commonly referred to as line-by-line reading, may also be referred to as the *zooming in* approach.

With as many as half of published proofs containing flaws, not necessarily vital, to the argument (Davis, 1972), a counterpart of *zooming in* forms the third strategy, the application of methods or *zooming out*. As Thurston expressed, "People are usually not very good in checking *formal correctness* of proofs, but they are quite good at detecting potential weaknesses or flaws in proofs" (1994, p.169, italics are Thurston's emphasis). Assessing proof via logical components constituted a natural validation technique. The process itself is aided in part by the visual cues and partitions generally present in the written form of the proof (Weber and Mejia-Ramos, 2011). An additional think-aloud study illustrated the occurrence of initial structural overviews of number theoretic proofs. This result could be attributed to blatant structural errors and content familiarity. Thus it is possible that this familiarity enabled immediate identification of the argument structure (Ko and Knuth, 2013).

A characterization of proof validation independent of the aforementioned strategies dichotomizes the process (Inglis, Mejia-Ramos, Weber, Alcock, 2013). The positive characterization of an individual validation reconstructs and establishes the sequence of statements forming the proof. The negative characterization reduces the validation process as an endeavor to identify errors. In order to test these characterizations, a nontrivial premise is relied upon. The supposition asserts that, in both characterizations, a comparable and direct relationship exists between confidence and goal achievement. The goal is the completion of the task according to your characterization, either finding an error or reestablishing the argument. Higher participant confidence in invalid assertions and the ability to readily justify them as such resulted

in the conclusion that negative characterizations serve as the primary characterization of the individual validation process.

We now express doubts about the validity of this test. The claim that positive characterization validations concluding with invalid will lack "justification beyond reporting that they failed to reconstruct the proof" (Inglis, Mejia-Ramos, Weber, Alcock, 2013, p. 273) is lacking. This is the premise for a decreased confidence level in validators asserting invalid from a positive characterization of proof. The finding of a logical error, however, is grounds for the cessation of argument reconstruction. Therefore, high levels of confidence should be attainable and expected with either characterization. Secondly, there is the issue of implicit warrants; allowable gaps are not standard. We argue that for any particular theorem, the statement of the suppositions followed by the conclusion of the theorem would not constitute a proof. Thus there exists a cut off, albeit not specified, of implicit warrant gap size which universally merits an assertion of invalid. Again high confidence from either characterization would be reasonable. These faults render the confidence-goal test invalid and the conclusions of such unjustified.

An eye tracking experiment of proof validation

The primary impetus to this current research endeavor follows an eye tracking study conducted by Inglis and Alcock (2012). This particular study received several significant response articles (Weber, Mejia-Ramos, Inglis, and Alcock, 2013), and the eye-movement data was used again, in part of a study published later (Alcock, Hodds, Roy, and Inglis, 2015). This series of articles is now discussed at length.

The initial article had three primary objectives: to empirically test the uniformity of individual proof validation at the expert level, to analyze the differences between novices and experts in attention allocation during the validation process, and to collect evidence supporting or

denying the occurrence of *zooming in* and *zooming out*. The first interest has already been discussed at length.

Novice participants were successful undergraduates having taken proof-based courses. The experts were all research-active mathematicians. Six proofs were presented for validation. The first four were elementary number theoretic proofs presented as student produced. These were the same proofs as used previously in the literature (Selden and Selden, 2003; Weber, 2008). The last two proofs were presented as proofs submitted to some mathematical journal. These proofs were more novel regarding usage in the study of validation; the number theory proof was used by Weber (2008), and the remaining proof was from undergraduate calculus. As hypothesized by Selden and Selden (2003), the novices spent significantly more time, proportionally, dwelling on the surface features of the proof. The notational and computational features of each proof written in LaTeX were collectively measured and classified as surface (Inglis and Alcock, 2012).

Based upon the claims from Weber (2008) that mathematicians take a precursory read through a proof to get an overview of the proof structure, Inglis and Alcock (2012) formulated a measure to test for *zooming out*. It was argued this initial read through, meant primarily to get an overview for the proof structure, would not constitute a significant portion of the validation duration. The initial reading (IR) ratio was computed by taking the proportion of elapsed time until first fixation on the last line to the total time of validation. The resulting mean of IR ratios were not significantly smaller than 50% as one would expect if an initial reading occurred. It was thus concluded that initial reading for overall structure did not occur (Inglis and Alcock, 2012).

Continuing to evaluate their third research question, eye movement analyses testing paths of attention were developed. *Zooming in* was characterized by a sequential reading order with

zooming out displaying non-sequential reading orders. Attentional changes were classified as within-line and between-line. Mathematicians displayed significantly more between-line saccades than novices. Upon further analyses, the difference stemmed from a greater number of back and forth saccades between consecutive lines. Investigation of warrant seeking (counts of line fixation sequences of the form $n \rightarrow n-1 \rightarrow n$ where line n required a warrant) on the novel number theoretic proof revealed significant differences between test groups. Both groups displayed the ability to recognize the need for warrants. The researchers thus concluded that mathematicians display greater between-line saccades through an attempt to infer warrants and connect consecutive lines in the proof. They acknowledged several key limitations to the analysis: implicit warrants can include multiple, non-consecutive lines or they may lie within a single line (Inglis and Alcock. 2012; Alcock, Hodds, Roy, and Inglis, 2015).

The response articles

Weber and Mejia-Ramos responded to the assertions made by Inglis and Alcock (2012) with a critique regarding the initial reading of proofs. This was followed by a rebuttal from Inglis and Alcock; both articles were published together (Weber, Mejia-Ramos, Inglis, and Alcock, 2013).

The initial critique by Weber and Mejia-Ramos raised two primary concerns. The first concern centered on the averaging of data across different tasks. Averaged data was utilized through several of the analyses presented by Inglis and Alcock (2012). The particular analysis in question, however, was the averaging of IR ratios across separate tasks. The strategy of a precursory read through for proof structure was denoted as the initial skim strategy. As argued by Siegler (1987), when attempting to identify strategies, the averaging of data across different tasks and trials can distort conclusions regarding the many aspects of individual performance. The

second concern is related to the validity of the IR ratio in testing for a quick initial read through for proof structure. The brevity, simplicity, and familiarity of the proofs studied may render the need for further examination pointless. In such a situation, an IR ratio close to 1 would not be unreasonable. Given the blatant errors present in the studied proofs, a validator may identify the error and upon completion of the initial reading assert the invalidity of the argument. This again would result in a large IR ratio despite the fact that initial skim-reading was the strategy employed. Weber and Mejia-Ramos exemplified their critiques through a reasonable hypothetical. Take a validator, who on three of the proofs found a blatant error and designated the proofs invalid. In each case, an IR ratio of around 0.9 would not be unexpected. On the remaining proofs, suppose this validator implemented an initial skimming strategy in order to identify proof structure. IR ratios centering on 0.2 would not be unexpected. Aggregation of these IR ratios would result in 0.55 indicating non-utilization of initial structural reading (Weber, Mejia-Ramos, Inglis, and Alcock, 2013).

Weber and Mejia-Ramos offered the utilization of distributions as a remedy for this issue. Having been granted access to the eye movement data by Inglis and Alcock, Weber and Mejia-Ramos looked at the distribution of the aggregated IR ratios to discover a bimodal nature supporting the possibility that the initial skim-reading strategy occurred in practice. Further analysis was conducted on the proof deemed most likely to have skimming occur. This analysis found that 8 of the 12 mathematicians had IR ratios of less than 0.35. At which point Weber and Mejia-Ramos concluded initial skimming as a strategy in practice is supported (Weber, Mejia-Ramos, Inglis, and Alcock, 2013).

The final rebuttal from Inglis and Alcock addressed two issues with response of Weber and Mejia-Ramos. The original IR ratio test was a traditional by-subjects analysis, meaning the

distribution of interest was the distribution of IR ratio means, which was found to be normal. Thus they assert their test was indeed valid; it just does not conclude the strategies implemented on individual proofs. They argue analyzing the single distribution of IR ratios formed by aggregating all the task-individual pairs violates independency. Additionally, they argue the presence of early fixations at the end of the proof could be meaningless without further inquiry. They may occur after a blink or through a head movement. The absence of a fixation until a certain point, however, ensures no attention was paid at least until that particular point in the validation. Applying this critique would render their support of practical use of an initial skim strategy unjustified (Weber, Mejia-Ramos, Inglis, and Alcock, 2013). This series of articles exploring proof validation confirmed many conclusions drawn from verbal protocols and retrospective interviews, but it left questions regarding the general methods of proof validation. In order to build an initial framework of diagram usage in proof validation, we visit diagram usage in a closely related field of study: problem solving.

Diagram Usage in Problem Solving

Diagrams and other visualizations play a major role in understanding and formulating mathematics. However, diagram utilization in the realm of proof validation is not well understood. Yet within the realm of proof construction, the utilization of visual arguments has merited significant research (Zazkis, Weber, and Mejia-Ramos, 2016; Komatsu, Jones, Ikeda, and Narazaki, 2017). Substantial work studying diagram utilization within problem solving has been conducted. Going beyond understanding just the key differences in the utilization, key research has been conducted in cueing to facilitate learning (Madsen, 2013; Rouinfar, 2014; Agra, 2015; Wu, 2016). The body of work studying diagram usage in problem solving provides insight into the mental processes taking place during the validation process.

The evaluative process of proof validation is vital to the construction of rigorous proofs. Continual evaluation of the construction must occur to insure the production of a satisfactory proof. Working with secondary school students taking geometry, proof validation within proof construction was studied (Komatsu, Jones, Ikeda, and Narazaki, 2017). Exploration of the validation process was conducted through local counterexamples (counterexamples contradicting certain lines not the purported theorem). Students had to identify the counterexamples and then modify their argument to account for them. Constructing diagrams within the parameters of the proof and justifying warrants constituted the difficulties exhibited by these secondary students.

Two types of visual inferences may be made in the context of a proof with associated visualizations. Perceptual inferences are based upon the appearance of the visualization. A property of the visualization, say a graph of a function, is recognized and then applied to the formal entity, say the function itself. Deductive inferences are based upon the necessary characteristics of visualizations containing a particular property. Through acknowledging some visual property of a formal entity, a further visual property may be asserted. The majority of mathematics majors reject the use of perceptual inferences in proof, while permitting the use of deductive inferences (Zhen, Weber, and Mejia-Ramos, 2015).

With a close proximity to proof validation, the persuasiveness of visual argument was studied using Young's Inequality, a theorem from undergraduate calculus (Inglis and Mejia-Ramos, 2008). The study, consisting of two experiments, compared novice and expert persuasion using solely visual evidence versus the same visual evidence with descriptive text. The text merely described the image; no supporting arguments were included. It was found that the accompaniment of the descriptive text increased the persuasiveness of the argument in both groups. The second experiment ensured that the results of the first were not due to the text

drawing attention to the key features of the visualization. Arrows indicating the key features were now included in the non-description presentation. The results from the first experiment were replicated in the second. Overall, however, novices tended to be less convinced by the visual evidence than the experts. This indicates a general skepticism about visual argument among undergraduate students that is less present in the research community.

The interplay between this general skepticism and the presence of visualizations is explored in a study about the effects of illustration in solving true-false problems from vector calculus (Nyström and Ögren, 2012). While the presence of the illustrations played no significant role in the overall performance, the presence of the illustration increased the likelihood of participants asserting the presented problem is indeed true. So while students display a general skepticism to relying solely on visualization, the presence of visualization, regardless of the true nature, bolsters their belief in the statements.

Another related study examined the use of tabular data in the context of college algebra utilizing eye tracking (Johnson, 2015). The study discovered that higher dwell time on relevant values in the tables was indicative of correct solvers. Incorrect solvers had a tendency to spend more time on table labels, but most significantly a notable group of students never went beyond the table labels in the course of solving the problem. This work, in part, stemmed from work conducted in the realm of physics education.

Originally interested in the differences between novice and expert approaches to diagram usage in conceptual physics problems, physics educational researchers have studied these differences in correct and incorrect solvers. Similar to the findings of Johnson (2015), correct solvers of conceptual physics problems spent significantly more time on the thematically relevant areas of the diagrams (Madsen, Larson, Loschky, and Rebello, 2012; Madsen, 2013).

Using ScanMatch (Cristino, Mathôt, Theeuwes, and GilChrist, 2010) similarity scores, comparisons within and between solver types (correct and incorrect) were made. Significant differences in the distributions of these comparisons indicate overall differences in scan paths by solver type. There, however, were instances where incorrect-correct comparisons scored higher than the incorrect-incorrect or correct-correct comparisons. This indicates that the scan paths of some incorrect solvers were closely related to those of correct solvers (Madsen, 2013).

Research Questions Revisited

Having presented the details and subtleties of the literature, we revisit the research questions. This exposition will provide greater depth and present the hypotheses tested in chapter 4.

Does *zooming out* occur as a legitimate practice in proof validation?

The evidence for the practice of *zooming out* proffered via verbal protocols needs verification and extension through various means of exploration (Selden and Selden, 2003, 2015; Weber and Mejia-Ramos, 2011, 2013; Ko and Knuth, 2013). Initial skimming constitutes a single method from the *zooming out* strategies; one cannot assert the absence of *zooming out* by testing a single method, only the absence of that particular method. Yet there is merit in the IR ratio, as developed by Inglis and Alcock (2012), for testing the claims of an initial structural read through as asserted by Weber (2008).

The valid critiques of the measure, however, must be addressed. Given proofs with sufficient complexity and length without blatant errors from a relatively unfamiliar domain, the IR ratio, when analyzed from the distributional perspective, validly measures the occurrence of an initial skim. Increased complexity and length combined with decreased familiarity (Weber, 2008; Ko and Knuth, 2013) address the issue of validators recognizing and confirming proof

structure without effortful processing. Furthermore the absence of blatant errors prevents high IR ratios resulting from quick invalidation following an initial read. Remedying the familiarity concern associated with the IR ratio was the primary concern. Complexity, length, and proper errors could be achieved in almost any mathematical domain.

Schoenfeld hit upon the solution in referencing Euclidean geometry: "After a 10 year hiatus, many of the specific facts and procedures that the faculty once knew have faded from memory" (Schoenfeld, p. 3, 1985). Research-active mathematicians could previously compare the structure with already known valid arguments that establish the results and not actively engage in the validation process. So this addresses an additional threat to the validity of measuring expert validation processes. As Thurston wrote, "When the idea is clear, the formal setup is usually unnecessary and redundant" and "People familiar with ways of doing things in a subfield recognize various patterns … for certain concepts or mental images" (Thurston, p.7, 1994). It is further noted that "the students will remember more of the basics and thus have the initial advantage over the faculty" (Schoenfeld, p. 3, 1985).

Thus by using suitably complex proofs from Euclidean geometry without blatant errors, we test for the strategy of initial skim-reading. Each individual theorem is tested separately. Both the mean IR ratio and the distribution will be analyzed in order to reach a conclusion. A conclusion of initial skim-reading occurring would require further eye movement analysis to confirm that the fixation occurred while reading the end of the proof and not by accident.

The *zooming out* strategy may present itself in various implementations. The validation process concluding in an assertion of invalid may provide the insight guaranteeing usage of *zooming out*. Think-aloud studies have identified the major strategies of both novices (Selden and Selden, 2003; Ko and Knuth, 2013) and experts (Weber, 2008). Termination of the proof

validation upon identifying an error is reported as a strategy implemented at both levels. Selden and Selden (2003) reported that novices used the mental techniques of rumination and rethinking before drawing firm conclusions regarding these particular statements. Experts resolve uncertainties through examples and deductions made within the confines of the proof setup. They, however, never explicitly mention these utilizing of the remainder of the proof. The closest behavior discussed in the literature lies in the critique of the IR ratio by Mejia-Ramos and Weber as discussed earlier (Weber, Mejia-Ramos, Inglis, and Alcock, 2013). It is possible that this sort of behavior just went unnoticed or was not reported by participants. However, once identifying a flaw or questioning a conclusion and continuing through the remainder of the proof may serve a couple of purposes. The first purpose is to confirm the flaw. The second purpose is to gauge the importance of that conclusion. Both purposes are implementations of *zooming out*. This is addressed in greater detail in chapter 4.

Research Question 1: Does *zooming out* occur as a legitimate practice in proof validation?

1. IR ratio and distribution - competing hypotheses
    a. Hypothesis: Given the longer, more complex proofs not containing blatant errors, initial skim-reading, as proposed by Weber (2008) and classified as *zooming out* (Mejia-Ramos and Weber, 2011, 2013), is generally implemented in the proof validation process.
    b. Hypothesis: Initial skim-reading is not generally implemented in the proof validation process thus corroborating the claims of Inglis and Alcock (2012).
2. Eye movement after encountering flaw or doubt - competing hypotheses
    a. Hypothesis: Given the absence of continued forward progression after errors in the think-aloud proof validation literature (Selden and Selden, 2003; Weber,

2008; Ko and Knuth, 2013), continued forward progression after encountering an error will not generally occur.

b. Hypothesis: As alluded to by Mejia-Ramos and Weber (Weber, Mejia-Ramos, Inglis, and Alcock, 2013), continued forward progression after encountering an error will generally occur.

Does overall validation performance vary by mathematical expertise?

Many aspects of the validation process vary with mathematical experience. Mathematical maturity effectuates overall competency and has been clearly established (Selden and Selden, 2003, 2015; Alcock and Weber, 2005; Inglis and Alcock, 2012; Ko and Knuth, 2013). Experts display the ability to focus on relevant material while novices readily attend to surface features (Selden and Selden, 2003; Inglis and Alcock, 2012). Furthermore, mathematical maturity yields increased warrant seeking patterns (Inglis and Alcock, 2012, Alcock, Hodds, Roy, and Inglis, 2015). These differences have been identified by measuring particular aspects of the validation process without any overall temporal or sequential reference. This is not to negate the importance of identifying these differences in particular actions. It is an acknowledgement that the presence of these differences does not imply a fundamental difference between the validation processes of experts and novices. Whether these differences result from different validation processes or skill iniquity is an open question. We seek to address this question through a direct comparison of scan paths, which encode entire validation processes. The basis of this comparison is the ScanMatch algorithm implemented through its toolbox in MatLab (Cristino, Mathôt, Theeuwes, and GilChrist, 2010). For further analysis, these normalized similarity scores can be converted to dissimilarity scores. Given pairwise dissimilarity scores, algorithms, like k-medoids and agglomerative nesting, provide key information about process proximities.

Dissimilarity within and between groups provide additional information relating the different validation processes. This set of analyses is conducted at two levels: the purported theorem level and the experimental level. At the experimental level, dissimilarity scores are viewed as five dimensional dissimilarity vectors. Through the application of a norm, an overall dissimilarity score is computed for use in the above analyses.

Research Question 2: Does overall validation performance vary by mathematical expertise?

1. ScanMatch within-group and between-group dissimilarity score comparisons (conducted at two levels)
    a. Hypothesis: Given experts' warrant seeking behavior (Inglis and Alcock, 2012) coupled with a greater trust in visual arguments (Inglis and Mejia-Ramos, 2008), within-group disparity will be significantly smaller than between-group disparity.
2. ScanMatch clustering (conducted at two levels)
    a. Hypothesis: Given experts' warrant seeking behavior (Inglis and Alcock, 2012) coupled with a greater trust in visual arguments (Inglis and Mejia-Ramos, 2008), clusters will be homogeneous with respect to mathematical maturity.
3. Statement page duration
    a. Hypothesis: Due to an inclination to construct their own outline of a proof, experts will spend significantly more time on the statement screen.

Does the utilization of diagrams vary by mathematical expertise?

Utilizing Euclidean geometry on the basis of familiarity also introduces the use of diagrams. Diagram usage in the realm of proof validation has received little attention. In an attempt to study validation and counterexample across various domains of mathematics, geometry has played a small role (Ko and Knuth, 2013). Geometry has received more attention

in proof construction and comprehension (de Villiers, 2004; Komatsu, Jones, Ikeda, and Narazaki, 2017).

The strides made regarding diagram usage in problem solving will help us frame diagram usage in proof validation. We will briefly review the contributing results. Despite having an overall skepticism of visual argument (Inglis and Mejia-Ramos, 2008), along with a possible aversion to its use in college algebra (Johnson, 2015), students are more inclined to believe statements given an illustration (Nyström and Ögren, 2012). Furthermore, students reject graphical perceptual inferences while accepting graphical deductive inferences (Zhen, Weber, and Mejia-Ramos, 2015). Participants spending proportionally more time on thematically relevant areas are more likely to answer correctly (Madsen, Larson, Loschky, and Rebello, 2012; Madsen, 2013; Johnson, 2015).

Research Question 3: Does the utilization of diagrams vary by mathematical expertise?

1. Proportion of fixation time on diagram and text
    a. Hypothesis: Given the skepticism of visual arguments displayed by novices and their general focus on equations and symbols (Selden and Selden, 2003; Inglis and Alcock, 2012), experts will spend significantly more time on the diagram while novices spend significantly more time on the text.
2. General attentional changes including the diagram
    a. Hypothesis: Having a greater trust in visual argument (Inglis and Mejia-Ramos, 2008) and appreciating a deeper understanding of mathematics (NRC, 2000), experts will make significantly more attentional changes within the diagram and between the text and diagram.
3. Warranted attentional changes

a. Hypothesis: As supported by Inglis and Alcock (2012), experts display warrant seeking behavior and are more willing to trust visual argument (Inglis and Mejia-Ramos, 2008); this will result in experts making significantly more attentional changes from warranted lines of the proof and also between warranted lines and the diagram.

Chapter 3 - Methods

Pilot Study

Euclidean geometry presented advancement in the study of proof validation through diagram usage and by remedying some of the shortcomings of previously provided proofs. We sought to provide proofs that provided enough complexity and novelty to warrant sincere validations that were not immediately nor readily reduced to recollection or obvious error. Given the limited resources pertaining to proof validation in Euclidean geometry, the construction of our purported theorems and proofs needed to be tested. We therefore conducted a pilot study to test that our proposed proofs elicited these sorts of active validations. Various measurements were implemented to facilitate the experimental design of the eye tracking study. The pilot study was run to also discover subtleties deserving fuller exploration.

Participants

The pilot study was conducted on 13 participants during the summer of 2016. Due to the small pool of possible participants, the pilot study was conducted on participants unlikely to be able to participate in the eye tracking study. This was discerned through two criteria: soon-to-be departure from the area or necessity of prescription glasses. The later criterion stems solely from possible issues obtaining accurate tracks with the eye tracking machine given strong corrective lenses. Participants were recruited via email or through conversation in person and then classified by mathematical experience. This classification consisted of four tiers: novice undergraduate (2 males), early graduate student (2 females, 3 males), late graduate student (4 males), and faculty (2 males). To qualify as a novice undergraduate, the participant needed previous or concurrent enrollment in an undergraduate level proof-based course. Participants were classified as early graduate students via one of two ways: enrollment as a graduate student

in mathematics or enrollment as an undergraduate having successfully taken more than one graduate course. Late graduate students were those participants who had passed qualifying exams and were actively working on research. Faculty consisted of research-active professors and postdoctoral researchers. Contrasting the traditional expert/novice classifications, the intent of this finer striation was to gauge the maturation of mathematical ability. Participants were paid $15.

Materials

The materials consisted of seven purported theorems from Euclidean geometry. Each purported theorem had three associated pages. The first page consisted of the theorem statement with a diagram visualizing the theorem statement. The second page repeated the theorem statement, provided any necessary lemmas, and presented the purported proof with the associated diagram(s). The final page was a replica of the second with the addition of answer prompts. Progression through any of the pages never resulted in the loss of available information.

Purported proof and theorem development

Considering this was preparation for an eye tracking study, the purported theorems and proofs were designed to fit the parameters of such a study. The three phase presentation of the purported proof was motivated by several factors. Having a statement page devoid of argument provided the participant an opportunity to understand or study the statement without influence. Furthermore, presenting the proof across several pages would not promote the most natural validation processes. Pertinent information needed to be readily available at all times. Continual flipping between screens would elicit unnecessary noise. And finally, the experimental design minimized the risk of unexpected complications and simplified data analysis. Balancing both the experimental goal of providing longer, more complex proofs and this experimental design was a

primary concern. Furthermore, utilization of the entire screen was not optimal as eye tracking accuracy wanes along the borders, near the monitor casing. Considering all of these factors, proofs were written in a fairly terse style. This terse style was accomplished by leaving certain justifications implicit. This method is both necessary and common in mathematical practice (Alcock, Hodds, Roy, and Inglis, 2015). Our terse style required frequent use of implicit warrants. To address the possible negative consequences of this style (Alcock and Weber, 2005; Komatsu, Jones, Ikeda, and Narazaki, 2017), purported theorems and proofs were developed to minimize the use of any advanced geometric arguments. Less memorable arguments were addressed by providing lemmas which were to be understood as established and true. Our aim was to leave no implicit warrant without a reasonable means of establishing it.

