ASSESSING ITP STUDENTS' ARGUMENT VALIDATING ABILITY:

FRAMING, DEVELOPING AND VALIDATING

A PILOT ASSESSMENT

by

Joshua B. Fagan

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Mathematics Education
August 2019

Committee Members:

Alex White, Chair

Kate Melhuish, Chair

Sharon Strickland

Keith Weber

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

This work is dedicated to my family – Katelyn, Lisa, Alison, Michael, Emily and Rebecca – who has faithfully seen me through the last five years. Their love and support have made it all worthwhile. May we never lose sight of what is most important in life and ever be willing to give our all to love and support one another.

AKNOWLEDGEMENTS

First and foremost, I am thankful for the hand of God in my life. He has blessed and inspired me throughout this undertaking. Not enough can be said of His grace and mercy in my life. I am forever grateful for His love for me. I would also like to thank my wife, Katelyn, for not only loving me and being patient with me through this ordeal, but for reading and editing this entire work from beginning to end. I know it was not fun to read, but dare we say it is finally over? Finally, I would like to thank the faculty and staff at Texas State University for their dedication and professionalism.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

In this paper I discuss the process of creating a closed-form multiple-choice assessment of students' ability to validate mathematical proofs at the introduction to proof (ITP) level. This process involved: (1) creating and validating a cohesive framework of common validity issues (CVI) in proof writing as a basis for assessment creation through a mathematician survey ($N = 228$) and two focus groups ($N = 4$ & $N = 7$); (2) creating and piloting an open version of the assessment as a means to create distractors for the closed assessment; (3) creating, piloting ($N = 187$) and analyzing the results from the closed form assessment; and (4) conducting interviews with student participants after the pilot to determine the characteristics of the process that students took during the pilot. The results of the processes offer an assessment that, with some refinement, can authentically measure students' ability to validate mathematical arguments from a number of perspectives in the ITP setting.

I. Introduction

Argumentation and proof are indispensable practices for all aspects of mathematics (de Villiers, 1990; Hanna, 2000; Rav, 1999; Stylianides, Stylianides & Weber, in press). For most of the twentieth century research which focused on mathematical argumentation and proof focused on how students construct arguments. This trend shifted in the first decade of the twenty-first century as researchers renewed awareness for other skills related to the argumentative process, most especially that of comprehending and validating proofs, skill which in the past were largely neglected in empirical studies (Alcock & Weber, 2005; Mejía-Ramos & Inglis, 2009; Selden & Selden, 2003). Proof validation is a significant facet of the mathematical practice for both students and mathematicians and as such is deserving of empirical consideration.

The skill of validating proofs is an aspect related to the reading of proof (Mejía-Ramos & Inglis, 2009) and as such is a cognitive activity requiring the reader to focus not only on the logical aspects of the proof, but also the use of statements and representation, and the overall structure of the proof (Ko & Knuth, 2013; Selden and Selden, 2003; Mejía-Ramos, Fuller, Weber, Rhoads, & Samkoff, 2012; Weber & Mejía-Ramos, 2011). Much of the research on the reading of mathematical proofs focuses on the skill of proof validation (Inglis & Alcock, 2012). These research tracts focus on such varying topics as the skills students need to validate proofs (Alcock & Weber, 2005; Weber & Alcock, 2005); the strategies mathematicians and students use in a variety of mathematical settings to validate proofs, and how they actually go about said evaluations (Inglis & Alcock, 2012; Ko & Knuth, 2013; Moore, 2016; Morris, 2007; Weber, 2008; Weber & Mejía-Ramos, 2011); and how conviction is affected by a teacher's or student's beliefs

about the validity of purported proofs (Knuth, 2002; Segal, 1999; Weber, 2010; Weber, Inglis, & Mejía-Ramos; 2014).

Two of the most common refrains from research on proof validation are that: (1) proof validation is an important skill for students to gain, however making these evaluations presents a considerable challenge for students (e.g., Inglis & Alcock, 2012; Selden & Selden, 2003; Weber, 2010); and (2) mathematicians often do not agree on the validity of purported proofs (e.g., Inglis, Mejía-Ramos, Weber,& Alcock, 2013; Moore, 2016). The existing research highlights the situation facing mathematics education regarding proof validation is at odds with itself: students need to learn how to validate proof but struggle, yet, teachers and mathematicians are inconsistent in their own practice of validation due to contextual constrains which exist with regards to proof in mathematics generally. More understanding is needed in order to remedy the variance in the practices of teaching and learning of proof validation.

Despite the research efforts surrounding proof validation, there does not exist a comparable and perhaps parallel field of proof comprehension: no formal assessment exists on any level, nor is there an understanding of the domains for which one might even begin to assess such a construct. The three ways in which the assessment for proof comprehension are important to mathematics education (Mejía-Ramos, Lew, de la Torre, & Weber, 2018) could similarly be true for proof validation if such an assessment existed. In mirrored fashion, an assessment of proof validation might be able to: (1) offer a tool for teachers to better understand what their students know; (2) offer students a map to better focus their attention to important aspects of proof, leading to deeper understanding; and (3) offer researchers a tool to better understand aspects of proof

validation and uncover new related competencies. A proof validation assessment (PVA) offers a tool not only to improve classroom outcomes, but also to further researcher into this often difficult-to-grasp skill.

Significance

Because proof validation seems to be connected to other areas of proof, like proof comprehension and proof construction (see Mejía-Ramos, Fuller, Weber, Rhoads, & Samkoff, 2012; Selden & Selden, 1995, 2003), it is important to have a tool to measure students' ability to validate proof to further research efforts in this field. If proof validation is in fact strongly connected to these other aspects dealing with proof, then an assessment which measure students' ability to validate proofs would be a useful tool in first confirming there is in fact some sort of relationship between these other proof constructs, but also in determine the strength of these relationships. As Mejía-Ramos et al. (2018) suggest concerning their proof comprehension assessments, assessing proof validation "could be important for evaluating the effectiveness of mathematics instruction" (p. 4). Researchers suggest a link between validating and constructing proofs (Powers, Craviotto & Grassl, 2010; Selden & Selden, 2003) and similar claims exist about the connection between validation and proof comprehension (Alcock, Bailey, Inglis, & Docherty, 2014), however, these claims need stronger empirically evidence to fortify and generalize the theory. Furthermore, focusing on the Introduction to Proof (ITP) level is a good first step in the process of understanding the effects proof validation has on learning proof generally and gives a good starting point for helping map our understanding about proof and proof validity at the university level.

A tool, like the one proposed for this study, could help identify teaching and instruction habits, as well as proof learning and reading habits. This research tool can be expanded upon to focus on other areas of advanced mathematics beyond the ITP setting in order to help further clarify these questions throughout the undergraduate and even graduate mathematics tracts too.

Because creating a research tool necessitates the involvement of mathematicians to identify the essential aspects of proof with regards to proof validity, a strongly supported framework for understanding aspects of proof validity emerges. This framework, situated in this instance in the ITP context, could be scaled to help researchers better identify aspects of proof which are important markers in curricular materials, classroom presentations, and graded feedback, all of which may further support the development of proof validating competencies in the classroom environment. This knowledge could build new understandings on how to teach intuitively and overtly proof-base mathematics.

The norm concerning mathematical instruction and assessment at the advanced undergraduate level is often describe in terms of *definition-theorem-proof* presentations which are broken up over the semester by *state-and-prove* assessments (see Conradie & Frith, 2000; Davis & Hersh, 1981; Dreyfus, 1991; Weber, 2004) . The definition-theorem-proof/state-and-prove model of teaching and assessing has considerable drawbacks, the first of which is that only part of the curriculum is ever explicitly addressed in class.