The purported theorems and associated proofs were developed with reference to geometry textbooks (Davis, 1949; Smart, 1998). Four of the purported proofs were intentionally constructed to be invalid. Blatant errors within the body of the proof were avoided. The logical errors consisted of not justifying a statement whose justification may use the theorem statement and failing to prove the entirety of the theorem statement. The latter error occurred in three different fashions: failure to prove an if-and-only-if statement, failure to prove a direct assertion of the theorem, and failure to realize the argument is degenerative in a particular case included in the theorem statement. It was thought that the recognition of this latter sort of errors resulted primarily from a *zooming out* perspective.

Diagrams were constructed with two central tenets. Firstly, diagrams were constructed using GeoGebra (https://www.geogebra.org) using geometric construction techniques. While they were constructed to show desired cases, they were not manipulated to hoodwink participants visually. Secondly, construction aimed to facilitate progress through each proof. Angle and

polygonal reference order tried to promote a continuity of the references and continued progression. References were not intended as points of confusion or obstacles. Congruencies, found in both the assumptions and the proof, were notated in the diagram. The text was structured so as to provide a visual connection or separation of statements. See Appendix A for the purported theorems and proofs used in the pilot study.

Table 3-1 Pilot study: list of purported general theorem statements with intended logical issues

Purported Theorem	General Statement	Intended Proof Issues
Practice	If a diameter bisects a chord, then it is perpendicular to the chord.	Argument degenerates and is indeed false when the chord is also a diameter.
Angle Bisector	An internal angle bisector divides the opposite side length proportionally to the adjacent sides.	None.
Angle Isosceles	Given an isosceles triangle, the sides opposite the equal angles have equal lengths.	Stating that an angle bisector intersects the opposite side perpendicularly requires justification. Justification may use the theorem statement
Bisector Isosceles	If two internal angle bisectors are equal, then the triangle is isosceles.	None.
Inscribed Angle	An inscribed angle is equal to half of the central angle.	None.
Miquel Point	If an arbitrary point is taken on each side of a triangle, the three circles determined by each vertex and the two points on the adjacent sides have a common point of intersection within the triangle.	Proof fails to prove the point lies within the triangle, which is indeed false.
Ptolemy	A quadrilateral is cyclic if and only if the product of the diagonals is equal to the sum of the products of the opposite sides.	None.

Procedure

Each participant partook in an individual session lasting between 45 and 75 minutes. To ensure time commitments were not exceeded, participants did not initiate any new validations after 60 minutes. Despite having not completed all of the validations the remainder of the study was conducted following the usual format. Each session was conducted in the Center for Quantitative Education and both audio and visual were recorded. The video camera was focused on the computer screen running the experiment. The experiment was built using Experiment Builder (http://www.sr-research.com/index.html). The sessions consisted of three parts: the instructional phase, the validation phase, and a brief concluding interview.

After informed consent was obtained, participants were instructed to think aloud throughout the entirety of the session. A brief discourse ensued ensuring the participant's understanding of proof validation and the structure of the session. They were instructed to take as long as they needed to reach informed decisions. They were further told that breaks would be available between each theorem. The instructional phase concluded with an example and a last opportunity to ask questions. The example did not vary by participant.

The proof validation phase consisted of six proof validations. The seventh proof served as the example provided in the instructional stage. The six purported theorems and associated proofs were split evenly by intended validity assertions. We thus provided three proofs meant to be asserted as valid and three meant to be deemed invalid. In order to address order bias, the 6 purported theorems were presented in varying orders determined by a series of Latin Squares. As mentioned before each purported theorem had three associated screens: statement, proof, and answer. Participants progressed through each screen at their own pace. See Appendix A for an example of the three screens. They were told that once they reached a decision regarding a

proof's validity and were ready to answer, they needed to progress to the answer screen immediately. Once on the answer screen, the participant indicated their assertion of valid or invalid via a keyboard press of V or N. They were instructed to then justify the conclusion. The entirety of the proof and statement were available for reference during this justification. If needed, the interviewer would seek clarification. No performance feedback was given during the session.

Each session concluded with an interview seeking further knowledge about diagram usage, validation processes, consistency, and attention allocation. See Appendix B for the pilot study protocol. This protocol provides a complete list of questions used in the interviews. If interested in performance feedback, it was supplied after the interview was completed. Participants were then paid and led back to the hall.

Refinements and prompts for the eye tracking experiment

The analysis of the pilot study began with a complete transcription of each session. Transcriptions were analyzed seeking to identify concerns and struggles. These comments were classified and tallied by theorem and were then incorporated into future design. Addressing these comments resulted in the alterations to several proofs to ease flow and remove undesired struggles. It also led to the incorporation of an instructional page that provided certain troublesome definitions and symbols.

An analysis of the assertions followed. As previously mentioned, we constructed the proofs intending to elicit a particular assertion. These validity assertions were measured in two manners: first by how well they matched our intended responses, and second by judging the reasonableness of their justifications. This latter judgement referenced passages from the transcript where participants were working through their cited problems or the logical flaws we

knew to exist. The effectiveness of the purported theorems was assessed by comparing the two assertion measures. Large discrepancies between the two indicated the need for further review.

Table 3-2 Overall assertion measures by purported theorem from pilot study

[†]See Appendix A for details on each purported theorem and proof

[*]Each of these theorems was left uncompleted once

Purported Theorem[†]	Intended Validity	Accuracy	Justification
Angle Bisector[*]	Valid	58.33%	50.00%
Angle Isosceles	Invalid	38.46%	76.92%
Bisector Isosceles	Valid	53.85%	53.85%
Inscribed Angle[*]	Valid	66.67%	66.67%
Miquel Point[*]	Invalid	25.00%	8.33%
Ptolemy[*]	Invalid	41.67%	0.00%

Concerning discrepancies were found on the three theorems intended to be invalid. For two of the theorems, Miquel Point and Ptolemy, participants failed to notice the logical flaws. Not a single participant noticed that the proof from Ptolemy proved only one way of the if-and-only-if statement. A single participant concluded that the proof from Miquel Point was invalid because it failed to prove the point lay within the triangle. In both cases, recognition of the logical error stemmed from a full understanding of the theorem statement. Modifications to both theorems were necessary before conducting the eye tracking experiment.

The significant increase in the justification score occurred on Angle Isosceles which was invalid due to lack of justification and the possible use of the statement itself. Participants, however, were asserting valid after independently proving the claim without relying on the theorem statement. This was unanticipated. Given this discrepancy, length, and general commonality of the argument, Angle Isosceles was not used in the eye tracking experiment.

We measured the effect of fatigue by calculating overall accuracy by order rather than proof. The accuracy on the sixth proof dropped significantly compared to the previous proofs.

This observation, coupled with the time measurements, provided sound reason to conduct the eye tracking experiment with five purported theorems.

The interviews indicated that the majority of participants (11 of 13) were consistent in their proof validation methodology across all validations. Self-proclaimed diagram vs text utilization varied across groups. Utilizing the transcripts to delve into the validation process, several interesting, but inconclusive, observations were noted. These observations played a role in the final construction of the eye tracking experiment and analyses. Unsolicited, a couple of participants explained that belief in the theorem altered their validation process. Ignoring the particular theorem, behavioral counts provided some intrigue. There were 14 instances where proof validation concluded upon the realization of a flaw. Almost an equal number of instances (11) occurred where participants identified troublesome lines, completed the proof, and then returned to those lines.

The results of the pilot study provided insight into the further development of the eye tracking study design and analysis. It informed decisions regarding number of purported theorems, proof construction and layout, further questions, and analyses. With these effects in mind we continue with the methodology of the eye tracking experiment.

Eye Tracking Experiment

Participants

The eye tracking experiment was conducted during the fall semester of 2016; 46 participants agreed to participate, only 41 were successfully run through the experiment. These participants were unable to participate for one of two reasons: visual acuity or ability to have their eye movements tracked. In the case that I was unable to achieve an accurate calibration, a graduate student in Psychology with extensive experience from the eye tracking lab would

conduct a calibration. If he was unsuccessful only then were participants rejected with eye tracking issues as the basis. A more detailed explanation occurs in the upcoming procedure section. Participants were again paid $15.

Participants were again sorted using the four tier classification of novice undergraduate (13 male, 5 female), early graduate (7 male, 2 female), late graduate (5 male, 2 female), and faculty (4 male, 3 female). Graduate students and faculty were again recruited via email or in person conversations. As school was in full session, the recruitment of novice undergraduates was conducted through a short exposition in advanced proof-based mathematics courses. See the following table for the courses fitting this description. With the approval of the instructing professor, the premise, requirements, and payment of the experiment were presented to students at the beginning of class. Interested students received a scheduling sheet upon which they put their name, email, and typical weekly availability.

Table 3-3 Courses used to recruit novice undergraduates

Advanced Proof-Based Undergraduate Courses with Instructor Approval (Fall 2016)		
Advanced Calculus I	Dynamics, Chaos, and Fractals	Introduction to Algebraic Systems
Discrete Mathematics	Foundations of geometry	Introduction to Complex Analysis

Materials

With the conclusions drawn from the pilot study, there were six purported theorems used in the eye tracking experiment. Each was again presented in the three screen format: statement, proof, and answer. The same practice example was used without modification. Having rejected the use of the Angle Isosceles theorem, the remaining five purported theorems from the pilot study were presented in modified form. Changes ranged from subtle changes in descriptive lines

to the addition of a lemma to the insertion of a flawed argument. See Appendix A for the full theorems and proofs used in the experiment.

Table 3-4 Eye tracking experiment: list of purported general theorem statements with intended logical issues

Purported Theorem	General Statement	Intended Proof Issues
Practice	If a diameter bisects a chord, then it is perpendicular to the chord.	Argument degenerates and is indeed false when the chord is also a diameter.
Angle Bisector	An internal angle bisector divides the opposite side length proportionally to the adjacent sides.	None.
Bisector Isosceles	If two internal angle bisectors are of equal length, then the triangle is isosceles.	None.
Inscribed Angle	An inscribed angle is equal to half of the central angle.	None.
Miquel Point	If an arbitrary point is taken on each side of a triangle, the three circles determined by each vertex and the two points on the adjacent sides have a common point of intersection within the triangle.	Proof provides an invalid argument asserting the point lies within the triangle, which is indeed false.
Ptolemy	If a quadrilateral is cyclic then the product of the diagonals is equal to the sum of the products of the opposite sides.	None.

Given the difficulty of constructing convincing invalid proofs and the results of the pilot study, only one proof, Miquel Point, was successfully modified as objectively invalid. Given that the point need not lie within the triangle, the argument cited the visual properties of the provided diagram and applied them the quadrilaterals in all cases; it stated two quadrilaterals were convex. This flaw was thought to be identified through two means. Visual manipulation of the arbitrary points would provide cases where the point did not fall within the triangle. Secondly, the argument presupposes that the Miquel point fell “between” the arbitrary points on the vertices. If

this were not the case, the cited quadrilateral would be a crossed quadrilateral and would thus require further formulation.

Procedure

Based on the pilot study data, participant sessions were expected to be completed within an hour. Despite this expectation, the starts of individual sessions were scheduled at least an hour and a half apart. Each session occurred in the same Visual Cognition Eye Tracking Lab in Bluemont Hall. Each session began by obtaining informed consent during which the purpose and method of the experiment were discussed. General information about the participant was then collected; these included age, gender, and native language. Participants needed to demonstrate a visual acuity of at least 20/30 with corrective lenses. This was measured using the program FrACT (http://www.michaelbach.de/fract/index.html). Once this general information was collected and noted in the experiment log, our focus turned to the eye tracking portion of the experiment.

To help ensure the accuracy of our tracks, a chin rest with forehead brace was used. The chin rest and chair were adjusted to the participant's comfortability. The forehead brace ensured a viewing distance of 54.5 cm. An explanation of the importance of a consistent head placement throughout the experiment, especially during the validation process, was presented. The participants were reminded of their ability to take breaks between theorems. To help facilitate progression, participants were encouraged to match head placement as well as they could. This was to minimize the need for recalibration; a drift check was conducted before each theorem to determine if recalibration was needed. Once the participant was comfortable and aware of these particulars, a 9-point calibration-validation process ensued.

The eye movements were recorded using an EyeLink 1000 Plus (http://www.sr-research.com/index.html) set to sample at 1000 Hz and is accurate up to 0.5° of visual angle. The setup requires two computers; one to run the eye tracking application and the other to display the experiment. While capable of tracking two eyes simultaneously, a single eye was tracked for this experiment. In general, the participant's dominant eye was tracked. This calibration-validation process continued until the average difference was less than 0.5° of visual angle with no particular instance differing by more than 1°. After ensuring an accurate track through the calibration-validation process, a Zoom H2n Handy Recorder was set to record the session. Recording the entire session minimized the risk of missing later justifications and altering participant validation processes.

The experiment, built again by Experiment Builder, consisted of three sections: the instructional phase, the proof validation phase, and the answer bias phase. This third phase was unannounced as knowledge of its existence could alter engagement throughout the experiment. The first phase presented further written explanation of experimental method and intent. As per the pilot study, a symbol and definition screen were presented which ensured participant understanding of notation indicating similarity and the definition of subtended. The phase ended with an example to solidify participants regarding the structure of the validation stage. Questions were encouraged throughout the entire instructional phase. Questions regarding performance, however, were not entertained. No eye movement or response data was collected throughout this phase.

The second phase consisted of all the proof validations. This was the only phase in which eye movement data was collected. There were five validations occurring in this phase; performance feedback was not given during the experiment. If interest persisted to the end of the

phase three, feedback was provided. Each validation began with a drift check. Further calibration was conducted if needed. The validation occurred throughout three screens: statement, proof, and answer. When the participant was ready to begin the validation process, the experimenter initiated the statement page. Within the validation, the participant controlled the progression. While there was no loss of available information through progression, certain safeguards were implemented through the experimental design to prevent accidental progression. This preserved the integrity of our eye movement and timing measurements. Participants were encouraged to take as long as necessary to reach an informed decision. The only expectation was immediate progression from proof screen to answer screen upon being ready to provide an answer. Answering occurred via a keyboard press of V or N. Participants then justified their assertion with the entire proof visible on the screen. Clarification was sought by the experimenter if necessary. Participants were told that if the justification for a valid assertion was that they found nothing wrong, providing that justification was not necessary.

The presentation of purported theorem and proof was balanced using a series of Latin Squares. Original plans supported up to 80 participants. Eight unique squares provided different viewing orders for the first 40 participants; this list of orders was then duplicated for the remaining participants. This balancing was implemented in order to combat order bias.

During the third phase, only response data was collected. This third phase was novel to the eye tracking experiment. Due to the unsolicited indications from the pilot study that belief in the purported theorem altered validation processes, we sought to explore this notion. We thus presented each theorem statement, in the same order as seen during the validation process, questioning their initial intuitions or beliefs about the purported theorem. Participants indicated their initial belief with a keyboard press of T or F.

Upon the completion of the third phase, participants were presented with a debriefing form (See Appendix B) and asked if they had any questions. The questions were answered to the best of the experimenter's ability. If interested in seeing the results of the experiment, long-term email information was collected. Participants were then paid and escorted back to the lobby.

Chapter 4 - Analysis

Our approach to data analytics sought to avoid the contention surrounding the eye tracking experiment conducted by Inglis and Alcock (2012). As it is unknown whether validation methods are consistent from proof to proof, the loss of information from averaging data across validations was not acceptable without justification. Therefore we conducted our analyses on each theorem separately. On several occasions aggregation was deemed reasonable and necessary. The overall accuracy and justification scores were computed across validations via means. In our attempt to compare overall validation processes (see research question two), aggregation of data was necessary. Firstly in order to maintain independence, average within and between dissimilarities were computed to be analyzed. And secondly, by treating the dissimilarity data from all five validations as five dimensional data, mathematical norms were necessary to compute the dissimilarity between overall validation processes for further analysis. After analyzing very general fixation time proportions and within-diagram attentional changes, the loss of specificity in the data was deemed reasonable. We therefore, in addition to individual purported theorem calculations, compared the average proportions of these measures. These are discussed in detail later in the chapter.

With a total of 41 participants, data collection goals were not met. The initial goal was to run nearly 80 participants (about 20 per classification) through the eye tracking experiment. This expected quantity would provide enough information to analyze the maturation process. The desired statistical strength was not present with 41 participants spread across four classifications. Given that three classifications failed to reach a count of ten participants, our analyses were conducted in the classical division of expert and novice. The division of the classifications was not a difficult one. The majority of participants labeled early graduate students (5 of 9) were in

their first semester as graduate students. An additional participant was still actually an undergraduate, and the remaining three were just starting their second year of graduate school. We therefore combined the novice undergraduates and early graduate students and relabeled them as novices. The previous literature qualified participants as expert given several different standards: research-active mathematicians (Inglis, Mejia-Ramos, Weber, and Alcock, 2013), PhD students (Mejia-Ramos, 2013), or academic mathematicians (Inglis and Alcock, 2012). We therefore found it reasonable to combine our previous division of late graduate students and faculty into a single classification, expert. With this reclassification, we proceeded with two test groups: novices (20 male, 7 female) and experts (9 male, 5 female).

Preparing the Data

Upon the completion of data collection, the process of compiling the data began. For each participant in addition to the recorded audio file, the experiment returned a series of results files. Each of these files contained a certain aspect of the data: eye movement data, validation response and timing data, and belief responses regarding the purported theorem based solely upon the statement. These separate files were then incorporated into a more suitable form.

The assertions and justifications of each participant were transcribed and timestamped by purported theorem for analysis. Participants rarely provided justification for valid assertions. The justifications for invalid assertions generally identified the particular statements that rendered the judgment. Issues with particular statements were then tallied by mathematical experience. On rare occasion (4 of 205 validations), the participant altered an assertion of invalid to valid during the justification process. These four occurrences were spread across four participants and two of the purported theorems. The assertion accepted for these four validations was dependent upon the analysis. Since the participant, without any feedback, altered the assertion during the

justification, the altered answer was accepted for the accuracy and justification analyses. For the purposes of eye movement analyses, the original assertion was retained. This was due to the fact that participants were instructed that progression through to the answer screen must take place only when an answer is ready to be given. The participants in these cases had reached a conclusion and readily asserted that the proof was invalid. The eye movement data collected while on the proof screen represented that of an assertion of invalid. While eye movement data was collected on the answer page, the aggregation of that data into a single validation attempt would be questionable and was therefore not done.

The eye movement data files were compiled using Data Viewer (http://www.sr-research.com/index.html). Data Viewer provides various useful analytics tools. Through this program, experimenters design the AOIs for each screen. Note that AOIs can be constructed before or after the collection of data. AOI construction is discussed in further detail later in this section. The various measurements conducted during eye tracking are all reported in the data file. Data Viewer catalogs all of the measurements from each data file and allows you creation of reports containing only the measures of interest. These reports are exported as Excel sheets for further analysis. Eye movements of entire validations can be reviewed at adjustable playback speed. There is also the capability of creating fixation and duration heat maps to identify features of increased usage.

The remaining results files consisted primarily of participant responses to validation prompts and the initial belief in the purported theorem statements. We first address the initial belief data. The integrity of the initial belief measure was drawn into question during the data collection phase of the experiment. Upon entering the third phase of the experiment, one participant responded to the directions with “I was assuming they were true, that you were giving

me true statements" (GeoP33, 55:52). It is probable that this honest response was provided due to a greater familiarity between us than with the majority of other participants. This casts doubts upon the integrity of initial belief data, and it was thus removed from further analysis. The other response files provided additional information regarding the time spent in each stage of the validation process. While the time spent on the answer screens was measured and reported, this data was included only to facilitate in the experimental design; it was never meant to be analyzed. The time spent on the answer screen depended upon both experimenter and participant and was thus not a useful measure. These files were then compiled as Excel files for further analysis.

Verifying and cleaning the eye movement data

During each validation, the experimenter monitored, from the second computer, how well the eye tracker was tracking. There were two primary cues for the quality of the track. First, the second screen displayed participant eye movements through the motion of a dot. The dot would disappear when the track was lost and reappear upon regaining the track. A written status of the track was also displayed prominently on this second screen. Given that blinking occurs naturally and frequently, the loss of a track is not in and of itself a problem. The concern is prolonged tracking absences. There were four instances, involving three participants in total, with significant tracking issues occurring during a validation. The experimenter made note of the issue in the experiment's logbook without notifying the participant so as not to alter the validation process. After the conclusion of that particular validation, prior to the drift check, the experimenter recalibrated the participant. The eye movement data for these validations was removed from any further analysis. All non-eye movement data, however, remained in our analyses as the integrity of the validation itself and hence responses were not undermined.

Monitoring eye movements concurrent with the validation process was the first quality assurance check. To check for possible anomalies or systemic errors, distributions of two different measurements from the proof screen were assessed by expectancy. The first measure was saccade length, a measure of the distance between fixations. In general, these distributions were bimodal in nature. This was to be expected given the arrangement of the displays. The shorter saccade lengths occur naturally during the reading process, and the saccades between text and diagram result in the longer lengths. Abnormal recurrences of saccade lengths were not manifest in our data. Fixation duration distributions did not supply any concerns regarding our data. The general distributions were unimodal skewed right as expected for fixation duration distributions. Given these distributions, the integrity of our data was not drawn into question.

Before eye movement analyses were conducted, the eye movement data needed to be cleaned. The cleaning was conducted by participant classification. In keeping with the original experimental design, the data was cleaned separately for each of the four original classifications. Using fixation duration as the basis, the extreme ends of the data were removed. The fixations occupying the bottom 1% and the top 1% in fixation duration were removed from the data. This was the lab standard meant to remove random and possibly erroneous fixations. This was the final stage of data preparation.

Construction of AOIs

Before we discuss the analyses conducted, we return to the topic of AOIs. These AOIs facilitate understanding of the eye movement data by classifying each fixation based upon its location. Design of the AOIs incorporate and further enable the research goals. Recall AOIs can be implemented before or after the data collection process. This continual ability to alter fixation classifications enables researchers to conduct further tests when the initial data presents the need.

We created the AOIs used in this experiment prior to the execution of the experiment. The accuracy of the eye tracker was incorporated into the design in order to preserve measurement integrity. The intent of our study was to address contentious results regarding validation methods, to provide insight into the differences in the validation process with respect to mathematical maturity and to better understand diagram usage. The layout of the proof screens already incorporated subtle visual cues for the structure of the proofs. The segmentation of the text into separate AOIs further continued down this path. Individual AOIs were composed of the individual ideas or statements of the proof. These consisted of, at a bare minimum, a line of text, and maximally a statement formed across several lines. The majority consisted of a single line of text. These structural chunks provided fixation content without resorting to a word-by-word analysis. Theorems and lemmas were each treated as separate single AOIs.

Previous work in diagram usage partitioned the diagrams into thematically relevant and irrelevant areas (Madsen, 2013; Johnson, 2015). This tactic was only possible because of the availability of space and the existence of static areas of relevancy. In the validation process of proofs with diagrams, features have a dynamic relevancy. This relevancy is doubly layered. The relevant content at a given time is dependent upon the line of the proof being referenced. Furthermore, it is primarily about drawing connections between varying aspects of the diagram. So relevancy would be more closely related to particular paths rather than stationary fixations in particular areas. Having to balance the amount of content presented on the proof screen with diagram size, the presented diagrams were not large enough for a fine segmentation. AOIs differentiating the finer details would not fall within the accuracy of measurement. Coupling the dynamic nature of relevancy with concerns of accuracy, entire diagrams were treated as single AOIs. Each diagram used in the Inscribed Angle proof formed a separate AOI. We thus

conducted our analyses with textual AOIs that identified different statements and AOIs indicating the use of a diagram. When conducting AOI analyses, the fixations not residing in a predetermined AOI was classified as nonAOI.