It is widely agreed that in advanced undergraduate mathematics courses, one of the biggest overt focuses is on improving students' proof writing ability (e.g., Weber,

2001). Despite this emphasis, one of the most common activities students take part in, especially in the classroom, is the *reading* of proofs. This is due to the fact that the definition-theorem-proof paradigm is the most common pedagogical approach for the presentation of formal proofs (see Davis & Hersh, 1981; Weber, 2004) and it is through the presentation of written proofs professors most often afford instructional explanations (Lia, Weber & Mejía-Ramos, 2012). Because proofs as a pedagogical tool are meant to convince, explain, stimulate understanding, and enable mathematical discourse and critique, (de Villiers, 1990; Hanna; 1990; Hersh, 1993; Knuth, 2002; Mejía-Ramos et al., in press; Weber, 2010) it seems the curriculum in advanced undergraduate mathematics consists of writing proofs and also aspects of proof reading. One important aspect of proof reading is proof validating (Mejía-Ramos & Inglis, 2009) which is linked to the ability to construct proofs (Powers, Craviotto & Grassl, 2010; Selden & Selden, 2003) and is a major factor in promoting critique and mathematical discourse.

A second shortcoming of the pedagogical model of definition-theorem-proof/state-and-prove is in how it leads to a suboptimal learning experience as it assesses students' abilities to recall and prove theorems. Conradie and Frith (2000) outline some of the drawbacks of the state-and-prove assessment strategy, all of which can be inferred to affect learning in one form or another:

1. In preparing for these tests, students typically learn to memorize proof structures and tricks, if not entire proofs outright.
2. In grading, the professor learns what their students memorized or failed to memorize, not necessarily what students know, understand, or do in terms of proof and conceptual understanding.

3. The professor's feedback – or lack thereof – afford students little opportunity for further learning.

From a learning standpoint, this is a grim situation as students are ultimately judged more on their ability to memorize, not understand or apply. Even if a professor has the best intentions, these assessments often fail to test what students actually know and can do, so a professor has difficulty making accommodations throughout the course based upon students' performance. Also,, if students hope to gain further insight into how their knowledge and ability align with expectations, they might be left guessing as grading these assessments is contextually based upon the preference of the professor grading them (Moore, 2016) making the grades themselves somewhat arbitrary from class to class.

The final shortcoming of the definition-theorem-proof/state-and-prove cycle stems from the fact that proof reading – and of particular interest in this research – proof validating is part of the curriculum but not directly assessed. In constructing proofs as part of an assessment, some amount of validating passively occurs, as Selden and Selden (2003) suggest that:

> One constructs a proof with an eye toward ultimately validating it and may often validate parts of it during the construction process. In fact, the final portion of a proof construction is likely to be validation of that proof. That is, each process, validation and proof construction, entails the other. (p. 6)

Thus, as part of an assessment involving proof construction, students also validate, even if it is to assess if what they have memorized is presented – in what the student perceives – as a sound argument. Ultimately, if validating is not blatantly part of the curriculum,

then validating occurs – even in the setting of an assessment – much the way it is learned, implicitly and passively without specific regard to how it relates to the curriculum or mathematics in general.

As validating is not a deliberate part of the curriculum currently, there is little hope students become proficient in this practice. Because of this, feedback becomes the only means for a student in a definition-theorem-proof/state-and-prove class to learn how their conception of proof stands up against the standards of mathematical practice at large. But as was pointed out, feedback in this regard may be insufficient for the needs of the student (Conradie & Frith, 2000), and thus their conceptions of proof are shaped without specific knowledge about where they are deficient in their ability and understanding.

Research Goal

Mathematics education in the twenty-first century has seen considerable growth in terms of researcher's ability to assess individuals for conceptual and pedagogical knowledge about mathematical topics, skills, and understandings (e.g., *GTCA* – Melhuish, 2015; *LMT* – Hill, Ball & Schilling, 2008; *MQI* – Learning Mathematics for Teaching Project, 2011; *PCA* – Carlson, Oehrtman, Engelke, 2010; Proof Comprehension Assessment – Mejía-Ramos et al., 2018). This work needs to continue, especially at the tertiary level to support efforts to understand learning outcomes, and reform practice to increase opportunities to learn. To this end, having a validated instrument measuring proof validating competencies could open new venues for large scale research at the advanced undergraduate level. As such, the goal of this research is to build an assessment for the advanced undergraduate level, focused on proof validation that is robust enough to

serve as a research tool to further our understanding of students' proof competencies. Put plainly, the primary goal is to:

> Develop a cohesive analytic framework and use it to construct a closed-form, multiple-choice assessment which measures students' abilities to validate deductive mathematical arguments at the introduction to proof (ITP) level.