General Analyses

The initial analyses conducted were general in nature. They were conducted to provide a framework for the future analyses that would directly answer research questions. Fixation duration and saccade counts form the basis for many of the analyses conducted. General trends in this data could influence these analyses. In comparing novices and experts we are interested in the differences of attention generally and specifically. Without accounting for the general, the specific loses meaning. Take for example, a blind comparison of fixation time in a specific area that results in a find of significant differences. Does this reflect the possibility of a general difference in validation time or that there is significant difference in the perceived importance of this specific area? We thus conducted analyses comparing total validation time and total saccade counts. Validation time was computed as the total time spent viewing the proof screen. The results of a mixed factorial ANOVA analysis are presented in Table 4-1. Recall the removal of four validations across three participants, the saccade analysis was run without the data from these participants.

Table 4-1 Results of mixed factorial ANOVA for proof validation times and saccade totals

*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

Effect	Proof Validation Time		Saccades Made During Validation	
	F	p	F	p
Purported Theorem	F(2.51,97.95)=14.25†	<0.0001*	F(2.18,78.54)=15.26†	0.0001*
Participant Type	F(1,39)=3.02	0.09	F(1,36)=2.44	0.13
Purported Theorem*Participant Type	F(2.51,97.95)=1.39†	0.25	F(2.18,78.54)=1.55†	0.22

Significant main effects were found by purported theorem. These differences were not unexpected given the various degrees of length and complexity presented across proofs. Differences between participant types failed to reach significance. The analyses presented in Table 4-1 do not show the general behavior of the data. General trends, while not necessarily statistically significant, provide additional insight informing the decisions of future analyses. Viewing times and saccade counts were thus analyzed by individual theorem seeking general trends. Given the individual theorem approach, these averages were computed including the validations from the previously excluded participants. This pattern of behavior is consistent throughout our analyses. The results of this further investigation are presented in Figure 4-1 and Table 4-2.

Figure 4-1 Bar graphs displaying average measures of time spent and attentional change during the proof validation process.

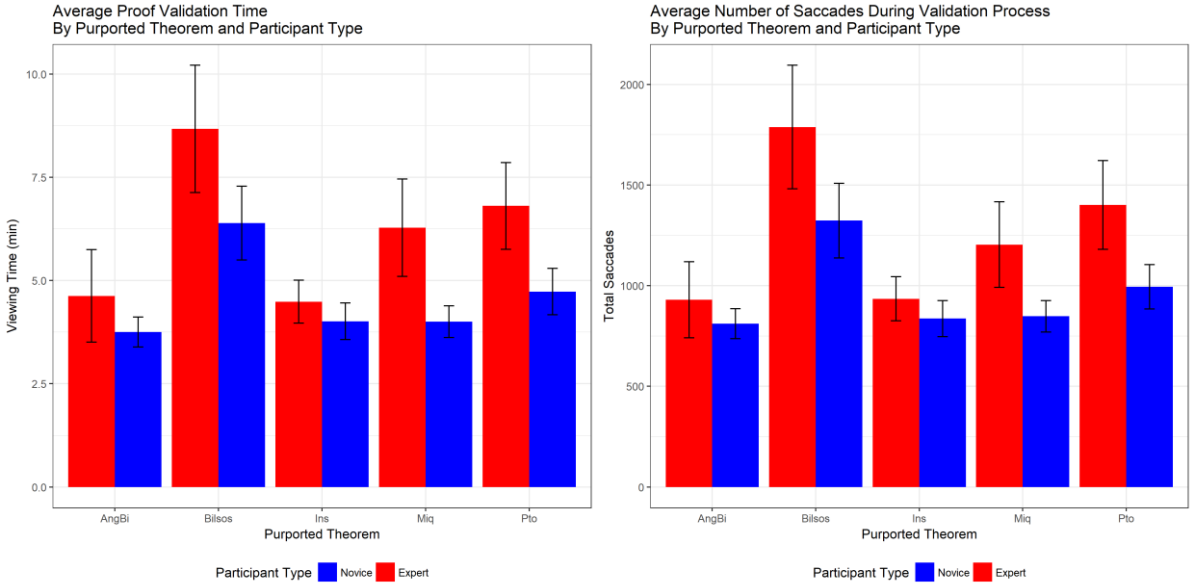


Table 4-2 Average validation time and total saccade counts

Purported Theorem	Measure	Novice	Expert
Angle Bisector (AngBi)	Viewing Time	3 min 45 sec	4 min 38 sec
	Total Saccades	812	931
Bisector Isosceles (BiIsos)	Viewing Time	6 min 23 sec	8 min 40 sec
	Total Saccades	1323	1789
Inscribed Angle (Ins)	Viewing Time	4 min 1 sec	4 min 29 sec
	Total Saccades	837	935
Miquel Point (Miq)	Viewing Time	4 min 0 sec	6 min 17 sec
	Total Saccades	848	1204
Ptolemy (Pto)	Viewing Time	4 min 44 sec	6 min 48 sec
	Total Saccades	995	1401

The general trend of both measurements warranted attention. In both particularly similar cases, experts consistently measured longer validation times and more saccades. Given the visual similarity, further inquiries were made to ensure the accuracy of the data and calculations. Since saccade and fixation counts are directly related, the mean fixation duration distributions by participant type were used to verify this relationship. As they shared a similar distribution and their means were close, the similarity in viewing time and saccade counts is not unexpected.

Given that experts generally spent longer times validating and made more saccades than novices, further analyses utilized normalized measurements. Instead of comparing fixation duration in a specific area, the comparison is made between proportions; the proportion of time spent in that area to the total time spent fixating during that validation. Normalized measurements enable the comparison of attentional importance in the validation process.

Validation Accuracy

The accuracy of individual validation attempts has received a significant amount of attention. Early researchers supported a standard of uniformity in validation conclusions within the realm of expert mathematicians; their assessment of novices deemed their performance at chance. Subsequent research has been conducted both questioning and confirming earlier thoughts. Novice performance continues to yield similar results, while the uniformity of mathematicians falters in recent research. Given the continual interest and the benefit of repeated results, we conducted analyses on the accuracy of our validators.

Alcock, Hodds, Roy, and Inglis (2015) caution against claiming proofs exude "obvious" validities. This caution is well-advised given various previous studies. Our goal in designing these proofs was to provide nontrivial proofs with particular intentions of the validity. The mental litmus test of intended validity assertions was belief that a simple, but genuine, interaction with a participant, as described by Selden and Selden (2003), would conclude in agreement. Given the small population, this mental litmus test remained as such, a mental barrier. We compute participant accuracy by comparing provided assertions with intended assertions as percentages.

Having the justifications provided vocally by participants enabled further analyses. Not only could we analyze the accuracy of participant assertions, we could also incorporate the

justifications for these declarations. In conjunction with the assertions, each justification was considered for reasonableness and rendered right or wrong. These were then tabulated to identify both overall and specific performance. Novices and experts were compared by counts of matching assertions for accuracy and right justifications for justification. Given the likelihood of small frequency counts, the Fisher exact test was used to make the comparisons.

Table 4-3 Results from Fisher exact tests on accuracy and justification counts

*Indicates a significant difference at the $\alpha = 0.05$ level.

		Percent Accuracy			Percent Justification		
		Novices	Experts	p	Novices	Experts	p
	Overall	59.26	74.59	0.087	52.59	72.86	0.029*
Purported Theorem	Angle Bisector	59.26	100	0.007*	59.26	100	0.007*
	Bisector Isosceles	62.96	35.71	0.115	59.26	35.71	0.197
	Inscribed Angle	70.37	100	0.035*	70.37	100	0.035*
	Miquel Point	29.63	71.43	0.019*	0	50	0.0001*
	Ptolemy	74.07	64.29	0.719	74.07	78.57	1

Looking at the overall percentages, note that novices perform marginally better than chance but not substantially better. This is in line with the reports of previous studies. Looking at the individual novice accuracies, the consistency displayed indicates novices perform perhaps better than chance in certain circumstances. Here we recall that Miquel Point is the only objectively invalid proof presented. It is possible the increased performance on the proof intended for valid assertions was due to a proclivity for valid assertions.

To test for this possibility, we computed the bias measure, *c*, from Signal Detection Theory (MacMillan and Creelman, 2005). This measure is determined from the hit and false alarm rates, which correspond to the rate of correct valid assertions and the rate of incorrect valid assertions, respectively. The magnitude of the measure indicates the strength of the bias while the sign indicates the type of bias. With *c* values of -0.4829 and -0.0543, respectively, both the novice and expert groups displayed a liberal bias, an inclination towards valid assertions. Given

the magnitudes, however, the expert bias is slight. The novices display a much more prevalent bias. Given that more proofs were intended valid, this bias may account for the better performance. Another possibility is the mathematical maturity present in the novices we studied. The novice group consisted of early graduate students and undergraduate students from upper division courses, possibly making it the most mature novice group studied.

Using overall uniformity levels rather than accuracy percentages, previous studies report levels of uniformity of 75% and 81% (Inglis, Mejia-Ramos, Weber, and Alcock, 2013; Inglis and Alcock, 2012, respectively). The uniformity level displayed by our expert participants was 80%. The individual theorems provide a similar tale to findings of Inglis and Alcock (2012). In both studies, experts showed uniformity on several theorems and dissention on the rest. Typically 4-5 participants dissented with the majority. Our expert group consisted of 14 participants compared to the 12 experts from Inglis and Alcock (2012). In either case, 4-5 is not negligible. Our findings thus bolster the conclusions from the literature.

Justification results indicate participant difficulty in the invalid argument presented in Miquel Point. None of the novices provided reasonable justifications for the purported theorem and only 50% of experts were successful in doing so. The invalid line presents an invalid perceptual inference. It claims certain quadrilaterals will always be convex, implying they will not be crossed quadrilaterals. While this is the presented case, this is not generally the case. So properties of the visualization were applied to the general case where they ceased to hold. Given that perceptual inferences are generally not accepted by novice validators (Zhen, Weber, and Mejia-Ramos, 2015), participants must have failed to recognize the occurrence of such a perceptual inference. It has been shown novices primarily focus on surface features. It is possible this perceptual inference, requiring a deeper understanding, was not recognized due to the

superficial process of novices. This, however, doesn't account for the poor recognition on behalf of experts. Nonetheless, performance differed significantly between novices and experts in 3 of the 5 proofs.

Analyses to Answer Research Questions

Having presented several general analyses that frame our future investigations and confirm the results of previous studies, we turn our attention to the questions motivating this research. Recall our interests lie in resolving contentions regarding *zooming out*, comparing overall validation processes, and understanding the use of diagrams. Each research question is addressed with its appropriate analyses in the following sections.

Research question 1: *Zooming out*

The notion of *zooming out* refers to the viewing of a proof as an application of methods rather than series of inferences building to a conclusion. The general emphasis resides in the logical components of the proof. During the course of a proof validation, the manner in which the validator views the proof may vary. Different views of proof result in different validation processes. One such process entails an initial reading of the proof to gauge its overall structure (Weber, 2008). This is called an initial skim-reading. Previous evidence for the occurrences of *zooming out* relied upon retrospective interviews and verbal protocols. With the dawn of eye tracking, such methods need not be solely relied upon. Inglis and Alcock (2012) originally sought to test for initial skim-reading through their IR ratio. Contentions arose through separate analyses of the same eye movement data (Weber, Mejia-Ramos, Inglis, and Alcock, 2013). We sought to address the issues raised through these analyses.

In order to ensure the possibility of an initial skim-reading to occur, we constructed longer, more complicated proofs from a subject field with mutually less familiarity. This

prevented participants from immediately recognizing commonly structured proofs. Furthermore, these proofs did not present the blatant errors that would render simplified validation processes. Having addressed these primary concerns in our experimental design, the secondary concerns of the early presence of fixations may be addressed. If early fixations are commonly present as shown by IR ratio distributions, whether these are incidental or part of an initial skim is easily determined through the playback mode of Data Viewer. So the IR ratios in conjunction with the underlying distributions fully enable us to test for the general occurrence of initial skimming as a strategy. The IR ratio is computed by computing the ratio of fixation time through the first fixation on the last line of the proof and the total fixation time of the validation. The IR ratio measure cannot be calculated on trials without a single fixation at the end of the proof. We therefore had to exclude a total of 13 sets of eye movement data from 12 participants over 4 theorems in this analysis. The breakdown of participants by type is show in Table 4-4.

Table 4-4 Results from IR ratio testing for initial skim-reading

Purported Theorem	Novices		Experts	
	IR Ratio	# Excluded	IR Ratio	# Excluded
Angle Bisector	66%	0	69%	0
Bisector Isosceles	71%	1	71%	3
Inscribed Angle	80%	5	84%	1
Miquel Point	55%	0	47%	1
Ptolemy	67%	2	71%	0

The IR ratio distributions were created using a bin width of 10% and are reported in Appendix B. The only distribution with a decent number of IR ratios less than 50% was the Miquel Point. The vast majority of these resided in the 30-50% range which doesn't support an initial skim-reading. Given the calculated IR ratios of both groups never drop significantly below 50% we conclude that initial skim-reading is not a general strategy implemented in this setting.

Continuing to analyze the data for a general use of the *zooming out* strategy, we turn to validations producing the invalid assertion. As witnessed in the pilot study, there were a nontrivial number of instances where participants identified an error or troublesome argument, continued through the rest of the proof, and returned to the error later. Continuing through the entire proof was not prompted; participants were only told to spend as much time as needed to make an informed decision. To the best of our understanding, there are two reasons for continuing the proof after identifying a possible error. The first arises from uncertainty regarding the legitimacy of the possible error. In this case continued progression through the proof could mean a couple of different things. Firstly, continued progression through the proof provides further information regarding that particular argument, whether it is simply a better understanding of the situation or something more subtle. Similar arguments made later may prove clearer, providing the necessary insight for this troublesome argument. The likelihood of clearer exposition later in the proof is terribly low, as repetitious arguments become terser with use. Furthermore, the conclusion of the argument will be used later in the proof, and this usage may provide the insight. A second reason for continued progression through the proof lies in trying to understand the role that particular claim plays in the overall argument and deciding if the error is able to be remedied.

Given these reasons for continued progression through a proof after identifying a possible error, we further argue that these reasons exhibit a strategy of *zooming out*. In the first case, participants questioned whether or not they encountered an error and then use the structure of the proof to better understand the identified argument. The participant in the second situation is attempting to construct the logical chunk and determine how essential that line is to the logical chunk. Now one may say, a validator might be looking for a more obnoxious error to rely upon

rather than hedge bets on a single error. If a participant is not continuing due to the second reason, an identification of an unquestionable error terminates the validation. There thus is an uncertainty about the identified argument and either subconsciously or consciously further progress through the proof is relative to that identified argument.

Attributing this sort of behavior to the use of a *zooming out* strategy, we identified participants that displayed this behavior. This process began with tabulating each validation terminating with an assertion of invalid. Pairing these assertions with their associated justification allowed us to pinpoint where in the proof an issue was identified. The proportions of the total fixation duration by AOI provided the check for further progression. Given the existence of incidental fixations, fixation proportions less than 1% were considered as incidental for the purposes of gauging continued progression. To be classified as continued progression, one of two qualifications had to be present. The first was the completion of the entire proof. In the second qualification a specific error needed to be stated in the justification, and given this error, further progression through the proof was evident. It however terminated before completing the proof. There were two cases in which a classification could not be made. If the cited error occurred at the end of the proof, continued progression was impossible. If no specific error was cited and the validation terminated without the completion of the proof, it was uncertain as to whether the termination occurred at or after the error. These account for the discrepancies in total assertions of invalids presented in the following Table 4-5. See Appendix B for the breakdown of continued progression counts and unclassifiable cases.

Table 4-5 Results from classifying validations asserting invalid by continued progression

Purported Theorem	Assertions of Invalid	
	Novice	Expert
AngBi	12	0
BiIsos	10	9
Ins	9	0
Miq	8	10
Pto	9	6

Purported Theorem	Continued Progression		Termination at Specified Error	
	Novice	Expert	Novice	Expert
AngBi	11	n/a	0	n/a
BiIsos	10	7	0	2
Ins	7	n/a	0	n/a
Miq	6	7	0	1
Pto	9	5	0	1

Analysis of the results presented in Table 4-5 identifies a couple of interesting details. The majority of validators, both novice and expert, when granting an invalid assertion utilize further progression through the proof to better frame their understanding of the error in general. This confirms that both novices and experts partake in the strategy of *zooming out*. As put by one of the pilot study participants:

> If I can't justify a line, it doesn't make sense to go line by line on the rest of the proof. […] Looking forward gives you a picture of the whole proof and then it's easier to go back to the line that you have a hard time with. […] Maybe there was something you didn't quite catch but because it is in the larger context, it's easy to see. (GeoPP12, 38:17)

Furthermore, the absence of validations terminating at an error indicates novices exude one of two characteristics. They either lack confidence in their ability to identify legitimate errors or they have a firm grasp that understanding is continually reframed through the reading progress. This concluded our analyses of the first research question.

Research question 2: The validation process

Various differences between novices and experts within the validation process have been identified. These differences measure actions throughout the entirety of the validation process with neither temporal nor sequential incorporation. The noted differences remain significant and

have pedagogical consequences. Without the defining aspects of sequence or time, conclusions about overall validation processes cannot be reached. Novices and experts either engage in similar processes with differing abilities or they partake in different processes entirely. In order to draw these sorts of conclusions, validations, as a whole, must be analyzed. Using the raw fixation location and duration data, scan paths represent the entire validation processes. These scan paths were compared using the ScanMatch toolbox in MatLab (Cristino, Mathôt, Theeuwes, and GilChrist, 2010). Taking two scan paths as inputs, ScanMatch returns a normalized similarity score. We then converted this similarity score into a dissimilarity score by subtracting the score from 1. By comparing each possible participant pair, we constructed dissimilarity matrices for each purported theorem. These matrices formed the primary basis for our analyses seeking to answer our second research question.

These analyses were conducted at two levels: the individual purported theorems and the entire experiment consisting of all five validations. The first level required no manipulation of the dissimilarity matrices. In the second level, the dissimilarity data was treated as a vector in some five-dimensional space. Each component contained the one-dimensional dissimilarity between two participants on a single theorem. To compute the overall dissimilarity between a pair of participants, a norm was applied to that particular dissimilarity vector resulting in a scalar dissimilarity. Without a justifiable reason for one norm over another, two commonly used norms were applied in these analyses: the L_1 norm and L_2 norm. The L_1 norm is the sum of the component dissimilarities, and the L_2 norm is the Euclidean distance (dissimilarity). Given that three participants had eye movement data removed from certain validations due to bad tracking, these participants were either removed from analysis or treated separately as discussed later in this chapter.

With our interest residing in the comparison of overall validation process, the general proximity of the groups was a main interest. Comparisons of within-group dissimilarities to between-group dissimilarities were thus made to gauge this general proximity. Given that dissimilarities associate participants pair-wise, for each participant we computed the within-group and between-group mean dissimilarity scores. Within-group and between-group average scores were the then separated by participant type. These groups were given a short-hand notation: E-E--within the expert group, E-N--between groups using expert averages, N-E--between groups using novice averages, and N-N--within the novice group. Note: E-N and N-E have the same mean but differ in element count. A one-way ANOVA was then used to compare these distributions by theorem and participant type. Given our interest in comparing the similarity of validation processes, within-group dissimilarities (i.e. E-E to N-N) were not tested.

Figure 4-2 Bar graphs displaying averages of mean dissimilarity by within-group and between-group, and purported theorem

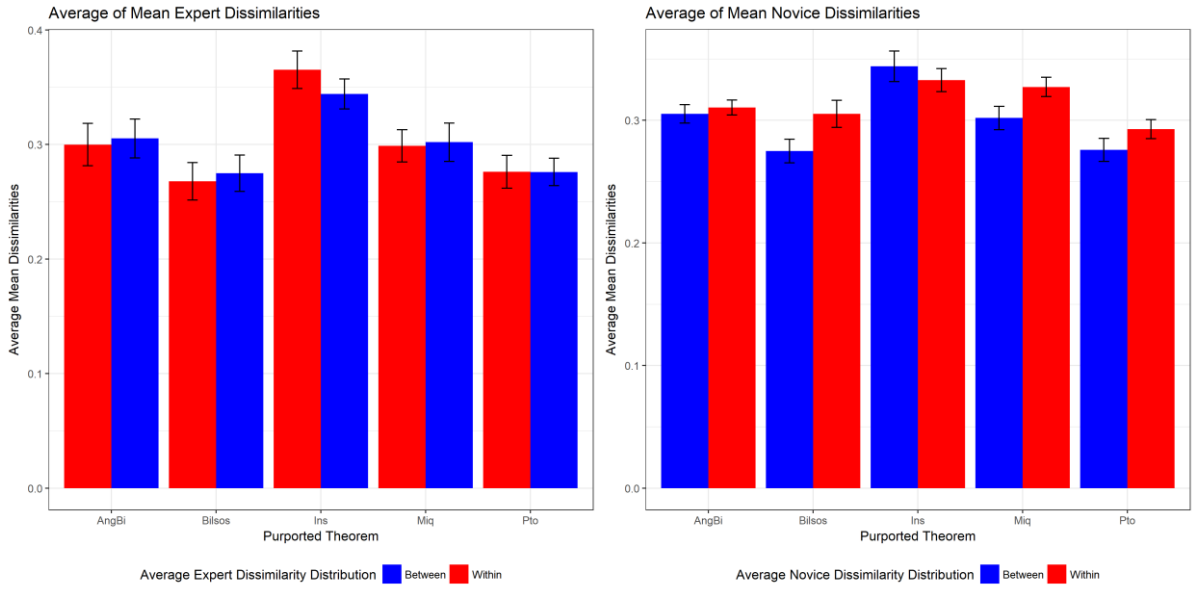


Table 4-6 Results from one-way ANOVA comparing within-group and between-group dissimilarities

*Indicates a significant difference at the $\alpha = 0.05$ level.

Purported Theorem	Comparison of E-E and E-N mean dissimilarities		Averages of Mean Dissimilarity			Comparison of N-N and N-E mean dissimilarities	
	F	p	E-E	Between	N-N	F	p
AngBi	F(1,26)=0.046	0.832	0.299	0.305	0.310	F(1,52)=0.285	0.596
BiIsos	F(1,26)=0.098	0.757	0.267	0.274	0.305	F(1,50)=4.284	0.043*
Ins	F(1,24)=1.014	0.324	0.365	0.344	0.332	F(1,50)=0.514	0.477
Miq	F(1,26)=0.02	0.887	0.298	0.301	0.327	F(1,50)=4.204	0.045*
Pto	F(1,26)=0	0.993	0.276	0.275	0.292	F(1,52)=1.925	0.171

The analysis shows that in general the dissimilarities within-group and between-group fail to be significantly different. There were two cases reaching a significant difference. In both of these cases, there was more dissimilarity within the novice group than there was between the novice and expert groups. This trend occurs on 4 of the 5 purported theorems. This indicates the possibility that collectively novices are closer to experts than they are to themselves. The general pattern indicates that the overall validation processes conducted by novices and experts fail to be significantly different. In a similar fashion, analysis of the normed dissimilarities was conducted. Recall these normed dissimilarities stem from viewing the dissimilarity data as a five-dimensional data set. Given our interest in comparing overall validation processes, we removed participants with missing data from the analysis.