II. Proof Validation Literature Review

The mathematics education literature on proof covers a wide range of topics from the philosophical treatise on the purpose and character of proof as well as reasons for proving (e.g., de Villiers, 1990; Rav, 1999) to the more pragmatic inquiry focused on understanding the conceptions teachers and students have about proof (e.g., Harel & Sowder, 1998; Healy & Hoyles, 2000; Knuth, 2002; Morris, 2007). Additionally, researchers studied pedagogical issues and the role of the teacher in proof related activities and settings (e.g., Herbst, 2002; Lai, Weber, & Mejía-Ramos, 2014; Lew, Fukawa-Connelly, Mejía-Ramos, & Weber, 2016; Rowland, 2002).

Most important to this study, researchers focused on the types of mathematical *activities* that are associated with proof in an educational setting. In exploring proof-related activities, research on proof and argumentation can be broken down into three general categories of student competencies: (1) proof construction; (2) proof reading; and (3) proof presentation (Mejía-Ramos & Inglis, 2009). Most research focused on proof construction, while considerably less focused on the latter two tracts, especially proof presentation. The literature explored here naturally focuses on proof validation, an aspect of proof reading, giving special attention to what has been explored related to mathematician's perspectives on validity and student's capabilities in judging mathematical arguments.

Research on proof validation breaks down into subgenres based upon the population which each study was focusing on. These populations form three distinct bands: (1) K-12 students, K-5 teachers and preservice teacher (PST), (2) undergraduate mathematics majors including secondary PST; and (3) practicing mathematicians. I set

these groups in ascending order of mathematical background; the first group with the least amount of mathematical background, the second with greater indoctrination into proof culture, and of course the practicing mathematicians who regularly interact with and set the norms for the argumentative culture with the greatest depth of mathematical knowledge.

K-12 Students and Preservice Elementary Teachers

This set of research explores ideas like preservice elementary teachers (PSeT) conceptions of proof (Martin & Harel, 1989), school-aged students' conceptions of validity (Healy & Hoyles, 2000), PSeT process of evaluation of students' mathematical arguments (Morris, 2007), and framing of students' understanding concerning the structure of deductive proofs (Miyazaki, Fujita, & Jones, 2017). Taken as a whole, this set of research demonstrates that K-12 students and PSeT are deficient in their ability to consistently identify valid proofs. These results are unsurprising in some ways though, as this group represents a cluster of individuals who have had minimal to no indoctrination into advanced mathematics, especially those which deal with the norms of proof. The research suggests that PSeT seem to have difficulty in identifying valid deductive arguments, and in symmetric fashion, students too have difficulties in this regard.

The importance of this research for this study is that it fixes the understanding that once students complete their tract of K-12 mathematics education they are unindoctrinated in the art of proof reading and especially validating. This means that students who are in an ITP course are truly novice from the sense that they have not been formally introduced to proofs in any constructive way during either their K-12 education

or their courses leading up to their ITP classes[1]. Thus, when setting the aim for this study, to build an instrument which measures students' ability to validate arguments, the ITP classroom becomes a sufficient baseline for exploration. These students are novice in this regard and will serve to set the standard for what is or might be learned in the university setting.

Proof Validating

The study of proof validation was largely formalized by Selden and Selden in their 1995 study in which they focused on 61 students' ability to unpack the meaning of mathematical statements. In this study, Selden and Selden (1995) defined the term validation as, "The process an individual carries out to determine whether a proof is correct and actually proves the particular theorem it claims to prove" (p. 127). The main claim of this study was that students' ability to unpack mathematical statements was in some way linked to their ability to validate proofs. Furthermore, they claimed that beginning undergraduate students could not reliably unpack mathematical statements and therefore would not be able to reliably validate proofs.