Figure 4-3 Bar graphs displaying averages of mean normed dissimilarity by within-group and between-group, and norm

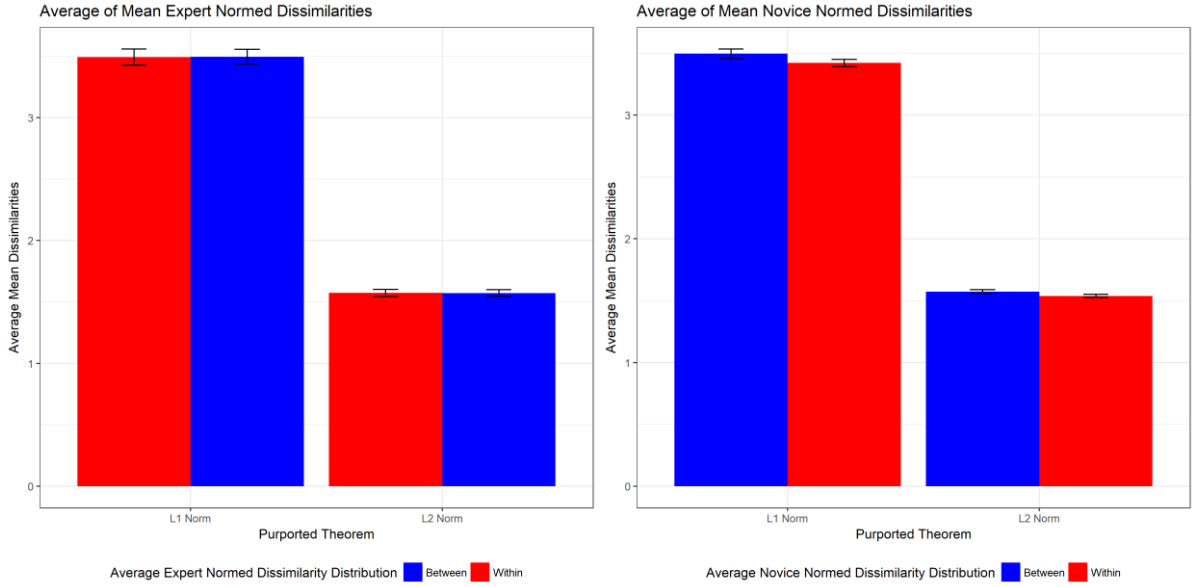


Table 4-7 Results from one-way ANOVA comparing within-group and between-group normed dissimilarities

Norm	Comparison of E-E and E-N		Averages of Mean Normed Dissimilarity			Comparison of N-N and N-E	
	F	p	E-E	Between	N-N	F	p
L_1	F(1,24)=0.001	0.979	3.493	3.495	3.421	F(1,48)=2.346	0.132
L_2	F(1,24)=0	0.986	1.573	1.572	1.537	F(1,48)=2.551	0.117

Again the analysis shows that the dissimilarities within-group and between-group fail to be significantly different. Given how similarly distributed these within-group and between-group dissimilarities are, the overall processes involved in the validation process must display a certain level of similarity.

Despite this overall similarity, the possibility of uniquely novice or expert validation processes existing was not eliminated. To provide a finer grained analysis in order to explore this possibility, clustering algorithms were applied.

Having the dissimilarity matrices from ScanMatch, the clustering algorithm of choice was k-medoids. This algorithm partitions data sets into k clusters, each of which is identified by

an exemplar, the medoid. The clustering problem is solved by trying to minimize the sum of dissimilarities between points and their associated medoid. The algorithm proceeds in the following form:

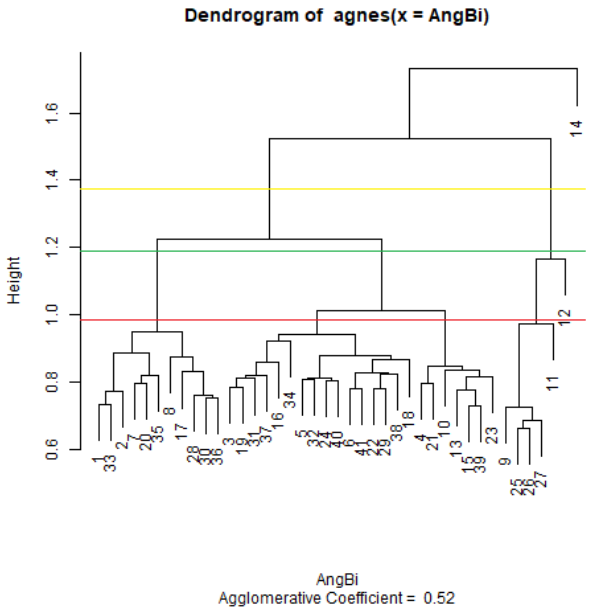
1. Identify k entities to serve as medoids (intentionally or randomly assigned)
2. Assign all entities to clusters using dissimilarity matrix to find which medoid each entity is the least dissimilar.
3. For each cluster, search for any entities within the cluster than reduce the average dissimilarity within the cluster if chosen as medoid. If an entity lowers the average dissimilarity, select it as the new medoid.
4. If at least one new medoid was selected, repeat steps 2 and 3. Else terminate algorithm.

There are two primary concerns with this algorithm. First, the number of clusters is an input into the algorithm. The second is that the algorithm can depend upon the initial choice of medoids. The latter issue may be addressed by running the algorithm several times and comparing the resulting clusters. The former issue may be addressed using another clustering algorithm: agglomerative nesting. The algorithm is of the following form:

1. Identify each entity as a separate cluster
2. Compute/update proximity matrix
3. Merge the two closest clusters
4. Repeat steps 2 and 3 until only a single cluster remains

There are several techniques for updating the proximity matrix; we used the group average technique. The dissimilarity between clusters was defined as the average of the dissimilarities between the entities in one cluster and the entities in another. The hierarchical clustering is then represented using a dendrogram which helps determine the number clusters.

Figure 4-4 Example dendrogram providing various possible cuts



At each level, clusters are being combined. The dissimilarity of clusters is represented through the height of the connections. Horizontal cuts through the dendrogram help identify possible clustering counts. The hierarchical nature also allows us to identify outlier data. Validations were removed as outliers if their individual cluster was joined with the cluster containing the rest of the group at the time. This is displayed in Figure 4-4, GeoPP14 (top right corner) was a singleton cluster until the very last combination during which it joined the entirety of the rest of the participants. There were a total of four outliers across all purported theorems. For a full reporting of dendrograms and outlier data see Appendix B.

The subjective nature of these cuts is of primary concern. Given that a strong majority of pilot study participants claimed to keep their proof validation process consistent, we decided to apply the constraint that the cluster counts must be reasonable for each purported theorem. This led to the decision to apply the clustering algorithm with two cluster counts: two and four. Three clusters also seemed plausible, but it failed to apply to one of the purported theorems. See appendix B for the full array of dendrograms with cuts. Our research interests reside in the

composition of clusters by participant type. The homogeneity of each cluster is presented in the Tables 4-8 and 4-9. Participant type homogeneity is calculated by taking the maximum percentage of participants from a single participant type within a given cluster.

Table 4-8 Homogeneity of mathematical experience and cluster size by purported theorem with 2 clusters

Purported Theorem	2 Clusters			
	First		Second	
	Homogeneity	Cluster Size	Homogeneity	Cluster Size
AngBi	82%	11	62%	29
BiIsos	78%	18	55%	22
Ins	71%	17	65%	20
Miq	52%	23	81%	16
Pto	59%	27	79%	14

Table 4-9 Homogeneity of mathematical experience and cluster size by purported theorem with 4 clusters

*Indicates a single participant differed in mathematical experience

Purported Theorem	4 Clusters							
	First		Second		Third		Fourth	
	Homogeneity	Size	Homogeneity	Size	Homogeneity	Size	Homogeneity	Size
AngBi	83%*	6	56%	9	69%	13	67%	14
BiIsos	60%	5	85%	13	60%	15	57%	7
Ins	56%	9	50%	6	85%	13	63%	11
Miq	69%	13	64%	11	80%	10	80%*	5
Pto	68%	19	78%	9	80%*	5	63%	8

Given that several clusters from the four cluster case only contained one member of differing mathematical experience, further analysis was conducted. We sought to see if the odd participant was an outlier. We thus computed the dissimilarities to the medoid in each cluster. The odd participant's dissimilarity was then compared to the average. If this dissimilarity exceeded the average it was deemed an outlier. The only case where this occurred was with GeoP12 in the first group of Angle Bisector; this participant produced the highest dissimilarity by a decent margin. The cluster in question actually consists of, with the exception of GeoP14,

the farthest right participants in Figure 4-4, the example dendrogram. The structure of this cluster provides a nice visualization for GeoP12's disparity.

We thus conclude that on the individual purported theorem level the clusters are primarily heterogeneous with respect to mathematical experience. The only possible exception would be the small cluster of five novice participants discussed above.

Perhaps as in the problem solving literature studying conceptual physics problems (Madsen, Larson, Loschky, and Rebello, 2012; Madsen, 2013), these clusters are better determined by answer rather than experience. We explored this possibility by computing cluster homogeneity by validation assertion.

Table 4-10 Homogeneity of validity assertion by purported theorem with 2 and 4 clusters
*Indicates a single participant differed in answer

Purported Theorem	2 Cluster Homogeneity		4 Cluster Homogeneity			
	First	Second	First	Second	Third	Fourth
AngBi	80%	67%	75%	83%	78%	57%
BiIsos	50%	55%	80%*	62%	53%	71%
Ins	82%	77%	78%	75%	80%	82%
Miq	57%	71%	69%	58%	80%	100%
Pto	70%	50%	84%	56%	60%	63%

For both calculations of homogeneity, by experience and by answer, we found that the large majority of clusters were heterogeneous. As in the previous analysis, one cluster's composition contained a single participant of differing experience when clustering into four groups. This cluster contained only six participants. In addition to this cluster, another cluster was homogeneous. This cluster contained five participants. Given that at most two clusters display homogeneity, the clustering is generally not determined by validation assertion.

Again trying to measure the validation process from the viewpoint of the entire experiment, we conducted a similar clustering analysis on the five-dimensional dissimilarity

scores. Given this analysis sought to identify clusters determined by overall validation processes, we included participants with missing data through a separation of processes. The norms were calculated using the available data. The missing data resulted in these participants displaying a uniformly lower dissimilarity with each other participant. If included with the initial clustering process, these participants would likely be classified as medoids. We thus conducted the clustering without these participants and then assigned them to clusters afterwards.

Our use of agglomerative nesting indicated that two or three clusters would be reasonable. It further identified a single outlier, GeoP14. This participant was removed for the clustering. Dendrograms with cuts are presented in Appendix B. The L_1 and L_2 norms resulted in the exact same clustering. We conclude with similar findings to the analysis by individual theorem. Clusters tend to be heterogeneous with respect to mathematical experience. We report these findings in Table 4-11.

Table 4-11 Homogeneity of mathematical experience and cluster size by norm and cluster count

Norm	2 Clusters				3 Clusters					
	First		Second		First		Second		Third	
	Homogeneity	Size	Homogeneity	Size	Homogeneity	Size	Homogeneity	Size	Homogeneity	Size
L_1, L_2	61%	28	83%	12	69%	16	82%	11	54%	13

The final analysis conducted sought to compare the initial stage of the validation process: the presentation of the purported theorem statement. The initial screen, for each validation, presented the purported theorem statement along with a visualization of the result. No inclination of proof structure was provided. The comparison of time spent viewing the statement screen is a testament to the relative depth of understanding desired before approaching the provided proof. The results of a mixed factorial ANOVA are presented in Table 4-12. Participants with missing data were removed from this analysis.

Table 4-12 Results of mixed factorial ANOVA for statement viewing time

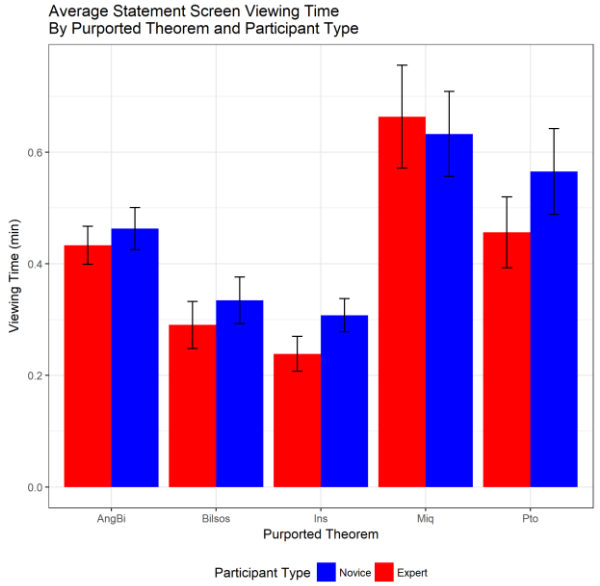
*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

Effect	Statement Viewing Time	
	F	p
Purported Theorem	F(2.61,101.88)=19.02†	<0.0001*
Participant Type	F(1,39)=0.50	0.49
Purported Theorem*Participant Type	F(2.61,101.88)=0.55†	0.63

A significant main effect was found by purported theorem. This difference is to be expected given the varying lengths and complexities of statements and visualizations across the different theorems. There failed to be a significant effect by participant type indicating that both novices and experts spend comparable time on the statement screen. To further gauge statement screen viewing time, analysis on the individual theorem statements was conducted. This latter analysis included the sound data of participants with missing data. The results are presented in Figure 4-5.

Figure 4-5 Bar graph displaying average time spent on statement screen by purported theorem and participant type



The general trend of averages indicates the possibility that novices spend more time on the statement screen than experts. The one exception occurred on Miquel Point, which is the most complicated statement presented in this experiment. This may be explained in terms of familiarity. It is possible that when an expert encounters a statement of a relatively unfamiliar nature more time is devoted to ensure proper understanding than when a novice encounters something similarly unfamiliar. This concluded the analyses for our second research question.

Research question 3: Diagram usage

Visualization provides important insight into the realm of mathematics. The mathematical community's understanding of the particular role it plays in proof is lacking. Important work has been conducted exploring the incorporation of visual arguments. The study of inferences drawn from visualizations has found that students do not condone direct utilization of the appearance of visualizations. The persuasiveness of standalone visual arguments is poor among novices and more acceptable in experts. This, however, is contrasted by the fact that novices are more inclined to affirm statements when provided an illustration. Nonetheless, the

role diagrams play in the proof validation process is not well understood. Our exploration of diagram usage begins by analyzing fixation durations and is then followed by studying attentional changes.

Through the use of AOIs in identifying fixations occurring in the diagram or in particular lines of text, we were able to compute the proportion of dwell time within two categories: text and diagram. Total dwell time was computed by summing the dwell time of each fixation on the proof screen and category dwell time was computed by summing the dwell time of each fixation within that particular category. The proportion of these provided a normalized measure as discussed in General Analyses. The results of a mixed factorial ANOVA for the proportions of diagram and text dwells times are reported in table 4-13.

Table 4-13 Results of mixed factorial ANOVA for proportional fixation time on diagram and text

*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

Effect	Proportion of Diagram Fixation Time		Proportion of Text Fixation Time	
	F	p	F	p
Purported Theorem	F(3.65,131)=30.05†	<0.0001*	F(3.4,122)=27.37†	<0.0001*
Participant Type	F(1,36)=5.41	0.03*	F(1,36)=4.44	0.04*
Purported Theorem*Participant Type	F(3.65,131)=1.57†	0.19	F(3.4,122)=1.96†	0.12

Significant main effects were found by both purported theorem and participant type. Differences in both proportions across purported theorems are expected. The complexity of diagrams, number of diagram references, and amount of written content varies across purported theorems. The main effect by participant type establishes differences between novices and

experts in at least one validation for each measure. These main effects were thus further analyzed using a one-way ANOVA on each of the purported theorems. When the homogeneity of variance was violated, Welch's t-test was conducted instead. The results of these analyses are presented in Figure 4-6 and Table 4-14.

Figure 4-6 Bar graphs displaying average fixation time proportions for diagram and text by purported theorem and participant type

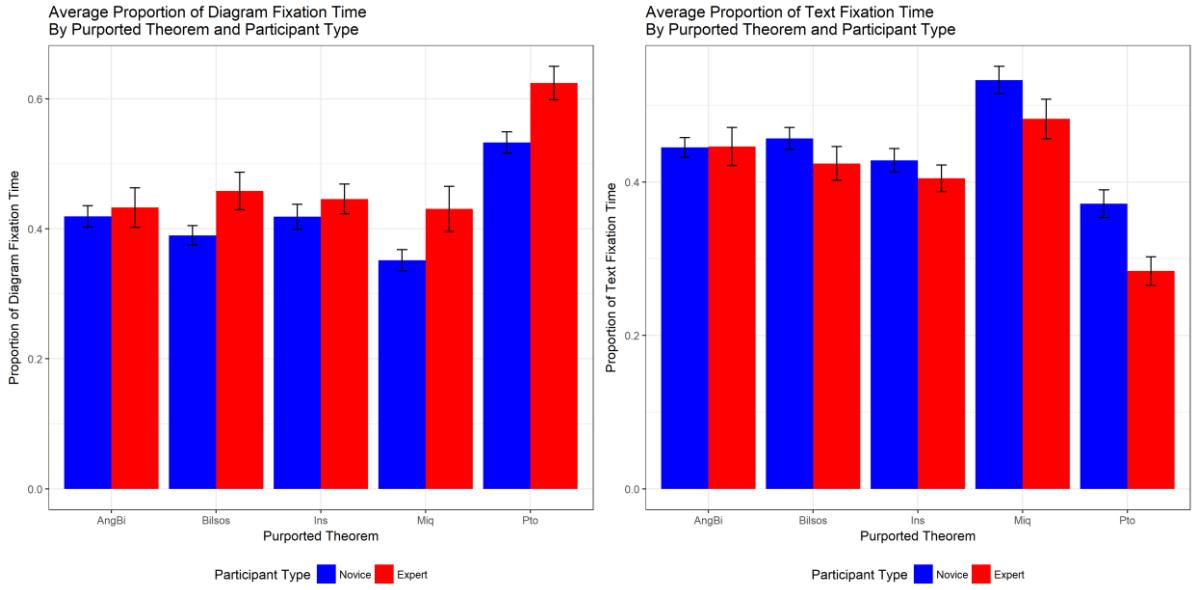


Table 4-14 Results of one-way ANOVAs[†] for average proportions of fixation time in diagram and text

[†]Welch's t-test was used when homogeneity of variance was violated (Miq-Diagram)

[*]Indicates a significant difference at the $\alpha = 0.05$ level.

Purported Theorem	Measure	Novice	Expert	F[†]	p
AngBi	Diagram	41.92%	43.29%	F(1,39)=0.189	0.666
	Text	44.51%	44.61%	F(1,39)=0.002	0.968
BiIsos	Diagram	38.99%	45.81%	F(1,38)=5.451	0.0249[*]
	Text	45.66%	42.40%	F(1,38)=1.671	0.204
Ins	Diagram	41.85%	44.58%	F(1,37)=0.737	0.396
	Text	42.81%	40.47%	F(1,37)=0.894	0.351
Miq	Diagram	35.16%	43.08%	t(18.87)=4.278[†]	0.0525
	Text	53.28%	48.21%	F(1,38)=2.723	0.107
Pto	Diagram	53.28%	62.43%	F(1,39)=9.689	0.00346[*]
	Text	37.14%	28.37%	F(1,39)=9.444	0.00385[*]

The results of the further analyses show that novices and experts show significant differences in diagram usage in two purported proofs, Bisector Isosceles and Ptolemy. A nearly significant difference was additionally shown in Miquel Point, p= 0.0525. The general trend displayed shows that experts tend to spend proportionally more time on the diagram. A significant difference was also shown to exist in the proportion of time spent on the text of Ptolemy. This was the only proof to reach significance despite novices displaying a general trend of higher proportions of time spent on the text.

Given the uniformity displayed within these overall trends and the general nature of these measures, we computed the average proportions of fixation time spent on the diagram and again on the text. These averages were then analyzed with the use of a one-way ANOVA for average proportional fixation time in the diagram and in the text. We found significant differences in both measures.

Table 4-15 Results of one-way ANOVA for average proportional fixation time on diagram and text

*Indicates a significant difference at the $\alpha = 0.05$ level.

Average Proportion of Fixation Time	Novice	Expert	F	p
Diagram	42.36%	47.97%	F(1,39)=6.984	0.012*
Text	44.55%	40.81%	F(1,39)=4.759	0.035*

The following analyses center on the attentional changes involved in the validation process. Saccades are determined by the motion of the eye and have both source and target locations. The source and target typically provide the information needed to assert an attentional change, but there are cases where one of the fixations occurs in blank space. We argue that blank space fixations occur for one of two reasons. Either the participant needs a brief reprieve from the validation process or the participant desires to further process information without the influences of visual stimuli. In both cases the omission of the blank space fixation would preserve the attentional change. In the case of mental reprieve, the participant went from focusing on one aspect to focusing on nothing to focusing on another. The fixation in blank space is thus of little consequence. In the latter case, the participant is continuing the mental processing and connecting of the information initiated during the last fixation just devoid of the visual distraction. Furthermore, this processing would be the instigator for the next fixation. Thus by omitting fixations located in blank space, we may accurately track attentional changes by classifying consecutive fixation locations.

Continuing to explore the incorporation of diagram usage in the validation process, we turned to attentional changes relating to the diagram. These attentional changes either connect the diagram with the text or connect components within the diagram. We thus classified attentional changes that included the diagram as either diagram-to-diagram, D-D or text-to-

diagram, T-D. As we are interested in the connecting of text and diagram, the T-D measure includes counts both of T-D and D-T attentional changes. The results of a mixed factorial ANOVA for the D-D and T-D measures are reported in table 4-16.

Table 4-16 Results of mixed factorial ANOVA for D-D and T-D proportional attentional changes

*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

Effect	Diagram-to-Diagram		Text-to-Diagram	
	F	p	F	p
Purported Theorem	F(3.51,126)=50.70†	<0.0001*	F(3.64,131)=25.52†	<0.0001*
Participant Type	F(1,36)=4.28	0.05*	F(1,36)=0.01	0.91
Purported Theorem*Participant Type	F(3.51,126)=2.17†	0.08	F(3.64,131)=5.25†	0.0009*

Significant main effects by purported theorem were found in both measures. As previously discussed these differences are not unexpected and are not further analyzed. The significant main effect of both participant type and purported theorem indicates that the relationship between participant type and the T-D measure is different across purported theorems. The interaction was not further analyzed. A significant main affect by participant type in the D-D measure was found. This indicates that on at least one purported theorem experts and novices displayed a significant difference in the proportion of within-diagram attentional changes made. This was further analyzed with the use of a one-way ANOVA for each purported theorem. When the homogeneity of variance was violated, Welch's t-test was conducted instead. The results of these further investigations are reported in Figure 4-7 and Table 4-17.

Figure 4-7 Bar graphs displaying average proportion of D-D and D-T attentional changes by purported theorem and participant type

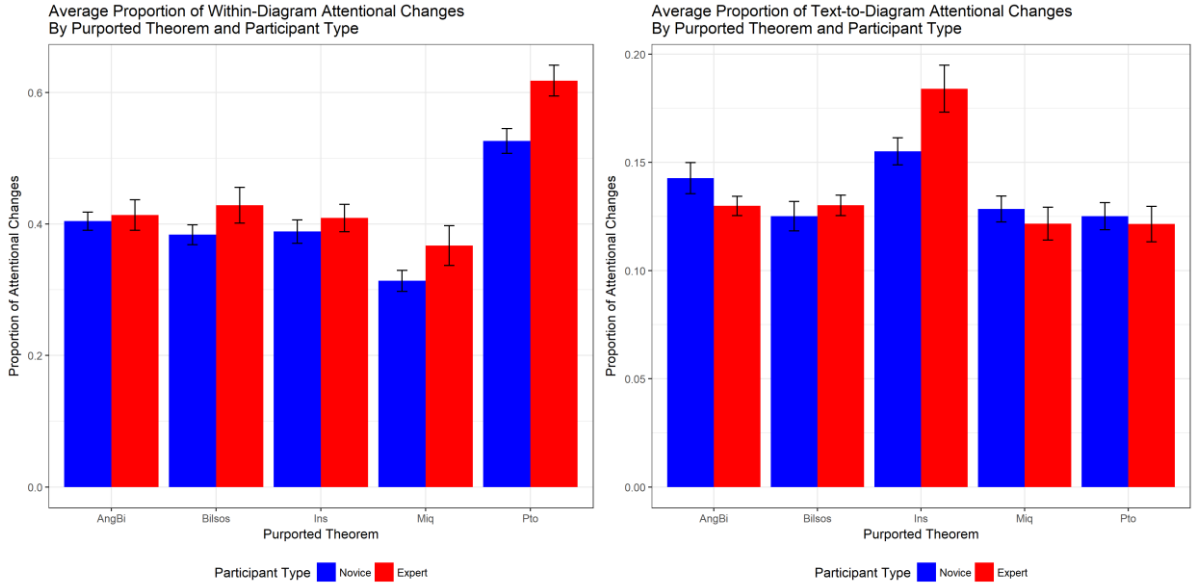


Table 4-17 Results of one-way ANOVAs[†] and average proportions of within-diagram attentional changes

[†]Welch's t-test was used when homogeneity of variance was violated (Miq)

[*]Indicates a significant difference at the $\alpha = 0.05$ level.