In a subsequent study, Selden and Selden (2003) tested this last claim, asking whether students could in fact validate proofs in a reliable manner. They did so by asking eight mathematics and secondary mathematics education majors from an ITP course to validate three different proofs. For a basis to this study, Selden and Selden (2003) gave a broad and generalized understanding of proof validation stating, "Here we focus on proofs as texts that establish the truth of theorems and on reading of, and reflecting on, proofs to determine their correctness. We call such reading and the mental processes

[1] This works under the assumption that an Introduction to Proof (ITP) course is in fact the university course where students are formally introduced to proof (David & Zazkis, 2017).

associated with them *validations of proof*" (p. 5). While this definition did little to clarify what those mental processes were, it importantly defined validating as a part of the process of reading a proof and connecting that reading to a series of cognitive efforts to determine the veracity of a proof. The results of this study suggest that students unaided are no better than chance at determining the validity of proofs, but with guided intervention can be led to be more reliable in their validation judgments.

Since Selden and Selden's (2003) study, research on proof validation moved away from asking if students could validate to how they validate and how that compares to how mathematicians validate. This tract of research broadens some to explore the more comprehensive conception of proof reading. Alcock and Weber (2005) had 13 undergraduate students explore a single proof to help determine what students attend to in validating. Their findings suggest that by the time they are in an analysis course there are specific aspects of proofs that students learn to focus on to aid in determine the validity of a proof. Moreover, they noted that students tend to focus on the veracity of statements rather than the tenability of the statements themselves, meaning students were more concerned with what was or was not true rather than what was or was not supported by the argumentative process employed in the proof.

Weber and Mejía-Ramos (2011) studied the process that mathematicians undertake in validating proofs, suggesting that mathematicians typically do so under the semblance of three different rationales; (1) the source of the proof, (2) an actual line-by-line checking of the proof, and (3) a review of the overall method or methods used to accomplish the proof. These results to some degree echo the findings of Heinze (2010) and those of Weber (2008) who found that mathematicians tend to evaluate proofs in a

two-step process. The first step consisted of checking the structure of the argument, typically by noting the proof technique (e.g. direct proof, proof by contradiction) employed in the argument, and then, if they found that satisfactory, the second step consisted of checking the proof line-by-line. Of Weber and Mejía-Ramos' (2011) third rationale, the authors discussed the idea that mathematicians may do two activities to explore the methods undertaken, that of zooming in and zooming out. The former is that of focusing on what may be problematic portions of the argument, while the latter is that of exploring the overall structure – what Selden and Selden (1995. 2003) refer to as first-level proof frameworks – to get at the heart of the main ideas of the proof.

Weber and Mejía-Ramos' (2011) results led Inglis and Alcock (2012) to explore if students took a different approach to validating proofs than those of mathematicians. Their approach to this research was novel as they employed eye-track technology to determine the process by which 18 first-year undergraduate students and 12 research-active mathematics approached validating proofs. Their results supported Selden and Selden's (2003) findings that these students are unreliable in their validating of proofs. Furthermore, they found that these students tended to fixate on the correctness of mathematical computations which echoed the finds published by Knuth (2002) almost a decade earlier. Inglis and Alcock (2012) also found that mathematicians were far more active in their zooming in than the students, and that neither group actively zoomed out, in contrast to the findings of Weber and Mejía-Ramos (2011).

These two research teams worked in unison to determine the standards that mathematicians had in evaluating proofs. Inglis, Mejía-Ramos, Weber, and Alcock (2013) surveyed 109 research-active mathematicians to determine areas of disagreement

in mathematicians' validity judgments. To do this, they presented these mathematicians with a single proof, which they asked the mathematicians to validate, as well as a critique of the proof which the authors asked the mathematicians to comment on, giving them the chance to change their minds. According to Inglis et al. (2013), "The results of this study provide empirical support for the claim that there is not universal agreement among mathematicians regarding what constitutes a valid proof" (p. 279). Ultimately, they found that mathematician do not hold universal agreement in the validity of proofs or more generally, what constitutes a valid proof and that in practice the standards they hold about what causes a proof to be invalid are also not universal. According to their study, the standards by which mathematicians validate proof are to some degree dependent upon the domain of mathematics in which the mathematician is involved.

Directly related to their interactions with students' proofs, Moore (2016) explored the proof grading and evaluating habits of four university professors to determine if there would be consensus in evaluating and scoring students proofs as well as explicating what these professors saw as the characteristics of well-written proofs. To accomplish these goals, Moore conducted one-on-one interviews where each professor was asked to talk aloud while scoring six student proofs, indicating how the proof could have been improved and assigning a score on a ten-point scale. This was followed by a question/answer session where Moore asked the professors about their grading habits generally, and the features they thought constituted a well-written proof. A year later Moore conducted a set of follow-up interviews with each professor to further understand the variance in results from the initial study where Moore used modified versions of three

of the proofs from the original study as well as presenting them each with a seventh proof.

The professors in Moore's study varied by three or more points on four of the six proofs in the initial study. This range actually increased in the follow-up study. Beyond overlooking details that would have led to different scores (i.e., performance errors) ultimately Moore's professors held vastly different schemas for grading which had the greatest impact on the final scores they gave to each proof. In characterizing well-written proofs, the professors identified four characteristics which they deemed as important: (1) logical correctness; (2) clarity which encapsulated a variety of meanings; (3) fluency of language; and (4) a demonstration of understanding.

Conclusion. This tract of research answered quite a few questions about what a small group of students can do in terms of validating a select set of proofs, what processes mathematicians and students undertake while validating, and what standards mathematicians may hold about the validity of proof. On the other hand, there are still many questions which this research does not answer. First and foremost, on the students side, most samples taken were small (e.g., $N = 8$, $N = 13$, & $N = 18$) and the data collection, while systematic in terms of methodology, did not employ quantitative measures and controls to make larger claims about the abilities student might possess with regards to validating. In some ways, this affects the generalizability of these finds and gives space for asking more pointed questions about students' ability in regard to validating. For instance, do students improve in their ability to validate arguments as they progress in an undergraduate mathematics course of study (i.e., as they take more classes in their degree)? To answer such questions, there needs to be a uniform and consistent

way of determining students' ability of validating on a larger scale data collection. Furthermore, in terms of the arguments themselves, while the proofs for these studies are certainly interesting and meaningful, it seems important for a large scale study of students' validating ability to have proofs which are selected and codified based on the types of mistakes that are present in the arguments and which mathematicians would agree to their validity, again in order to make larger claims about what students are and are not able to do, which no study to date has done.

With regards to the research on mathematicians, the research to date presents important findings about some basic norms and practices to which mathematicians adhere both in their reading and validating of proofs. What is less clear from the research is an overarching sense of how specific aspects of proof affect validity. For instance, there is a body of research focusing on errors in proof writing (e.g., Hazzan & Loren, 1996; Selden & Selden, 1987). What is unknown from the research base is how mathematicians view these errors in terms of validity and whether mathematicians see these errors as having similar or different effects on validity.

III. Common Validity Issues and Theoretical Background

To frame this research, I begin by outlining the definition of proof that will underlie my instrument, thus the idea of a valid or even invalid proof can be ascertained with minimal ambiguity. This defining process occurs in two steps, the first is to give general meaning to the aspects of proof, and second, to give sharper contrast to this research, I set the norms about each aspect in conjunction with the target population for the proposed instrument. I conclude the defining process by leveraging these constructs to define validating. Also included in this section is a presentation of the Common Validity Issue (CVI) framework which undergirds the creation of the assessment; the main goal of this study. Finally, I briefly discuss educational assessments, giving specific attention to validity – in this case content validity – and reliability.

Arguments and Proofs

To set the tone linguistically, I first will clarify the terminology used to identify what is valid and what is not. First and foremost, the term *argument* represents the body of all purported proofs regardless of their validity. Thus, to ascribe a series of logical (or illogical) statements as an argument is to remove any notion of validity from the conversation and is akin to the use of the term purported proof in other studies. Arguments are valid-neutral. On the other hand, though this may not be in common parlance, I take term *proof* to be self-evident in reference to an argument's validity. Identifying an argument as a proof is to remove its valid-neutrality and assert that it is, in fact, valid. Finally, the class of arguments that are not proofs, and thus not valid, I adopt the term *non-proof*. Non-proofs are arguments which do not prove for some reason or another; they are not valid. I adopt this terminology first because terms like purported

proof and valid proof are cumbersome, but also because I feel that the language I use is important and the term proof holds a special position as the only argument that proves. All other arguments may do other things, but ultimately, they fail to prove.