Purported Theorem	Diagram-to-Diagram proportion of attentional changes			
	Novice	Expert	F[†]	p
AngBi	40.42%	41.36%	F(1,39)=0.137	0.713
BiIsos	38.36%	42.84%	F(1,38)=2.482	0.123
Ins	38.84%	40.90%	F(1,37)=0.496	0.486
Miq	31.35%	36.71%	t(20.34)=2.444[†]	0.133
Pto	52.62%	61.81%	F(1,39)=8.627	0.00553[*]
Average	40.50%	44.81%	F(1,39)=4.59	0.0385[*]

The general trend of experts making proportionally more within-diagram attentional changes reaches significance only once across all purported theorems. Again due to the consistency of the trend across all purported theorems and the general nature of the analysis, we compared the average proportion of attentional changes within the diagram and found that they are significantly different. These results are also reported in Table 4-17.

Significant work has been conducted supporting the fact that novices have difficulties justifying warrants. The eye tracking study of Inglis and Alcock (2012), found that expert participants displayed more between-line saccades and attributed that to an increased presence of warrant seeking behavior. We thus ran attentional change proportion comparisons only including attentional changes with lines requiring warrants as sources. Two comparisons were made: warranted lines to diagram and warranted lines to other lines. The results of a mixed factorial ANOVA for each measure are reported in Table 14-18.

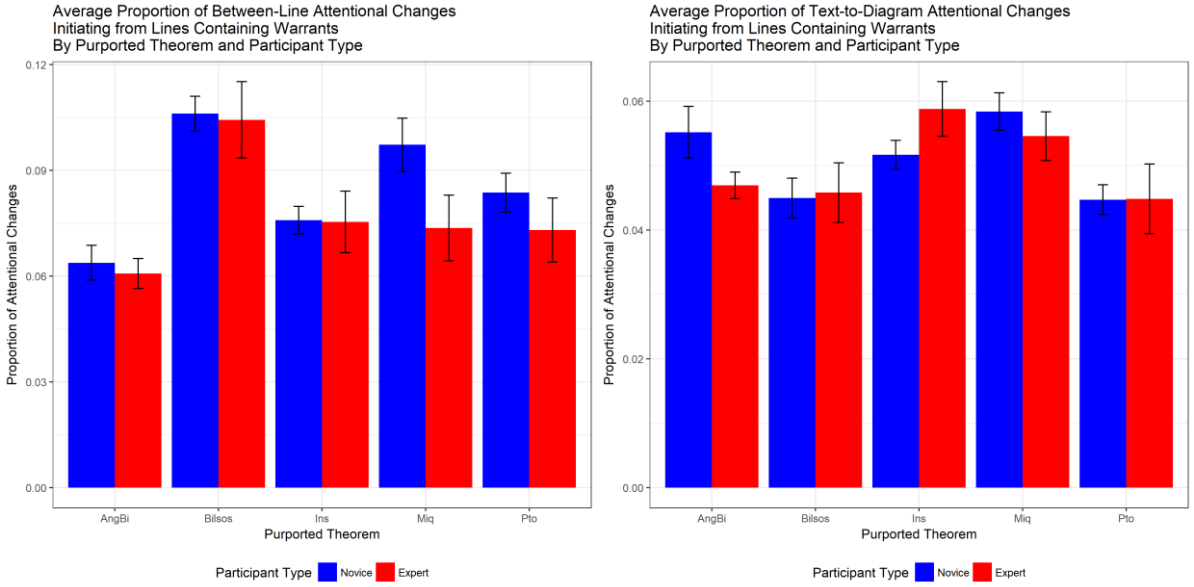
Table 4-18 Results of mixed factorial ANOVA for proportion of attentional changes initiating from a line requiring a warrant to either another line or to the diagram

*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

Effect	Warranted Between-Lines		Warranted Line-to-Diagram	
	F	p	F	p
Purported Theorem	F(3.75,134)=14.25†	<0.0001*	F(3.81,137)=0.34†	0.0001*
Participant Type	F(1,36)=0.97	0.33	F(1,36)=0.34	0.56
Purported Theorem*Participant Type	F(3.75,134)=1.11†	0.35	F(3.81,137)=2.29†	0.09

Figure 4-8 Bar graphs displaying the average proportions of warranted attentional changes between-line and text-to-diagram



Significant main effects by the purported theorem were again found, but not further analyzed. The lack of a significant effect by participant type contrasts the findings of Inglis and Alcock (2012) who found that experts make significantly more between-line saccades, in both proportion and raw count. The general trend of the between-line attentional changes even indicates the possibility of opposite of what they concluded.

One possible explanation is the introduction of the diagram. Attentional changes targeting the diagram may have several different intents. They may serve to identify and clarify the statements made in the particular line or they may serve in connecting this line to previous concepts. In essence, attentional changes may serve the same purposes as within or between-line saccades did in the Inglis and Alcock study. We however cannot distinguish between the two in our current setting.

The conclusion regarding warrant seeking behavior originated from the analysis of a single theorem addressing the issue in depth. Lines were identified as requiring a warrant or not. Warrant seeking was classified as a three fixation cycle that originated on a line requiring a

warrant, went to the previous line, and returned to the original line. These instances were then counted and compared.

In an attempt to thoroughly examine the possibilities, we conducted a similar warrant seeking analysis. To account for discrepancies in total saccade counts, we dealt with proportions instead of raw counts. We furthermore addressed one of their cited limitations by defining warrant seeking without the adjacency restriction. This was not a full remedy for it does not include the possibility of using multiple distinct lines in the justification of the warrant. To prepare for this analysis, consecutive fixations within the same AOI were treated as single fixations. This resulted in a sort of macro-level scan path indicating attentional changes. Warrant seeking triples were then identified. The total number of possible warrant seeking triples was calculated in order to compute the proportion. Further classification of the triples was made to indicate the use of the text or diagram in seeking the justification for the warrant. The interest was in the commonality of using the diagram to justify warrants. The proportion is the ratio of the number of warrant seeking triples using the diagram to the total number of warrant seeking triples. The results of a mixed factorial ANOVA for each measure is presented in Table 14-19.

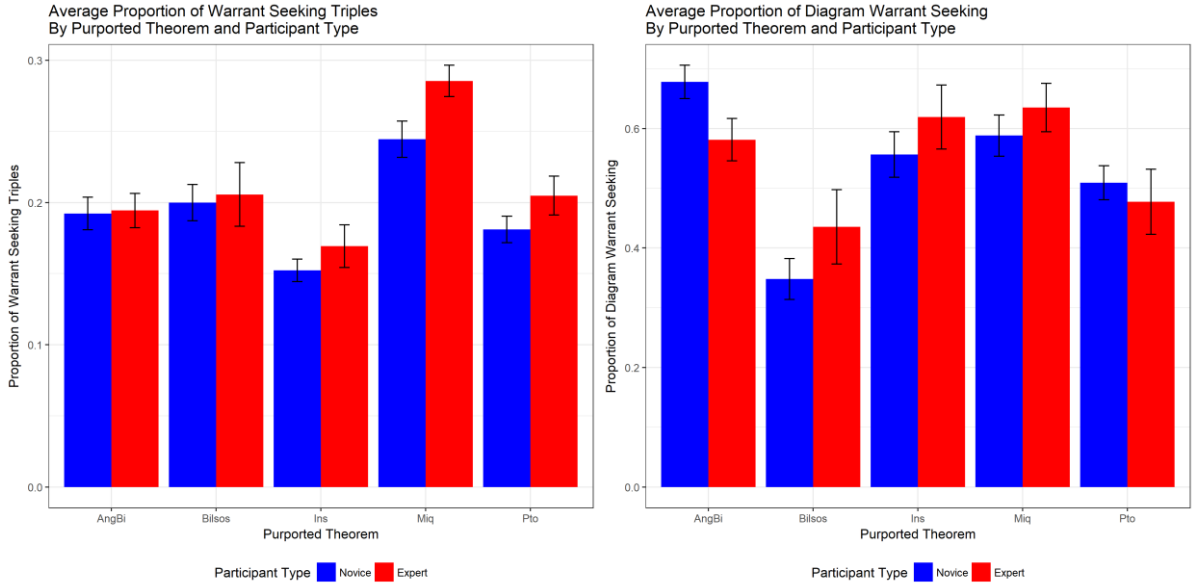
Table 4-19 Results of mixed factorial ANOVA for proportion of warrant seeking triples and use of diagram in those warrant seeking triples

*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

	Proportion of Warrant Seeking Triples		Proportion of Warrant Seeking Triples Referencing the Diagram	
Effect	F	p	F	p
Purported Theorem	F(3.57,128)=19.49†	<0.0001*	F(3.20,115)=14.62†	<0.0001*
Participant Type	F(1,36)=1.54	0.22	F(1,36)=0	0.99
Purported Theorem*Participant Type	F(3.57,128)=1.01†	0.40	F(3.20,115)=2.36†	0.07

Figure 4-9 Bar graphs displaying the average proportions of warrant seeking triples and the proportion of diagram use in their justification



Significant main effects by the purported theorem were again found, but not further analyzed. Again the lack of a significant main effect by participant type in warrant seeking triples is contrary to the findings of Inglis and Alcock (2012). The general trend does however indicate the same conclusion. The justification of warrants tends to be comparable by expertise and utilizes both the text and the diagram similarly. This analysis concluded our investigation into our third research question.

Unplanned Tangential Analyses

This chapter concludes with the presentation of two analyses that resulted from conducting our research question oriented analyses. The impetus for the first analysis is related to the proportions of textual and diagrammatic dwell time. It was noted that a nontrivial amount of time must be spent in the blank spaces of each theorem to result in presented averages. This was an unexpected occurrence as it was thought the two would vary but occupy close to the total

fixation time. We thus compared the dwell time in the nonAOI, the blank spaces of the proof screens.

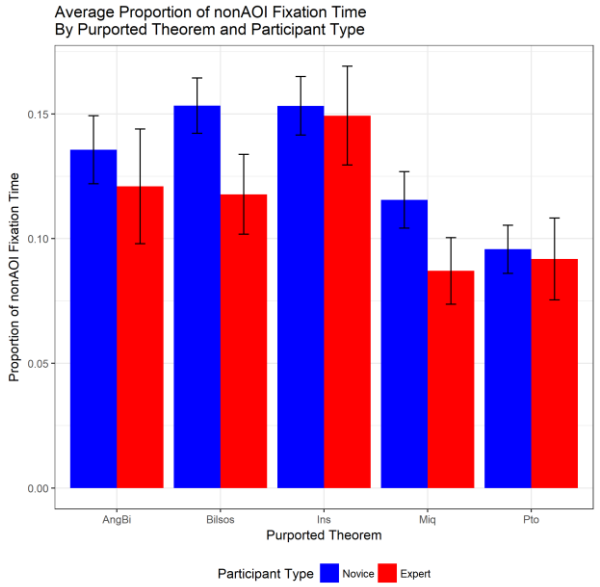
Table 4-20 Results of mixed factorial ANOVA for proportion of fixation time spent in blank space during the proof validation

*Indicates a significant difference at the $\alpha = 0.05$ level.

†Indicates use of the Greenhouse-Geisser Correction for sphericity

Effect	Proportion of Fixation Time Spent in Blank Space	
	F	p
Purported Theorem	F(3.49,125)=6.88†	0.0001*
Participant Type	F(1,36)=1.03	0.32
Purported Theorem*Participant Type	F(3.49,125)=0.55†	0.68

Figure 4-10 Bar graphs displaying the average proportions of fixation time spent in blank space



Significant main effects by the purported theorem were again found, but not further analyzed. While no significant main effect of participant type was reached, the overall nontrivial presence of blank space fixations and the general trend presented is interesting. As discussed previously, there are two primary explanations for these sorts of fixations: mental reprieve and

desire for visually unstimulated thought. Which is the source for the displayed disparity between novices and experts?

The second analysis stemmed from the process of identifying warranted statements within each proof. We decided to classify statement warrants as implicit or explicit. On a couple of occasions, lines were classified as both. See Appendix B for a full tabulation of line classifications. We decided to then compare these with our already tabulated justification data for assertions of invalid.

Table 4-21 Tabulations of asserted errors by classification of warrant type

Purported Theorem	Novice				Expert			
	Implicit	Explicit	Both	Unwarranted	Implicit	Explicit	Both	Unwarranted
AngBi	8	1	0	0	0	0	0	0
BiIsos	1	0	5	2	0	10	0	0
Ins	3	3	0	0	0	0	0	0
Miq	8	2	0	0	2	8	0	0
Pto	5	5	0	0	2	2	0	2

In justifying their assertion of invalid, novices sighted the majority of the time (30 of 44) a line containing implicit warrants. The struggle novices have with justifying implicit warrants is well documented and this bolsters that conclusion. This may also point to the fact that novices are more inclined to accept explicit justification. Take for example, not a single novice identified the logical error made in the explicit justification in Miquel Point. Experts, on the other hand, display a nontrivial inclination to lines containing explicit warrants (20 of 26). This could be reasonably explained in a couple of ways. The presence of a justification may indicate something nontrivial is occurring. Under that assumption, if the expert validator does not or cannot verify the justification, the validator may conclude, possibly from grading experience, that the author made an incorrect or too large of a jump in reasoning. Arguments generally left for implicit justification tend to have a familiarity associated with them that possibly lets the experts verify

them more readily and thus are cited less. These additional analyses conclude our discussion of our analysis of the data.

Chapter 5 - Conclusions

Overview of Research

The purpose of this research endeavor was to investigate the proof validation processes of expert and novice mathematicians. Participant eye movements, recorded from an eye tracker during each validation, served as our primary data set. Having been primarily studied using verbal protocols, the hypothesized strategy of *zooming out* was tested for its occurrence in practice. Our investigation sought to settle the considerable dissention in the literature surrounding this particular strategy. We examined the overall similarity of validation processes by mathematical maturity. We also expanded the overall scope of the proof validation literature by including a focus on the utilization of diagrams which has received little study within proof validation literature.

Research Questions Answered

Research question 1: *Zooming out*

Through the dissection of think-aloud studies and retrospective interviews, a seemingly thorough catalog of individual strategies implemented in the validation process has been composed. These strategies vary from use of example(s) to failure to find counterexamples to further deductive reasoning and construction of subproofs. These highly specific strategies were generalized, based upon how one views proof, into three classifications: proof as a cultural artifact, proof as a series of inferences (*zooming in*), and proof as an application of methods (*zooming out*) (Weber and Mejia-Ramos, 2011; Mejia-Ramos and Weber, 2013). The confirmation of zooming out techniques through alternative measurements was sought. Inglis and Alcock (2012) in the first direct comparison of novice and expert proof validations using eye tracking sought to test for the occurrence of a particular zooming out technique where

participants quickly read the entire proof to gauge the argument's structure: the initial skim strategy. The analyses conducted through a series of contesting articles concluded without resolution. Opposing authors, using the same data set, reached polar opposite conclusions. Our first research question asks if there is evidence supporting *zooming out* occurring in practice. We sought to answer this question by testing for two specific strategies attributed to the generalized *zooming out* strategy: initial skim-reading and continued forward progression after encountering an error.

Testing for initial skim-reading was one of the primary motivations for our research endeavor. We therefore, incorporated remedies to the many and valid concerns presented in the literature. We presented longer, complicated proofs from a mathematical field of relative unfamiliarity that were devoid of blatant errors. Furthermore IR ratios and their distributions were analyzed at the purported theorem level without aggregation distorting the possible conclusions.

Given the disparity in the literature, we adopted two competing hypotheses regarding the implementation of an initial skim-reading. Our first hypothesis formed its basis from the work of Weber (2008) and Mejia-Ramos and Weber (Weber, Mejia-Ramos, Inglis, and Alcock, 2013). We hypothesized that initial skim-reading would be implemented resulting in either right skewed distributions of IR ratios with averages significantly less than 50% or non-incidental bimodal distributions with one mode significantly less than 50%. We further hypothesized with the work of Inglis and Alcock (2012) that initial skim-reading would not be implemented resulting in unimodal distributions with averages not significantly less than 50%.

We found that the distributions of IR ratios were primarily unimodal with few participants exhibiting IR ratios significantly less than 50%. The few distributions displaying the

possibility for multiple modes (Angle Bisector, Miquel Point, and Ptolemy) each were distributions of novice validators with modes not significantly less than 50%. Given the proximity of these distributions to 50% IR ratios, identifying small IR ratios due to incidental or random fixations at the end of the proof was unnecessary. We thus reject the hypothesis that *zooming out* is generally implemented using an initial skim-reading strategy in the context of proofs from Euclidean geometry.

Despite providing longer proofs without immediately recognized mechanisms, it is possible that the sorts of proofs requiring or promoting an initial skim-reading were not provided. Firstly, proofs were required to fit entirely on one page with five proof validations within an hour. Secondly, proofs were written for an audience of both novices and experts. The study prompting initial skim-reading as a strategy consisted of two sets of validations: one of undergraduate number theory and one of advanced number theory. These latter proofs may have prompted the references to an initial structure-oriented read through.

Our second investigation into the particular usage of the *zooming out* strategy sought to test for a strategy not explicitly catalogued in the literature. The method is centered on what validators do when the truth of an individual statement is uncertain. The literature identifies several means by which validators build confidence or establish these uncertain statements. Some participants terminate the validation process upon identifying such a statement; others use examples and deductive reasoning to bolster belief (Selden and Selden, 2003; Weber, 2008; Ko and Knuth, 2013). Whether these actions include insight from the remainder of the proof is not clarified. Mejia-Ramos and Weber provide a hypothetical example illustrating the possibility of continued progression. Given a blatant error, an initial skim-reader may identify said error and upon completion of the proof terminate the validation quickly. They used this example to counter

the IR ratio, but it begs the question as to why would the reader upon identifying the error continue with the remainder of the proof. As discussed previously, continued forward progression after encountering an error displays use of the strategy of *zooming out*. Either the participant questions the legitimacy of the error and seeks further big picture perspective or is gauging the importance of that statement in the overall proof. Within our study, identification of such behavior is attainable only in cases were the validator asserted the proof to be invalid. The combination of verbal justification and eye movement enabled the checking for continued forward progress. Recall fixation proportions less than 1% were classified as incidental and were not used to support continued progression.

Again two competing hypotheses were adopted for this analysis. Given the lack of any explicit reference to such behavior in the literature, we hypothesized that continued forward progression after encountering an error or uncertainty would not occur. This is indicated by only incidental fixations occurring below the line cited as containing an error. In direct opposition, we hypothesized that, given the hypothetical from Mejia-Ramos and Weber, participants would engage in *zooming out* after encountering an error or uncertainty in a proof in order to better gauge the situation.

We found that for each theorem the majority of validators continued progressing through the proof after encountering an uncertainty or error. Keep in mind this analysis only used validators asserting that the proof was invalid and that the justification didn't cite the last line of the proof. We therefore reject the hypothesis that continued forward progression after encountering an error or uncertainty does not occur.

It is possible that recognition of the error is not immediate. This is more probable in our study than in the previous studies as we intentionally avoided blatant errors. This being the case,

forward progress may be expected. We therefore further classified the continued progress as those who continue on but cease shortly after the error and those who continued through to the end of the proof. Again for each purported theorem the majority of validators completed the entire proof. See Appendix B for the complete breakdown. It is possible that this merely indicates the perception that in order to conduct a proper validation a proof must be read in its entirety. This notion is unexpected and questionable but warrants further investigation. In the interviews conducted by Selden and Selden (2003), the interviewer did on occasion prompt students to continue the validation process indicating this notion may apply.

The overall interest in our first research question was about the practical nature of the *zooming out* strategy. We sought to test two specific implementations. Although initial skim-reading was not implemented in our Euclidean geometric proof validation attempts in general, the possibility for it occurring provided even more complicated and lengthier proofs remains. We have confirmed the general use of continued progression through a proof despite encountering an error or uncertainty. We thus conclude that *zooming out* is indeed a method utilized in practice.

Research question 2: The validation process

Significant work has been done analyzing particular differences in the validation processes of novices and experts. Differences in attention allocation and particular types of saccades have been noted (Inglis and Alcock, 2012). We have confirmed the long standing trend of novices doing poorly as compared to their expert counterparts in validation accuracy. These differences however fail to describe the overall disparity in validation processes. With neither temporal nor sequential referencing, these measure many isolated actions, not the fluid validation process as a whole. Our second research question asks about the overall validation processes and whether they display significant disparity.

The pursuit of this question required the ability to encode entire validation processes and then compare them in some meaningful manner. Validation processes can be encoded as scan paths, which are sequences of fixations and their associated durations. When given two pairs of these scan paths, the ScanMatch algorithm, similar to the algorithm comparing DNA, returned a normalized similarity score which we transformed into a dissimilarity score. These dissimilarity scores were collected to form a dissimilarity matrix for each purported theorem. Interested in not only the overall validation processes by purported theorem, but throughout each individual session, we also treated the dissimilarity data as five-dimensional data and computed overall dissimilarities at the experimental level. Two overall dissimilarities were calculated using two common norms.

For each purported theorem, we then calculated for each participant the average dissimilarity within-group and between-group. These within-group and between-group average dissimilarity sets formed the basis for our first investigation. The comparison of the within-group and between-group average dissimilarity distributions exhibits the general proximity of the overall validations by group.

We hypothesized that due to experts' warrant seeking behavior (Inglis and Alcock, 2012) and their greater trust in visual arguments (Inglis and Mejia-Ramos, 2008), expert validation attempts would be significantly different from novice attempts. This would be exemplified by a significant difference between the within-group and between-group distributions with the within-group exhibiting a smaller average.

We now report our findings by participant type first and then enter into the broader discussion. We found that expert within-group averages were generally comparable to the between-group averages. Any differences exhibited failed to reach significance across all

purported theorems. We found that novice within-group averages were bigger than the between-group averages in general (4 of 5 purported theorems). This trend reached significance on 2 of these purported theorems. This tells us that in general the experts are about as similar among themselves as they are with the novices and that the novices tend to be closer among the experts than they are among themselves. It is possible that in a larger study conducted with more participants significant differences would be found across more purported theorems in the novice comparison. The fact that novices tend to be closer to experts than themselves is an interesting result.

The distribution analysis conducted at the session level returned similar results for both norms. None of the subtle differences reached statistical significance. This indicates a general well mingling of participants regarding their overall validations throughout the whole session. Incorporating the results at the individual purported theorem level and session level, we do not support the hypothesis that experts and novices partake in significantly different proof validation processes.

These results speak only of the general proximity of expert and novice proof validations. The possibility of the existence of validation processes unique to participant type remains. To investigate this possibility, we clustered participant validations utilizing a combination of two techniques. The use of agglomerative nesting identified reasonable cluster counts and identified particular validations that were anomalous. These insights were applied in our application of the k-medoids algorithm for assigning clusters.

We hypothesized that due to experts' warrant seeking behavior (Inglis and Alcock, 2012) and their greater trust in visual argument (Inglis and Mejia-Ramos, 2008), clusters would be homogeneous with respect to mathematical maturity.

We found that homogeneity scores ranged from 50-85% on both levels of analysis: by purported theorem and by session. Clustering was conducted using both two and four clusters at the theorem level. In the latter case, we were clustering some 40 participants into four groups. This resulted in each theorem having at least one cluster with less than 10 participants. Given clusters of this size a single participant accounts for a significant portion of the group, if only a single participant differed from the majority (3 occurrences) we conducted further analysis to see if that participant was an outlier of that cluster. This was the case only once with a cluster of size 5. We therefore do not support the hypothesis that clusters are homogeneous with respect to mathematical maturity. Given that a nontrivial number of clusters displayed homogeneity scores around the 80% mark, a study with a larger participant base may find that there are clusters that consist primarily of a single mathematical maturity.

The combination of these two analyses yields the conclusion that the validation processes of novices and experts are not significantly different. As in experts do not clearly conduct one method of validation while the novices conduct another method. Our first analysis indicates that experts are just as close to novices as they are among themselves, and the novices are closer to the experts than they are among themselves. The second analysis indicates the possibility for unique validation processes being used by both groups. Interpreting both results simultaneously yields the possibility of participants implementing distinct methods that are more similar when conducted with expertise. Whether these methods converge or remain distinct as maturity increases is unknown.

We conclude the discussion of our second research question through a different lens. We have previously been investigating the overall validation process on the proof itself. The

approach to the validation process may differ. This initial approach occurs on the statement screen when participants first see the purported theorem statement.

We hypothesized that given experts' inclination to mull theorems over and discern how they would prove a statement, experts will spend significantly more time on the statement screens.

We found that the differences displayed by participant type in viewing the statement screen failed to reach significance. The general trend however alludes to the possibility for novices spending more time than the experts (4 of 5). The exception to this trend occurred on the most complicated and unfamiliar theorem statement, the Miquel Point. We therefore reject the hypothesis that experts spend significantly more time viewing and pondering the statement screen.

The exception may be explained through familiarity. It is possible that experts spend more time understanding that which they are not familiar than the novices do. Given the relative commonality of the other statements as far as structure, it is possible that near immediate understanding of the statement produced these lower averages.

The general interest of our second research question was the overall validation process. We found that overall validation processes of experts and mathematicians are comparable. Average dissimilarity distributions indicate a close proximity of expert and novice validations. The general heterogeneity of clusters by mathematical maturity further promotes this close proximity. And expert and novices generally spend a comparable time mulling over the statements of theorem. We thus conclude that expert and novice validation processes are not unique to the mathematical maturity of the validator.

Research question 3: Diagram usage

The use of visualization plays an important role in the learning and practice of mathematics. Various aspects of its use and persuasiveness have been studied, but its general role in the validation process remains unknown. Novices, having been taught that visualizations are not proofs, tend to display more of an aversion to trusting solely in a diagram than experts (Inglis and Mejia-Ramos, 2008). Novices however are swayed by the mere addition of an illustration to an argument (Nyström and Ögren, 2012). Our third research question asks about the differences of diagram usage between novice and expert validators. Diagram usage can be analyzed in two methods: prominence of the diagram in fixation time and in attentional changes.

The proportions of time spent on the diagram and text indicate the participant's allocation of attention and hence what is thought to be important and useful. Given that novices display a general skepticism for accepting arguments based solely upon visualizations (Inglis and Mejia-Ramos, 2008; Zhen, Weber, and Mejia-Ramos, 2015), we hypothesized that experts would spend significantly more time on the diagram than novices do. Furthermore, since novices focus primarily on surface features like equations and symbols (Selden and Selden, 2003; Inglis and Alcock, 2012), we hypothesized that novices would spend significantly more time on the text than experts do.

We found that the general trends matched well with our hypotheses. Experts tended to spend more time proportionally on the diagram while the novices tended to spend more time proportionally on the text. These trends reached statistical significance twice regarding the diagram (nearly a third time on Miquel) and once regarding the text. Only a single purported theorem, Ptolemy, reached significance for both. Given the overall consistency in the general trends presented in the proportions of time spent on the diagram and on the text and the general

nature of the comparisons, we computed average proportions of time spent on the diagram and text. These averages were statistically significant. We therefore conclude that experts spend a greater proportion of their time on the diagram while the novices spend a greater proportion of their time on the text.

Fixation duration is not the only way to measure the importance and use of diagrams. Attentional changes provide information for how information is digested and connected. Since mathematicians display greater trust in visual argument than novices (Inglis and Mejia-Ramos, 2008) and have a deeper understanding of mathematics (NRC, 2000), we hypothesized that experts will make more attentional shifts proportionally to the diagram and within the diagram than the novices.

We found that the general trend of within-diagram attentional change proportions matched our hypothesis. These differences however reached significance only once on Ptolemy. Given the consistency of the trends and the generality of our analysis, we compared the average proportion of within-diagram attentional changes over all the theorems and found the difference to be significant. We therefore conclude in agreement with our initial hypothesis that experts spend a greater proportion of the attentional changes drawing connections within the diagram.

No general trend is apparent in the text-to-diagram attentional changes and none of the differences were statistically significant. The two instances where experts exhibited a higher proportion of text-to-diagram attentional changes occurred on the two lengthiest proofs (3 or 4 lines longer than the rest) with the busiest diagrams. This discrepancy may be explained by the combination of two ideas. Novices look at previously coined surface features indicating that they may look at the diagram primarily as references. This would explain the higher proportion on the shorter proofs with less busy diagrams. They would have to continually reference the diagram

whereas experts would hold it in their mind better. On the other longer proofs with busier diagrams, experts needed to make more back and forth attentional changes because they couldn't hold it all in their heads.

Inglis and Alcock (2012) concluded that experts made significantly more between-line saccades than novices due to their warrant seeking behavior. Given this basis, we hypothesized that experts would make significantly more attentional changes proportionally from warranted lines either to the diagram or another line.

We found no differences of statistical significance in either the between-line attentional change proportions or the between line and diagram attentional change proportions. We therefore cannot support the hypothesis that experts display these behaviors in greater proportions than novices. The line to diagram attentional change proportions displayed similar characteristics to the unwarranted case discussed above. The attentional changes between diagram and warranted lines displayed a general trend with novices actually displaying a higher proportion. This is contrary to the findings of Inglis and Alcock (2012). A reasonable explanation for this difference, although statistically insignificant, could reside in the introduction of the diagram. Attentional changes to the diagram are meant either to understand that particular line or draw connections with other content. This means that the role of attentional changes to the diagram may serve the purpose of either within-line or between-line saccades as categorized in the literature. In our study, differentiating the two purposes is not possible.

Given the results of the previous analyses, we performed analyses similar to those conducted by Inglis and Alcock (2012) to test for warrant seeking behavior. Warrant seeking behavior is classified as attention change triples initiating on a warrant requiring line followed by an attentional shift away from that line and then returning to the original line. We conducted our

analysis again proportionally, while the literature used raw counts. We found no overall significant differences by participant type, but the general trend supports the conclusions of Inglis and Alcock (2012). Experts make more of these sorts of behaviors. We also failed to find a significant difference between expert and novice use of the diagram in these warrant seeking triples. The use of the diagram and other lines in these warrant seeking attempts varies across the purported theorems which is to be expected.

The interest of our third research question resided in the usage of the diagram during the proof validation process. We found that experts spend a greater proportion of their time fixating on the diagram while novices spend a greater portion of their time fixating on the text. Experts also devoted a greater proportion of their attentional changes to drawing connections within the diagram. Through our exploration of warrants, the general trend of our warrant seeking behavior results mimics those of Inglis and Alcock (2012), although ours fails to reach significance. We thus conclude that experts make greater utilization of diagrams than novices. They spend more time on the diagram and they dissect the diagram through more within-diagram attentional changes.

Limitations

In this section we expound upon the limitations of this particular research endeavor and how they might be addressed in the future. On several occasions comments were made about the deliberate terseness of the presented proofs. It is possible that the terseness resulted in an uncomfortable or unnatural validation process which in turn resulted in altered validation processes. A second possible factor leading to alteration of validation processes was the inability of the participant to use paper and pencil. It is not uncommon for mathematicians to utilize these during validation attempts. This, in combination with the brevity of our written proofs, may have

discouraged the construction of subproofs and exploration that would normally occurred. On several occasions statements regarding the desire for further exploration through different diagrams were made. Participants thought that individual construction of diagrams and references would have facilitated progression through the proof.

These overall limitations may be addressed through the implementation of a tablet which allows participants the use of writing materials and accurate tracking of participant progression. It may be possible to then display the tablet display on the screen to allow for a continuity of cognitive measurement. While the participant is writing, the cognitive processes would be known as they are recorded on the tablet, and then eye tracking would be recording the rest of the cognitive processes. This process would however greatly increase the complexity of the experimental design and analysis. A slightly more manageable solution could be the introduction of either progressive diagrams which change as the validator progresses through the proof or fixation-triggered highlighting within the diagram which highlights referenced entities.

The construction and presentation of convincing yet false arguments is another limitation of this study. We successfully constructed a single invalid proof for this research endeavor. All of the invalid proofs from the pilot study needed reworking. Future work should incorporate a more balanced design (not 4:1 valid to invalid ratio) to help account for bias as displayed by the novices.

The size of the study ended up being another limitation. During the clustering analysis, cluster sizes less than 10 occurred with a greater than desired frequency. Each participant had a significant effect on the overall homogeneity of these small clusters. The classification of primarily novice or expert would be reasonably attainable had cluster sizes been larger.

Future Work

With these limitations in mind, we now look forward to the possibilities of future research. In this research endeavor we tested for specific implementations of the *zooming out* strategy. Despite using longer and more complicated proofs, we did not find evidence for initial skim-reading in practice. The only evidence for this strategy stems from expert validators, so the use of an expert only oriented validation experiment with even longer proofs may establish its use. The second implementation of *zooming out* was the continuation beyond after encountering an error or uncertainty. This is the first explicit observation of such strategies in the validation process. It is possible that this is not a naturally occurring strategy but is the result of the terse structure without the ability to write. Further study into its occurrence is warranted. It is possible that it stems from the questionable and unexpected notion that in order to complete a proper validation one must read the entirety of the proof. If this strategy is indeed natural, we know little about its use. Is it primarily a strategy for overcoming impasse? Does it only occur after several attempts at resolution using the previous statements of the proof? Many different facets of study revolve around this strategy. The introduction of minimal verbal protocols would be instrumental in the identification of when participants encounter issues and how they proceed. This method could be beneficial for proof validation exploration in general.

Through the direct comparison of overall validation processes we have shown that novices and experts do not exhibit unique validation processes based upon their mathematical maturity. Our cluster analysis indicates distinct validation processes that grow in similarity as participants grow in mathematical maturity. These distinct validation processes are not unique to experience or assertion. Can we differentiate the various distinct processes and then identify

accurate predictors? As similarity of process seems to grow with maturity is there an idealized validation process?

We have seen that experts hold an increased importance of the information displayed in diagrams and that they look around the diagram to a much greater extent. How can we, as educators, elicit such behavior in our pupils and expedite the growth toward expertise? We furthermore must question the effectiveness of our presentations that incorporate visualizations. Is the diagram as useful as we think it is? Are we really indicating the priority of and effort required for understanding the problem through the diagram? If we are not, novices who attend more to the text will not fully apprehend the mathematical processes entailed.

Ptolemy is consistently among the purported theorems exhibiting statistical differences. Understanding why Ptolemy continually elicits these differences may provide key insight into the design of future studies and the validation process in general. Many of the validators (6 of 15) who asserted invalid for this proof struggled with the application of the lemma. This proof incorporates a vast array of techniques utilizing triangles, circles, angles, and algebra all relating back to a quadrilateral. Which aspects of this theorem elicited the noted differences? How can we readily identify from the plethora of possibilities theorems and proofs that will produce meaningful insights into the validation process?

This research endeavor provides a basis for various future works of research. We have identified a specific implementation of *zooming out* not found in the previous literature. It is closely related to how people solve problems with which they struggle. When and how often is it implemented? We further confirmed the existence of distinct overall validation processes which are not predicted by expertise or assertion. What are these distinct validation processes and how can we accurately predict their use? And lastly we identified key differences in novice and expert

utilization of diagrams. Do these differences reduce our perception of the effectiveness of diagram usage in the classroom? How can we ensure that students share a similar priority and utilization of diagrams in understanding mathematics? This research endeavor therefore prompts future study of the particular actions and holistic nature of proof validation, and effective mathematical pedagogy.

References

Agra, E. (2015). *A conceptual model for facilitating learning from physics task using visual curing and outcome feedback: theory and experiments* (Doctoral dissertation). Kansas State University, Kansas, USA.

Alcock, L., Weber, K. (2005). Proof validation in real analysis: inferring and checking warrants. *Journal of Mathematical Behavior*, 24, 125-134.

Alcock, L., Hodds, M., Roy, S., Inglis, M. (2015). Investigating and improving undergraduate proof comprehension. *Notices of the American Mathematical Society, 67(7)*, 742-752.

Azzouni, J. (2004). The derivation-indicator view of mathematical practice. *Philosophia Mathematica*, 12, 81–106.

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, 56A, 1053–1077. doi:10.1080/02724980244000729

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.doi:10.1016/0364-0213(89)90002-5

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3), 692.

Davis, P.J. (1972). Fidelity in mathematical discourse: is one and one really two? *The American Mathematical Monthly*, 79(3), 252-263

Dawkins, P.C., Weber, K. (2017). Values and norms of proof for mathematicians and students. *Educational Studies in Mathematics*, 95(2), 123-142.

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras*, 24, 17-24.

de Villiers, M. (2004). Using dynamic geometry to expand mathematics teachers' understanding of proof. *International Journal of Mathematical Education in Science and Technology*, 35(5), 703-724.

Davis, D. (1949). *Modern college geometry*. Cambridge, MA: Addison-Wesley Press Inc.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12), 1827-1837.

Glaholt, M. G., Rayner, K., & Reingold, E. M. (2012). The mask-onset delay paradigm and the availability of central and peripheral visual information during scene viewing. *Journal of vision*, 12(1), 9.

Hoffman, J., Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, 57(6), 785-795.

Inglis, M., Mejia-Ramos, J.P. (2008). On the persuasiveness of visual arguments in mathematics. *Found Sci*, 14:97-110. doi:10.1007/s10699-008-9149-4

Inglis, M., Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43(4), 358-390.

Inglis, M., Mejia-Ramos, J.P., Weber, K., Alcock, L. (2013). On mathematicians' different standards when evaluating elementary proofs. *Topics in Cognitive Science*, 5, 270-282. doi:10.1111/tops.12019

Johnson, J. (2015). *Investigating visual attention while solving college algebra problems* (Master's thesis). Kansas State University, Kansas, USA.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review*, 87, 329–354. doi:10.1037/0033-295X.87.4.329

Kaspar, K. (2013). *High- and low-level factors in visual attention* (Doctoral dissertation). University of Osnabruck, Germany.

Ko, Y-Y., Knuth, E. (2013). Validating proofs and counterexamples across content domains: practices of importance for mathematics majors. *Journal of Mathematics Behavior*, 32, 20-35.

Komatsu, K., Jones, K., Ikeda, T., Narazaki, A. (2017). Proof validation and modification in secondary school geometry. *Journal of Mathematical Behavior*, 47, 1-15.

Kowler, E., Anderson, E., Dosher, B., Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, Vol. 35, No. 13, 1897-1916.

MacMillan, N., Creelman, C. (2005). The yes-no experiment: response bias. In *Detection theory: a user's guide* (27-50). Lawrence Eribaum Associates.

Madsen, A. (2013). *Studies of visual attention in physics problem solving* (Doctoral dissertation). Kansas State University, Kansas, USA.

Madsen, A. M., Larson, A. M., Loschky, L. C., & Rebello, N. S. (2012). Differences in visual attention between those who correctly and incorrectly answer physics problems. *Physical Review Special Topics - Physics Education Research*, 8(1), 010122.

Mejia-Ramos, J.P., Weber, K. (2013). Why and how mathematicians read proofs: further evidence from a survey study. *Educational Studies in Mathematics, 85*, 161-173, doi:10.1007/s10649-013-9514-2.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9853.

Nyström, M., Ögren, M. (2012). How illustrations influence performance and eye movement behavior when solving problems in vector calculus. In Proceedings: LTHs 7:e Pedagogiska Inspirationskonferens

Posner, M., Peterson, S. (1990). The attention system of the human brain. *Annual Reviews Neuroscience*, 25-42.

Rav, Y. (1999). Why do we prove theorems? *Philosophia Mathematica*, 7, 5–41. doi:10.1093/philmat/7.1.5

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422. doi:10.1037/0033-2909.124.3.372

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62, 1457–1506.doi:10.1080/17470210902816461

Rouinfar, A. (2014). *Influence of visual cueing and outcome feedback on physics problem solving and visual attention* (Doctoral dissertation). Kansas State University, Kansas, USA.

Rota, G.-C. (1991). The concept of mathematical truth. *The Review of Metaphysics*, 44(3), 483-494. http://www.jstor.org/stable/20129055

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, 17, 759–769. doi:10.3758/BF03202637

Schoenfeld, A. (1985). *Mathematical problem solving*. San Diego, CA: Academic Press

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122, 166–183. doi:10.1037/0096-3445.122.2.166

Selden, A., Selden, J. (2003). Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education, 34*, 4-36. http://www.nctm.org/[ublivations/jrme.aspx

Selden, A., Selden, J. (2015). Validations of proofs as a type of reading and sense-making. In K. Beswick, T. Muir, and J. Wells (Eds.) *Proceedings of the 39th Conference of the International Group for the Psychology of Mathematics Education*, 4, 145-152.

Siegler, R. (1987). The perils of averaging data over strategies: an example from children's addition. *Journal of Experimental Psychology: General*, 116(3), 250-264.

Smart, J. (1998). *Modern Geometries* (5th ed.). Belmont, CA: Brooks/Cole Cengage Learning.

Thomas, L., Lleras, A. (2007). Moving eye and moving thought: on the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin and Review*, 14(4), 663-668.

Thomas, L., Lleras, A. (2009). Covert shifts of attention function as an implicit aid to insight. *Cognition*, 111, 168-174.

Thurston, W.P. (1994). On proof and progress in mathematics. *Bulletin of the American Mathematical Society, 30(2)*¸ 161-171.

Weber, K. (2004). Traditional instruction in advanced mathematics courses: A case study of one professor's lectures and proofs in an introductory real analysis course. *Journal of Mathematical Behavior, 23*, 115-133. doi:10.1016/j.jmathb.200.03.001

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education*, 39, 431–459. http://www.nctm.org/publications/jrme.aspx

Weber, K., Alcock, L. (2005). Using warranted implications to understand and validate proofs. *For the Learning of Mathematics*, 25(1), 34-38, 51.

Weber, K., Mejia-Ramos, J.P. (2011). Why and how mathematicians read proofs: An exploratory study. *Educational Studies in Mathematics, 76,* 329-344. doi:10.1007/s10649-010-9292-z

Weber, K., Mejia-Ramos, J.P. (2014). Mathematics majors' beliefs about proof reading. *International Journal of Mathematical Education in Science and Technology, 45(1).* https://doi.org/10.1080/0020739X.2013.790514

Weber, K., Mejia-Ramos, J.P., Inglis, M., Alcock, L. (2013). On mathematicians' proof skimming: A reply to Inglis and Alcock / Skimming: A response to Weber and Mejia-Ramos. *Journal for Research in Mathematics Education, 44(2),* 464-475. http://www.jstor.org/stable/10.5951/jresematheduc.44.2.0464

Wu, X. (2016). *Influence of multimedia hints on conceptual physics problem solving and visual attention* (Doctoral dissertation). Kansas State University, Kansas, USA.

Zazkis, D., Weber, K., Mejia-Ramos, J.P. (2016). Bridging the gap between graphical arguments and verbal-symbolic proofs in a real analysis context. *Educational Studies in Mathematics*, 93, 155-173.

Zhao, M., Gersch, T. M., Schnitzer, B. S., Dosher, B. A., & Kowler, E. (2012). Eye movements and attention: The role of pre-saccadic shifts of attention in perception, memory and the control of saccades. *Vision Research*, 74, 40-60.

Zhen, B., Weber, K., Mejia-Ramos, J.P. (2015). Mathematics majors' perceptions of the admissibility of graphical inferences in proofs. *International Journal of Research in Undergraduate Mathematics Education*, 2(1), 1-29.

Appendix A - Purported Theorems and Proofs

Figure A-1 Theorem: Practice - Both Studies

Theorem: If a diameter bisects a chord, then it is perpendicular to the chord.

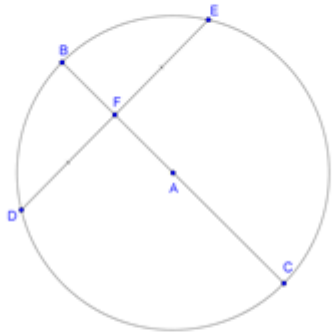


Figure A-2 Proof: Practice - Both Studies

Theorem: If a diameter bisects a chord, then it is perpendicular to the chord.

Proof:
Suppose DE is a chord on a circle centered at A.

Let BC be a diameter that bisects DE at a point F.

Construct the radii AD and AE.

Thus $\triangle DAE$ is isosceles.

$\triangle ADF$ and $\triangle AEF$ are congruent.

Thus $\angle AFD = \angle AFE$.

Note $\angle AFD + \angle AFE = 180°$.

Hence $\angle AFD = \angle AFE = 90°$.

Thus $BC \perp DE$.

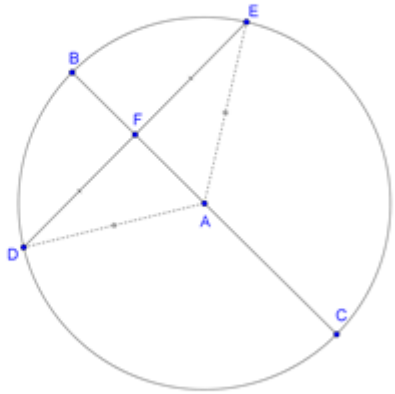


Figure A-3 Answer: Practice - Both Studies

Assert the validity of the proof

Theorem: If a diameter bisects a chord, then it is perpendicular to the chord.

Proof:
Suppose DE is a chord on a circle centered at A.

Let BC be a diameter that bisects DE at a point F.

Construct the radii AD and AE.

Thus $\triangle DAE$ is isosceles.

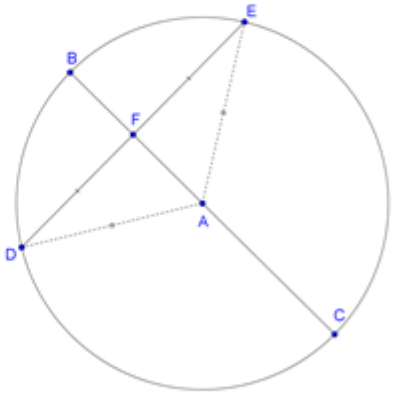
$\triangle ADF$ and $\triangle AEF$ are congruent.

Thus $\angle AFD = \angle AFE$.

Note $\angle AFD + \angle AFE = 180°$.

Hence $\angle AFD = \angle AFE = 90°$.

Thus $BC \perp DE$.



Valid press V Invalid press N
Please provide justification out loud

Figure A-4 Theorem: Angle Bisector (AngBi) - Pilot Study

Theorem: If, in $\triangle ABC$, AD is the internal angle bisector of $\angle A$, then $AB/AC = BD/CD$.

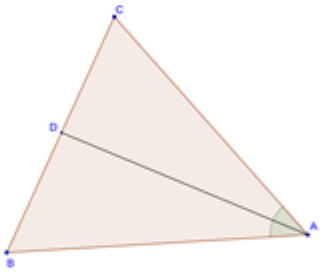


Figure A-5 Proof: Angle Bisector (AngBi) - Pilot Study

Theorem: If, in $\triangle ABC$, AD is the internal angle bisector of $\angle A$, then $AB/AC = BD/CD$.

Proof:

Extend the line segment AB.

Construct a line through point C that is parallel to segment AD.

Let E be the point of intersection.

As corresponding angles, $\angle CEA = \angle DAB$.

As alternate interior angles $\angle DAC = \angle ACE$.

Thus $\triangle ACE$ is isosceles and $AE = AC$.

Note that $\triangle BAD \sim \triangle BEC$.

$$BD/BA = BC/BE = \frac{BD + CD}{BA + AE}$$

Thus,

$$BD/CD = AB/AE = AB/AC$$

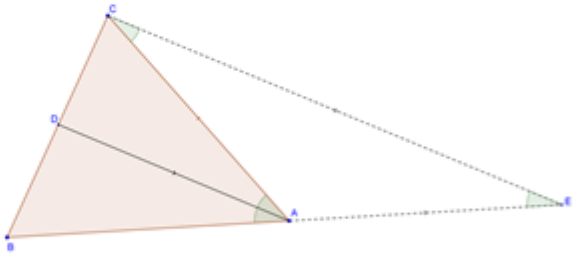


Figure A-6 Theorem: Angle Isosceles (AngIsos) - Pilot Study

Theorem: Given an isosceles triangle, the sides opposite the equal angles have equal lengths.

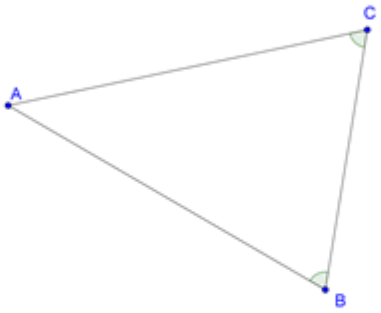


Figure A-7 Proof: Angle Isosceles (AngIsos) - Pilot Study

Theorem: Given an isosceles triangle, the sides opposite the equal angles have equal lengths.

Proof:
Construct AD, the angle bisector of $\angle BAC$ perpendicular to BC.

Note that $\triangle ADB \sim \triangle ADC$.

Since $\triangle ADB$ and $\triangle ADC$ share side AD,

the triangles are congruent.

Thus $AC = AB$.