Defining Proof

The idea of proof is nuanced in the mathematics education literature. These characterizations range from the overtly mathematical in nature (e.g., Healy & Hoyles, 2001; Knuth, 2002; Mariotti, 2000) where logic and deduction are stressed at the expense of all else, to the cognitive or social perspectives each focusing on aspects of conviction, and communal acceptance (e.g., Balacheff, 1988; Harel & Sowder, 2007). I adopt Stylianides' (2007) definition as it incorporates all three views previously mentioned (mathematical, social, and cognitive), but also explicates what might qualify as a proof. Stylianides defines proof thusly:

Proof is a *mathematical argument*, a connected sequence of assertions for or against a mathematical claim, with the following characteristics:

> 1. It uses statements accepted by the classroom community (*set of accepted statements*) that are true and available without further justification;
> 2. It employs forms of reasoning (*modes of argumentation*) that are valid and known to, or within the conceptual reach of, the classroom community; and
> 3. It is communicated with forms of expression (*modes of argument representation*) that are appropriate and known to, or within the conceptual reach of, the classroom community. (p. 291; emphasis in original)

From this definition the understanding is gained, as was previously mentioned, that a proof is a mathematical argument which is defined by three distinct characteristics

concerning statements, modes of argumentation, and representation (see Table 1). I take the phrase *mathematical statements* to represent the set of all acceptable statements. These mathematical statements represent the set of axioms, definitions, and theorems upon which a mathematical domain is built. When employed in an argument these mathematical statements are taken as true without further consideration. Modes of argumentation centers on the use of logic, reasoning, and the appropriate application of mathematical statements within an argument. Finally, modes of argument representation encapsulate ideas about the linguistic, symbolic, and diagrammatic nature of proof.

Table 1

Examples of the Three Components of a Mathematical Argument Mentioned in the Definition of Proof. Reprinted from "Proof and proving in school mathematics," A.J. Stylianides, 2007, Journal for Research in Mathematics Education, 38, p. 292. Copyright 2007 by NCTM.

Component of an argument	Examples
Set of accepted statements	Definitions, axioms, theorems, etc.
Modes of argumentation	Application of logical rules of inference (such as modus ponens and modus tollens), use of definitions to derive general statements, systematic enumeration of all cases to which a statement is reduced (given that their number is finite), construction of counterexamples, development of a reasoning that shows that acceptance of a statement leads to a contradiction, etc.
Modes of argument representation	Linguistic (e.g., oral language), physical, diagrammatic/pictorial, tabular, symbolic/algebraic, etc.

It is significant for situating this research as not just an educational activity, but also as a mathematic endeavor. By defining proof, the definition is appropriate for the classroom setting and also operational and acceptable in the mathematical world at large. In reference to this dual requirement Stylianides (2007) posited:

> Regarding the consideration of mathematics as a discipline, the definition requires that the accepted statements be true, the modes of argumentation be valid, and the

> modes of argument representation be appropriate. Regarding the consideration of students as mathematical learners, the definition requires that proofs depend on what is accepted, known, or conceptually accessible to a class- room community at a given time. (p. 294, emphasis in original)

The definition presented here walks a careful line between what is required of a mathematical argument to be a proof in the larger setting of the general mathematical discipline while simultaneously affording consideration to the social requirements that exist in the smaller setting of the mathematics classroom. For this research, this dual nature affords not only the ability to carefully view implications from an educational perspective, but for them to be potent and meaningful from a mathematics perspective.

Beyond simply making affordances for the various mathematical communities that exists, the definition of proof used in this research needs to be explicit enough to make evaluations of validity a straightforward and consistent effort. In agreement with Stylianides, Stylianides, and Weber (2017), I see this definition of proof as able to "support judgments about whether students' arguments meet the standard of proof, and if not, it can also support decisions about which specific components of students' arguments require development so as to better approximate that standard" (p. 5). Thus, once the normative nature surrounding acceptable statements, argumentation, and representation are well-defined, the definition affords a clear understanding of what is important in the context of checking for validity. Next, I discuss the general basis for setting the norms for proofs.