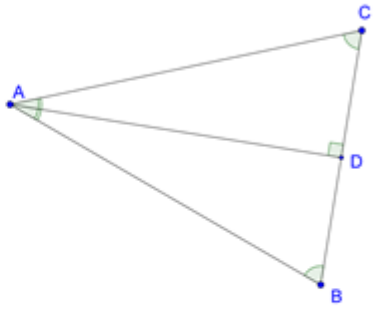


Figure A-8 Theorem: Bisector Isosceles (BiIsos) - Pilot Study

Theorem: If two internal angle bisectors are equal, then the triangle is isosceles.

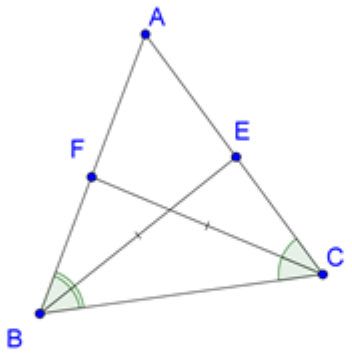


Figure A-9 Proof: Bisector Isosceles (BiIsos) - Pilot Study

Theorem: If two internal angle bisectors are equal, then the triangle is isosceles.

Proof:
Keep in mind that $BE = CF$.

Let us assume that $\angle B < \angle C$.

Hence $CE < BF$.

Construct a line through point B parallel to segment CE.

Construct a line though point C parallel to segment BE.

Let the intersection be the point H
and join points H and F.

Now $\triangle CFH$ is isosceles, so $\angle CFH = \angle CHF$.

$HB = CE < BF$ so $\angle BFH < \angle BHF$.

Now $\angle BHC = \angle BHF + \angle FHC$ and $\angle BFC = \angle BFH + \angle HFC$

Thus $\angle BHC - \angle BFC = \angle BHF - \angle BFH > 0$ meaning $\angle BHC > \angle BFC$.

Now $\angle BEC = \angle BHC$ so $\angle BEC > \angle BFC$.

Considering $\triangle BFK$ and $\triangle CEK$ we have $\angle BFC + \angle FBK = \angle KCE + \angle BEC$

Thus $\angle KCE - \angle FBK = \angle BFC - \angle BEC < 0$ meaning $\angle FBK > \angle KCE$.

This implies $\angle B > \angle C$. This contradicts our assumption.

Apply a similar argument for $\angle B > \angle C$. Thus $\angle B = \angle C$.

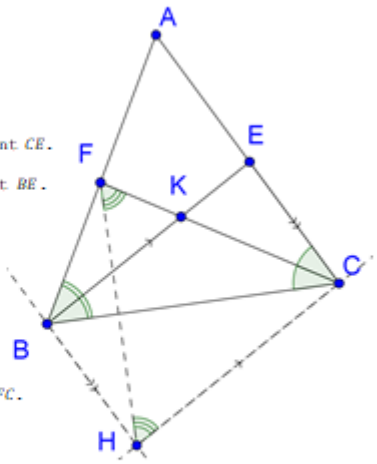


Figure A-10 Theorem: Inscribed Angles (Ins) - Pilot Study

Theorem: An inscribed angle is equal to half of the central angle. $\angle ABC = \frac{1}{2}\angle AOC$.

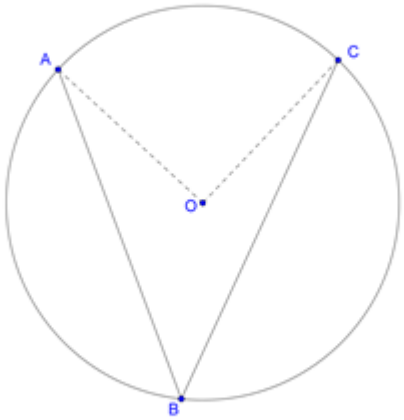


Figure A-11 Proof: Inscribed Angle (Ins) - Pilot Study

Theorem: An inscribed angle is equal to half of the central angle. $\angle ABC = \frac{1}{2}\angle AOC$.

Proof:
Case 1: Center on angle
Construct line segment OC.

$\triangle BOC$ is isosceles

Hence $\angle ABC = \angle OCB$.

$\angle AOC = 180° - \angle COB = 2\angle ABC$.

Case 2: Center inside angle
Construct the diameter BD and radii OA, OC.

Note that $\angle ABC = \angle ABD + \angle CBD$.

By case 1, it follows that
$\angle ABC = \frac{1}{2}\angle AOD + \frac{1}{2}\angle DOC = \frac{1}{2}(\angle AOD + \angle DOC)$.

Thus $\angle ABC = \frac{1}{2}\angle AOC$.

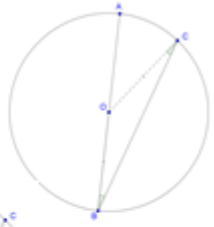
Case 3: Center outside angle
Construct the diameter BD and radii OA, OC.

Note that $\angle ABC = \angle ABD - \angle CBD$.

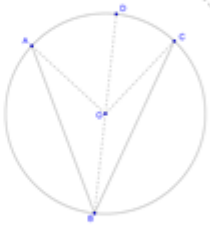
By case 1 it follows that
$\angle ABC = \frac{1}{2}\angle AOD - \frac{1}{2}\angle DOC = \frac{1}{2}(\angle AOD - \angle DOC)$.

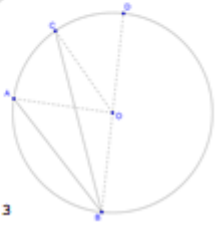
Thus $\angle ABC = \frac{1}{2}\angle AOC$.

Case 1



Case 2





Case 3

Figure A-12 Theorem: Miquel Point (Miq) - Pilot Study

Theorem: If an arbitrary point is taken on each side of a triangle, the three circles determined by each vertex and the two points on the adjacent sides have a common point of intersection within the triangle.

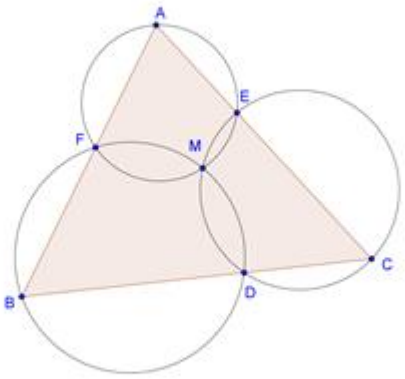


Figure A-13 Proof: Miquel Point (Miq) - Pilot Study

Theorem: If an arbitrary point is taken on each side of a triangle, the three circles determined by each vertex and the two points on the adjacent sides have a common point of intersection within the triangle.

Lemma: A quadrilateral is cyclic (can be inscribed in a circle) if and only if opposite angles are supplementary.

Proof:

Let the circles determined by points C, D, E and F, B, D intersect at M.

$\angle FMD = 180° - B$

And $\angle DME = 180° - C$

$\angle FME + \angle DME + \angle FMD = 360°$

Thus $\angle FME = B + C$.

Hence $\angle FME = 180° - A$.

$AFME$ is thus a cyclic quadrilateral

Meaning point M lies on the circle determined by the points A, F, E.

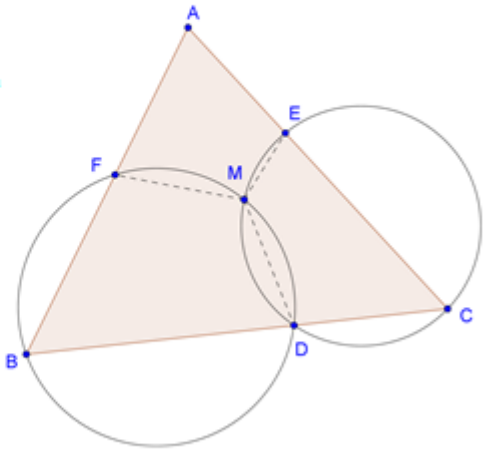


Figure A-14 Theorem: Ptolemy (Pto) - Pilot Study

Theorem: A quadrilateral is cyclic (can be inscribed in a circle) if and only if the product of the diagonals is equal to the sum of the products of the opposite sides: $ef = ac + bd$.

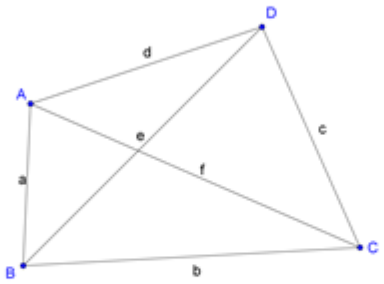


Figure A-15 Proof: Ptolemy (Pto) - Pilot Study

Theorem: A quadrilateral is cyclic (can be inscribed in a circle) if and only if the product of the diagonals is equal to the sum of the products of the opposite sides: $ef = ac + bd$.

Lemma: Inscribed angles subtended by the same arc are equal.

Proof:

Construct $\angle CDE = \angle ADB$.

Since $\angle ABD = \angle ECD$

$\triangle CDE \sim \triangle BDA$.

Thus $\frac{a}{e} = \frac{EC}{c}$.

Similarly, $\angle ADE = \angle BDC$

And $\angle CBD = \angle DAE$

So $\triangle BDC \sim \triangle ADE$.

Meaning $\frac{b}{e} = \frac{AE}{d}$.

Thus $bd + ac = e(AE + EC) = ef$.

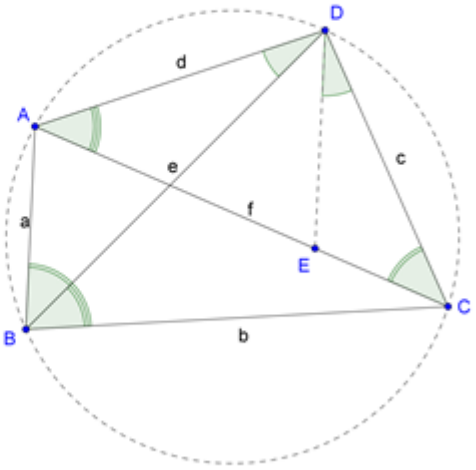


Figure A-16 Statement: Angle Bisector (AngBi) - Eye Tracking

Theorem: If, in $\triangle ABC$, AD is the internal angle bisector of $\angle A$, then $^{AB}/_{AC} = {}^{BD}/_{CD}$.

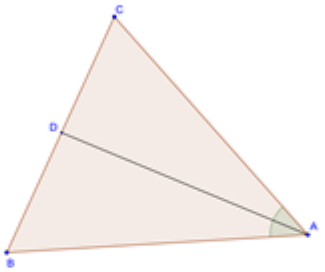


Figure A-17 Proof: Angle Bisector (AngBi) - Eye Tracking

Theorem: If, in $\triangle ABC$, AD is the internal angle bisector of $\angle A$, then $^{AB}/_{AC} = {}^{BD}/_{CD}$.

Lemma: If a line is parallel to one side of a triangle and intersects the other two sides, then it separates those sides into segments of proportional length.

Proof:

Extend the line segment AB.

Construct a line through point C that is parallel to segment AD.

Let E be the point of intersection.

$^{BD}/_{CD} = {}^{AB}/_{AE}$

As corresponding angles, $\angle CEA = \angle DAB$.

As alternate interior angles $\angle DAC = \angle ACE$.

Thus $\triangle ACE$ is isosceles and $AE = AC$.

$^{BD}/_{CD} = {}^{AB}/_{AC}$

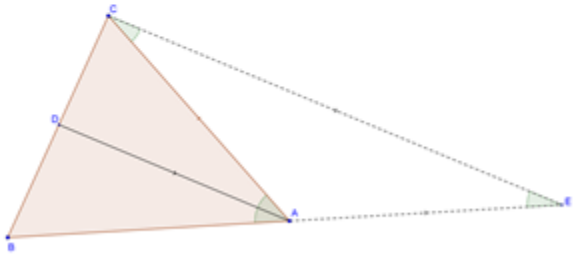


Figure A-18 Statement: Bisector Isosceles (BiIsos) - Eye Tracking

Theorem: If two internal angle bisectors are of equal length, then the triangle is isosceles.

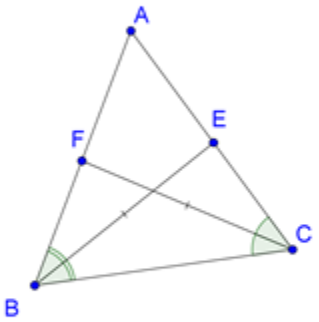


Figure A-19 Proof: Bisector Isosceles (BiIsos) - Eye Tracking

Theorem: If two internal angle bisectors are of equal length, then the triangle is isosceles.

Proof:

Suppose that $BE = CF$.

Let us assume that $\angle B < \angle C$.
Hence $CE < BF$, as longer lengths are opposite larger angles.

Construct a line through point B parallel to segment CE.
Construct a line though point C parallel to segment BE.
Let the intersection be the point H
and join points H and F.

Now $HC = BE = CF$, so $\angle CFH = \angle CHF$.
$HB = CE < BF$ so $\angle BFH < \angle BHF$.
Now $\angle BHC = \angle BHF + \angle CHF$ and $\angle BFC = \angle BFH + \angle CFH$
Thus $\angle BHC - \angle BFC = \angle BHF - \angle BFH > 0$ meaning $\angle BHC > \angle BFC$.
Now $\angle BEC = \angle BHC$ so $\angle BEC > \angle BFC$.

Considering $\triangle BFK$ and $\triangle CEK$ we have $\angle BFK + \angle FBK = \angle KCE + \angle KEC$
Thus $\angle KCE - \angle FBK = \angle BFK - \angle KEC = \angle BFC - \angle BEC < 0$ meaning $\angle FBK > \angle KCE$.
This implies $\angle B > \angle C$. This contradicts our assumption.
Apply a similar argument for $\angle B > \angle C$. Thus $\angle B = \angle C$.

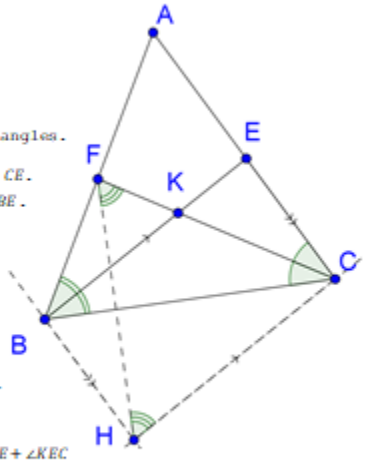


Figure A-20 Statement: Inscribed Angle (Ins) - Eye Tracking

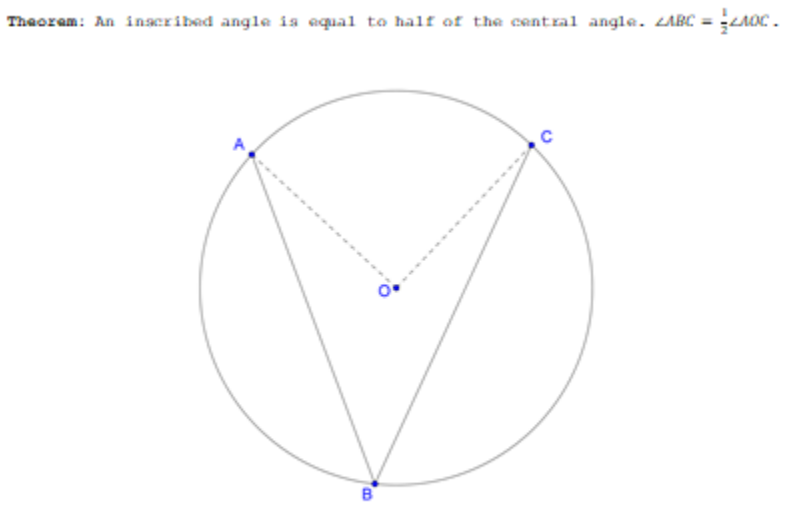


Figure A-21 Proof: Inscribed Angle (Ins) - Eye Tracking

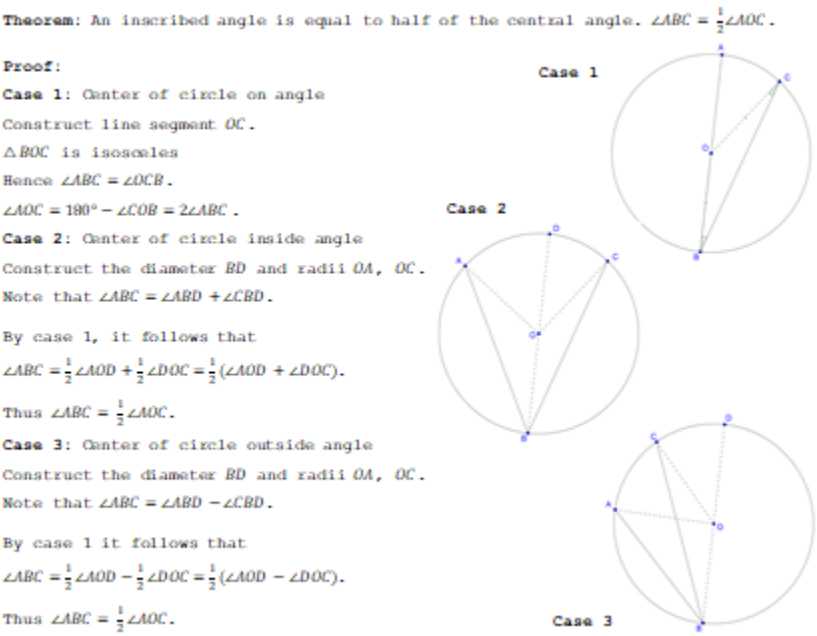


Figure A-22 Statement: Miquel Point (Miq) - Eye Tracking

Theorem: If an arbitrary point is taken on each side of a triangle, the three circles determined by each vertex and the two points on the adjacent sides have a common point of intersection within the triangle.

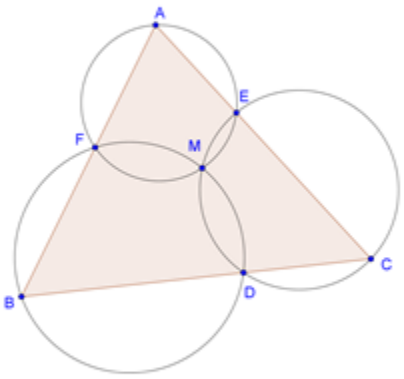


Figure A-23 Proof: Miquel Point (Miq) - Eye Tracking

Theorem: If an arbitrary point is taken on each side of a triangle, the three circles determined by each vertex and the two points on the adjacent sides have a common point of intersection within the triangle.

Lemma: A quadrilateral is cyclic (can be inscribed in a circle) if and only if opposite angles are supplementary.

Proof:

Let the circles determined by points C, D, E and F, B, D intersect at M.

Now $BFMD$ and $DMEC$ are convex and points E and F lie on the sides of $\triangle ABC$ thus M lies within the triangle.

$\angle FMD = 180° - B$

And $\angle DME = 180° - C$

$\angle FME + \angle DME + \angle FMD = 360°$

Thus $\angle FME = B + C$.

Hence $\angle FME = 180° - A$.

$AFME$ is thus a cyclic quadrilateral

Meaning point M lies on the circle determined by the points A, F, E.

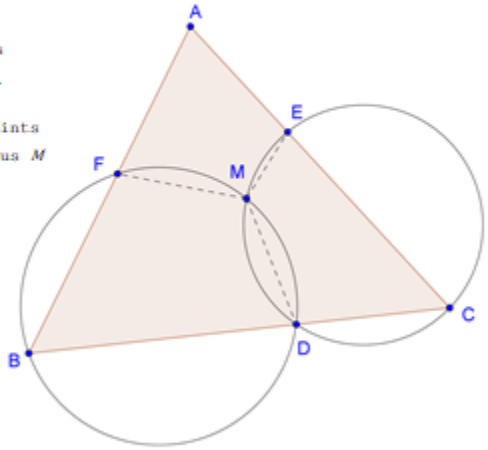


Figure A-24 Statement: Ptolemy (Pto) - Eye Tracking

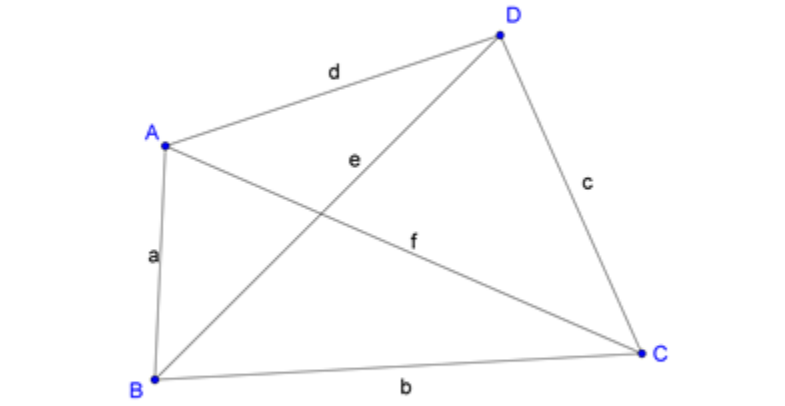
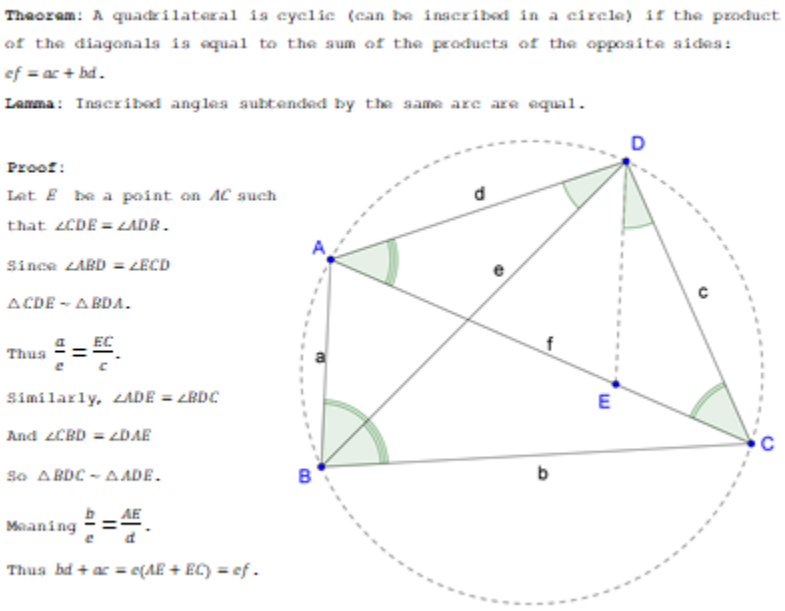


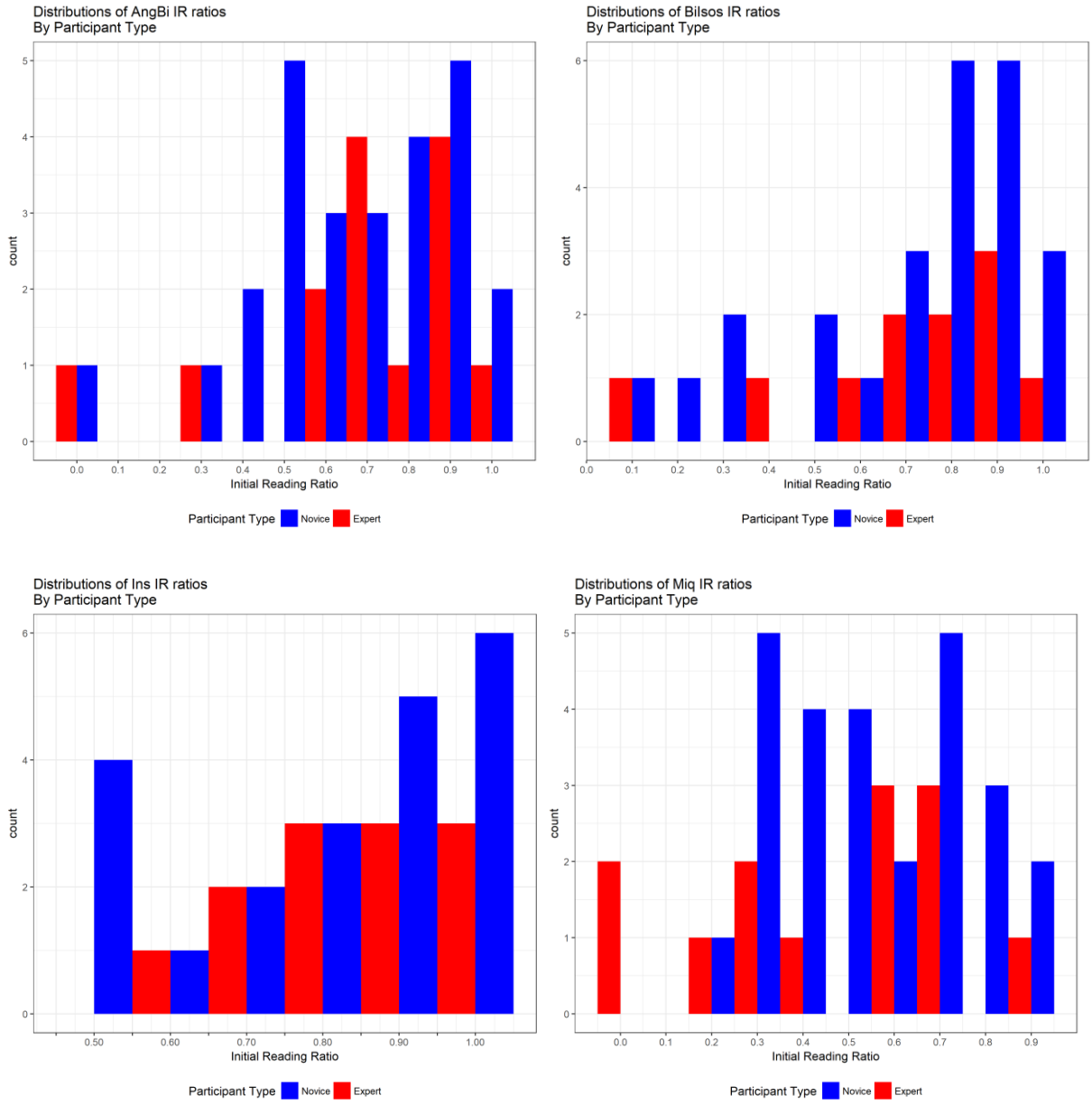
Figure A-25 Proof: Ptolemy (Pto) - Eye Tracking

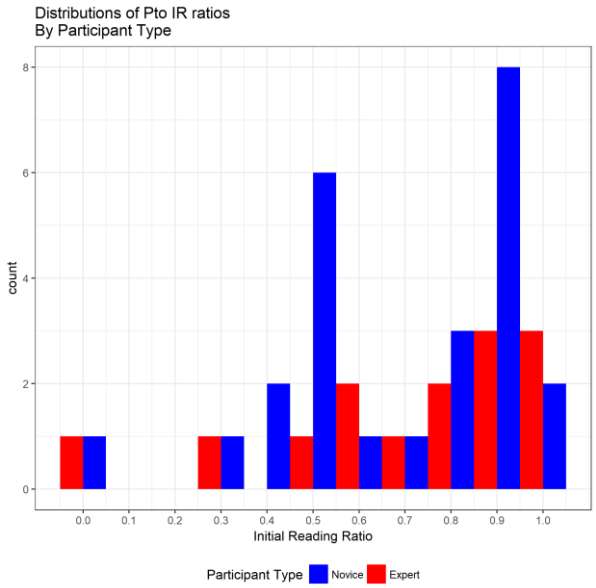


Appendix B - Supplemental Data

IR Ratio Distributions

Figure B-1 Histograms of IR ratio distributions by participant type





Continued Progression during Validations Asserting Invalid

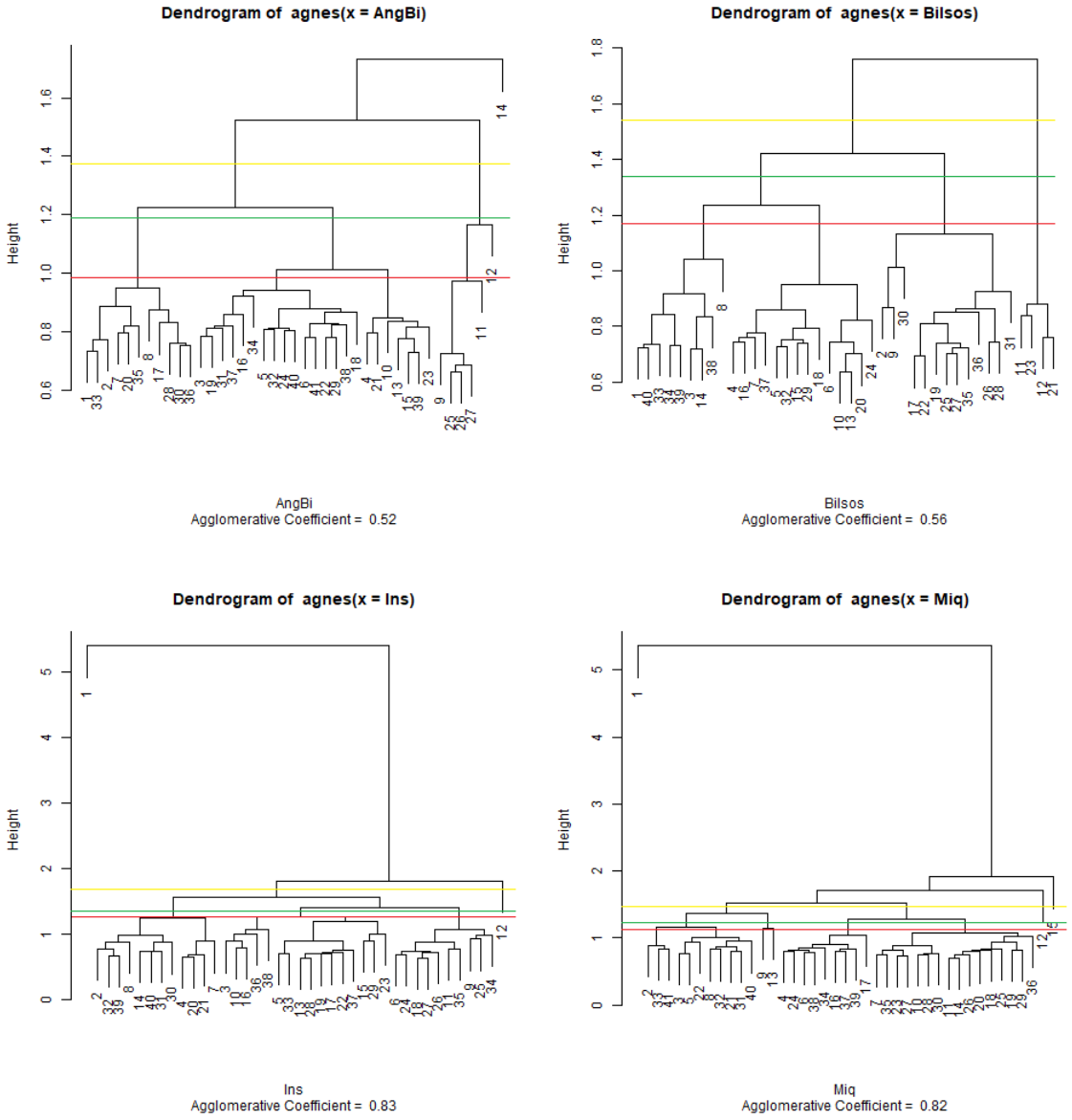
Table B-1 Tabulations of continued progression through and identification of error at the end of the proof

*Indicates the single occurrence of the termination of validation before the end of the proof in which the no specific error was identified in the justification. Participant type-novice.

Purported Theorem	Progression through entire proof		Progression terminates shortly after error		Specified Error at end of proof	
	Novice	Expert	Novice	Expert	Novice	Expert
AngBi	10	0	1	0	1	0
BiIsos	8	3	2	4	0	0
Ins*	5	0	2	0	1	0
Miq	6	6	0	1	2	2
Pto	4	4	5	1	0	0

Dendrograms

Figure B-2 Dendrograms of each purported theorem with reasonable cuts



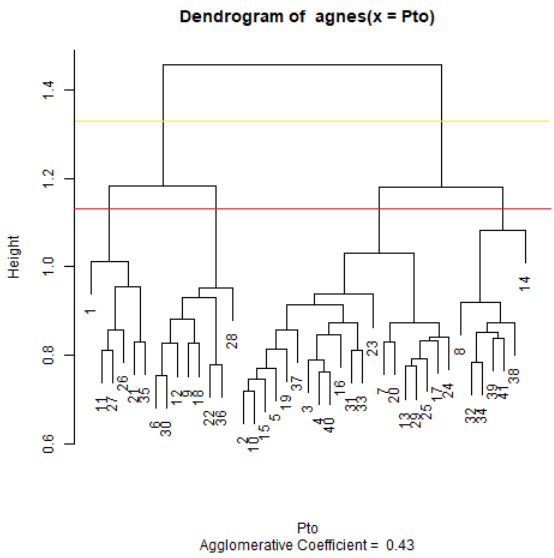
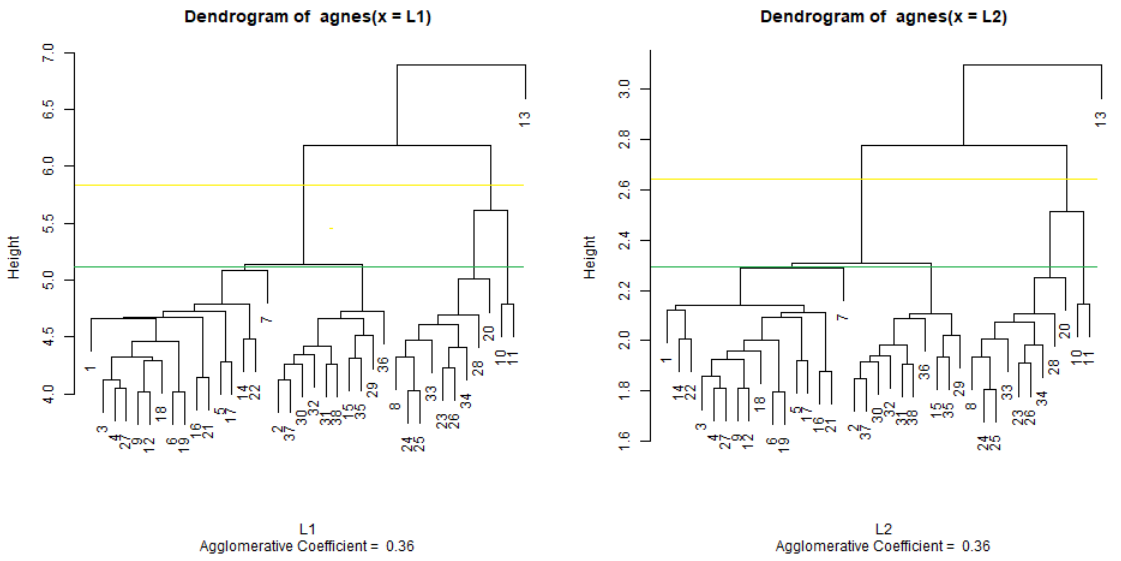


Table B-2 Identification table of outlier data regarding clustering by purported theorem

Numbers on dendrograms differ from participant number in certain cases due to bad tracking on several trials as discussed in Preparing the Data in Chapter 4.

Purported Theorem	Outlier participants	Number on Dendrogram
AngBi	GeoP14	14
BiIsos	n/a	n/a
Ins	GeoP2, Geop13	1, 12
Miq	GeoP13	13
Pto	n/a	n/a

Figure B-3 Dendrograms of normed ScanMatch dissimilarity vectors with reasonable cuts



Warranted Line Classifications

Table B-3 Table presenting the classification of the proof components

Classifications: C=construction/description, I=implicit warrant, E=explicit warrant, and B=both implicit and explicit warrants

Component	Class	Justification
AngBi 1	C	Construction
AngBi 2	C	Construction
AngBi 3	I	Does not justify existence of intersection
AngBi 4	I	Does not cite use of lemma
AngBi 5	E	Provides justification via assertion as corresponding angles
AngBi 6	E	Provides justification via assertion as alternate angles
AngBi 7	E	Thus indicates it follows from the previous statements and provides justification for final statement in line
AngBi 8	I	Cites neither use of lemma nor AngBi 4
BiIsos 1	C	Supposition
BiIsos 2	C	Additional supposition forming basis of proof structure
BiIsos 3	E	Explicit justification provided for claim in line
BiIsos 4	C	Construction
BiIsos 5	C	Construction
BiIsos 6	I	Does not justify existence of intersection
BiIsos 7	C	Construction
BiIsos 8	B	Unjustified claim justifies the following claim within the line

Component	Class	Justification
BiIsos 9	B	Unjustified claim justifies the following claim within the line
BiIsos 10	I	Makes two claims without justification
BiIsos 11	E	Thus indicates it follows from previous statement and provides justification for final statement in line
BiIsos 12	B	Unjustified claim justifies the following claim within the line
BiIsos 13	E	Cites two triangles justifying the resulting claim
BiIsos 14	E	Thus indicates it follows from previous statement and provides justification for final statement in line
BiIsos 15	E	This implies indicates it follows from previous statement and forms a contradiction
BiIsos 16	E	Explicitly lays out the symmetry of argument to form conclusion
Ins 1	C	Case Description
Ins 2	C	Construction
Ins 3	I	Unjustified claim
Ins 4	E	Hence indicates it follows from previous statement
Ins 5	I	String of claims without justification
Ins 6	C	Case Description
Ins 7	C	Construction
Ins 8	I	Unjustified claim
Ins 9	E	Claim cites usage of the first case
Ins 10	E	Thus indicates it follows from previous statement
Ins 11	C	Case Description
Ins 12	C	Construction
Ins 13	I	Unjustified claim
Ins 14	E	Claim cites usage of the first case
Ins 15	E	Thus indicates it follows from previous statement
Miq 1	I	Does not justify existence of additional intersection
Miq 2	E	Provides explicit argument for why M resides within the triangle
Miq 3	I	Uncited use of lemma
Miq 4	I	Uncited use of lemma
Miq 5	I	Unjustified claim
Miq 6	E	Thus indicates it follows from previous statements
Miq 7	E	Hence indicates it follows from previous statements
Miq 8	E	Thus indicates use of previous statements to provide explicit justification for conclusion
Pto 1	C	Construction/ duplication of angle
Pto 2	I	Uncited use of lemma
Pto 3	E	Since from previous line provides explicit justification for the following line
Pto 4	E	Thus indicates it follows from previous line
Pto 5	I	Unjustified claim

Component	Class	Justification
Pto 6	I	Unjustified claim
Pto 7	E	So indicates it follows from previous statements
Pto 8	E	Meaning indicates it follows from previous line
Pto 9	E	Thus indicates it follows from previous line

Appendix C - Informed Consent, Protocol, and Debriefing Forms

Pilot Study

Pilot Study Protocol

Location: Q-center
Timeframe: July 2016

Format: This study is to be a think-aloud study. To create an environment similar to the actual study, the participants will be in a quiet room (the Q-center) and will read from a computer screen.

Procedure:

I will set the room up 10 minutes before the scheduled experiment time. This entails

- Making sure the room is at a comfortable temperature
- Turning on the computer and initiating the experiment.
- Prepare for recording.
- Having the proper informed consent forms
- Having proper payment materials (money and form)

I will meet the participant outside in the hall way.

Complete the Informed Consent form (two copies-one for us and one for them)

Have them explain it back to you.

Ensure the following:

- Comfortable with validation
  - Proof validation is when one reads a mathematical proof with the aim of determining whether it is valid or not. You may think of it as deciding whether the proof establishes the claim of the theorem or not.
- Know the format
  - They may take as long as the need to make an informed decision.
- Know to think aloud

Start both recordings- audio and video

After the completion of the last proof, I will ask the following questions.

In general,

-Would you say the way you assessed validity varied from theorem to theorem?

-We had a statement page in addition to the proof page for each theorem, was this useful?

-Did this process differ from previous validation experience?

How would not being able to point affect this process?

For each theorem,

-How did you assess this proof's validity, what was the process?

-Did you find the diagrams helpful? If so, how did you utilize them?

-Did the text or diagram occupy most of your attention or was it pretty evenly split?

Upon completion, I will ask if they have any questions.

-This is a pilot study. I will be conducting further study in the fall. It can seriously mess up my research if these theorems and proofs are discussed with future participants. Please refrain from discussing this with other people related to math at KSU.

Stop recordings.

Fill out payment form.

Pay the participant.

Walk them outside

Kansas State University

Informed Consent Form

Project Title: Proof Validation in Euclidean geometry Using Eye Tracking-Pilot Study

Approval Date of Project: 01/01/2014
Expiration Date of Project: 12/31/2016

Principal Investigator: Dr. Andrew Bennett
Co-Investigator(s): Dr. Lester Loschky, Paul Flesher, John Hutson, Zachary Throneburg

Contact for Questions/Problems: 138 Cardwell Hall, KSU Math Dept
Phone: (785) 532-6750

IRB Chair Contact: Rick Scheidt, Chair on Research Involving Human Subjects, 203 Fairchild Hall, KSU, Manhattan, KS, 66506 (785)-532-3224

Cheryl Doerr, Associate Vice President for Research Compliance, 203 Fairchild Hall, KSU, Manhattan, KS, 66506 (785)-532-3224

Sponsor of Project: National Science Foundation

Purpose of Research: To study how different people validate mathematical proofs.

Procedures:
1) You will be asked to think aloud as you validate the proofs provided for a series of theorems and provide justification for your conclusions.
2) Each theorem has a statement page followed by the proof page, which has the theorem restated. There is then an answer page. You may assume the provided lemmas are correct.
3) At the conclusion, you will be asked several questions regarding the study.

4) The study will be videotaped-focused solely on the computer screen. Audio will be recorded.
5) Upon completion of the study, you will be paid $15 (SSN required).

Length of Study: 45 to 75 minutes, averaging 60 minutes
Risks Anticipated: No known risks are anticipated

Extent of Confidentiality: Participation is anonymous. A participant number will be used. Recordings will be kept securely. Transcriptions will be used to prevent voice recognition.

Is Compensation or Medical Treatment Available if Injury Occurs: Not applicable. No known risks.

Parental Approval for Minors: Only subjects 18 or older are included in the study.

Terms of Participation: I understand this project is research, and that my participation is completely voluntary. I also understand that if I decide to participate in this study, I may withdraw my consent at any time, and stop participating at any time without explanation, penalty, or loss of benefits, or academic standing to which I may otherwise be entitled.

I verify that my signature below indicates that I have read and understand this consent form, and willingly agree to participate in this study under the terms described, and that my signature acknowledges that I have received a signed and dated copy of this consent form.

Participant Name: _____

Participant Signature: _____ **Date:** _____

Witness to Signature: _____ **Date:** _____

Eye Tracking Study

Eye Tracking Study Protocol

Location: Bluemont
Timeframe: Fall 2016

Procedure:

I will set the room up 10 minutes before the scheduled experiment time.

- Turn on the computers (eye tracking(2) and vision)
- Check vision test chair placement
- Initiate the experiment.
- Prepare the audio recorder.
- Put proper informed consent forms(2) on clip board.
- Get out proper payment materials (money and form)
- Prepare the log book

I will meet the participant outside in the hall way.

Have them sit on the vision test chair.

Complete the Informed Consent form (two copies-one for us and one for them)

Have them explain it back to you.

Ensure the following:

- Comfortable with validation
  - Proof validation is when one reads a mathematical proof with the aim of determining whether it is valid or not. You may think of it as deciding whether the proof establishes the claims of the theorem or not.
- Know the format
  - They may take as long as the need to make an informed decision.

Do the vision test- FRACT

Fill out the participant information in the log book

Adjust the chair and chin rest with the participant.

Explain the importance of keeping the chin and forehead the same distance from the screen throughout session and study.

Begin the calibration(C)/validation(V) process:

Press ENTER to get the image on the display monitor

Use ARROW keys to switch image

Use mouse click thresholds

Center head (single eye) in camera view big room (little room)

- Minimize corneal glare with focus on the camera (light blue)
- Thresholds maximize the dark blue in pupil (not outside)
- Minimize the white orb around corneal glare (not outside)
- Ask them to look certain places (top L/R, bottom L/R) and keep track of eye movement and the bleeding of colors
- Start calibration
    - "Look at the center of the white circles, maintain eye contact with circles, when it moves follow it, but do not try to predict where it is going. Please keep your head stationary during this process."
    - Press space bar to begin calibration
- Want a square of green crosses, once good validate
    - If tilted left, rotate camera to the right
- Validation
    - Press space bar to begin the validation
    - Maximum less than 1
    - Average less than ½

Talk about drift check- Look at the center of the dot and then press space bar

Start audio recording

Press O to begin experiment; control + C aborts experiment

After they finish the practice problem, ask if they have any questions.

Upon completion debrief the participant by thanking them and asking them if they have any questions, and giving them the debriefing sheet. Tell them they can take the sheet if they would like to keep it, or leave it if they don't want it. They may email you or talk with you if they have any questions later.

Stop recordings.

Fill out payment form and pay participant

Kansas State University

Informed Consent Form

Project Title: Proof Validation in Euclidean geometry Using Eye Tracking

Approval Date of Project: 01/01/2014
Expiration Date of Project: 12/31/2016

Principal Investigator: Dr. Andrew Bennett
Co-Investigator(s): Dr. Lester Loschky, Paul Flesher, John Hutson, Zachary Throneburg

Contact for Questions/Problems: 138 Cardwell Hall, KSU Math Dept
Phone: (785) 532-6750

IRB Chair Contact: Rick Scheidt, Chair on Research Involving Human Subjects, 203 Fairchild Hall, KSU, Manhattan, KS, 66506 (785)-532-3224

Cheryl Doerr, Associate Vice President for Research Compliance, 203 Fairchild Hall, KSU, Manhattan, KS, 66506 (785)-532-3224

Sponsor of Project: National Science Foundation

Purpose of Research: To study how different people validate mathematical proofs.

Procedures:

1) We will begin with a vision test. You must see at or better than 20/30.
2) We will then calibrate the eye tracking equipment. If the accuracy requirements cannot be met, you will be dismissed from the study.
3) You are asked to validate the proofs (does the argument establish the claims of the statement?) that we have provided for a series of theorems. Once you have reached each conclusion you are to verbally justify.

4) For each theorem you will be presented 3 pages in the following order: Statement, Proof, and Answer. The theorem is repeated on each page.
5) You may assume the provided lemmas are correct.
6) At the conclusion, you will be asked several questions regarding the study.
7) Audio will be recorded
8) Upon completion of the study, you will be paid $15 (SSN required).

Length of Study: 45 to 60 minutes
Risks Anticipated: No known risks are anticipated

Extent of Confidentiality: A participant number will be used. Recordings will be kept securely. Transcriptions will be used.

Is Compensation or Medical Treatment Available if Injury Occurs: Not applicable. No known risks.

Parental Approval for Minors: Only subjects 18 or older are included in the study.

Terms of Participation: I understand this project is research, and that my participation is completely voluntary. I also understand that if I decide to participate in this study, I may withdraw my consent at any time, and stop participating at any time without explanation, penalty, or loss of benefits, or academic standing to which I may otherwise be entitled.

I verify that my signature below indicates that I have read and understand this consent form, and willingly agree to participate in this study under the terms described, and that my signature acknowledges that I have received a signed and dated copy of this consent form.

Participant Name: _____

Participant Signature: _____ **Date:** _____

Witness to Signature: _____ **Date:** _____

Proof Validation Study Debriefing:

Thank you for participating in the eye tracking study!

The purpose of this study is twofold.

- To gain insight into proof validation techniques and the differences between experts and novices
- To gain insight into diagram usage.

Understanding these important aspects of mathematics will enable further development of teaching strategies and help lay the ground work for online learning environments geared towards mathematical proof.

If you would like to receive a copy of the final report from this study, I would be happy to send you a link via email. The project should be finished up in 2018. Please let me know if interested.

If you are interested in the previous work in this area, I recommend beginning with the following papers:

Matthew Inglis and Lara Alcock, Expert and Novice Approaches to Reading Mathematical Proofs (2012)

Keith Weber and Juan Pablo Mejia-Ramos, On Mathematicians' Proof Skimming: A Reply to Inglis and Alcock (2013)

We ask that you refrain from discussing the content of the study with other possible participants.

Thanks again!

If you have any questions or concerns, please contact me, Paul Flesher, at pmflesher@ksu.edu or drop by my office CW 126.

Further contact information:

Dr. Andrew Bennett, Major Professor, bennett@ksu.edu

Dr. Lester Loschky, Head of Visual Cognition Lab, loschky@ksu.edu