The Norms for Proof

The norms for this research are set from the perspective of the commonly referred to university course, introduction to proof (ITP). The choice to set the ITP class as the standard is because it is the entryway into advanced mathematics and is the setting where most students first interact with proof at a high level. As such, the classroom community used in the broad definition of proof can be understood to be that of the ITP classroom, and in this way, I heavily leverage the work of David and Zazkis (2017) as they have taken great strides in defining what is meant by ITP from a research perspective. Using this perspective focuses the understanding of classroom community toward a curricular bias as David and Zazkis' (2017) work largely identifies the curriculum and curricular materials not necessarily the classroom environment itself. I justify this bias as textbooks and other curricular materials represent a bridge between the classroom and the intended curriculum (Thompson, 2014).

Furthermore, Stylianides (2007) concedes that the classroom community overlooks the individual and focuses more on what is within reach of the community and that which "Can comfortably be assumed and used publicly without further justification" (p. 293). Therefore, while the classroom community is ill defined, David and Zazkis (2017) give a thorough understanding of the intended curriculum and the texts involved, which combine understanding is enough for situating the norms about proof for ITP classroom community. Thus, when referencing the ITP curriculum, I do so in place of the classroom community.

What follows is discussion on the general meanings for this research in regard to acceptable statements, acceptable forms of argumentation, and acceptable argument

representation. Where appropriate, I will follow the general discussion with a set of norms based upon the ITP class as outlined by David and Zazkis (2017).

Acceptable Statements

As previously mentioned, the phrase *mathematical statements* represents the set of acceptable statements, which statements are the set of axioms, definitions, and theorems upon which a mathematical domain is built. Mathematical statements have two important roles as part of the mathematics register, (1) they establish meaning, and (2) they function as a basis for building new meaning (Halliday, 1978). The first aspect is straightforward because axioms, definitions and theorems by their very natures connote meaning. The second aspect points to the *use* of mathematical statements, and beyond being markers for what is known, they are the basis upon which new knowledge is built. For mathematics, the mode of creating new knowledge is precisely the argumentative process, and in terms of argumentation, mathematical statements play the role of justifying, or in the vernacular of Toulmin (1964) warranting.

Warranting is the cognitive process of either inferring or outright stating an axiom, definition, or theorem to justify claims from the basis of some data (Toulmin, 1964). Repeating this process forms an inferential chain and is the very essence of an argument, which I will speak of more in acceptable argumentation. For the purpose of specifying acceptable statements and their significance in argumentation, understand that the role mathematical statements play in argumentation is by offering meaning to build upon and function as warrants or justification in building new meanings.

I will not lay out a list of acceptable mathematical statements for use in building an assessment, as this does not seem possible at this time, but I will instead loosely define

the scope of what is acceptable for building the assessment based upon the curriculum which this assessment is based. The set of acceptable statements for this assessment arises from the ITP course. David and Zazkis (2017) introduce the Standard ITP course whose curriculum covers "Symbolic/formal logic, truth tables, propositions, quantifiers, methods of proof (including contradiction and induction), number systems, sets relations and functions, infinite sets, and cardinality" (p. 5). Fagan and Melhuish (2018) found that proof activities in Standard ITP classes most often fell in the mathematical domains dealing with basic number systems/theory, sets, relations, functions, and cardinality. It is from this curriculum that I define acceptable statements using the vernacular common to three of the most widely used Standard ITP texts from David and Zazkis' (2017) survey, namely *Mathematical Proofs: A Transition to Advanced Mathematics* (Chatrand, Polimeni & Zhang, 2013), *A Transition to Advanced Mathematics* (Smith, Eggen, St. Andre, 2014), and *Book of Proof* (Hammack, 2013).

The *set of accepted statements* underlying proofs in the validating measure consist of mathematical statements which make up the basic set of axioms, definitions, and theorems for the Standard ITP course dealing specifically with basic number systems/theory, sets, relations, functions, and cardinality. A proof in the context of this research is any argument that is built upon mathematical statements that are common to the curriculum dealing with basic number systems/theory, sets, relations, functions, and cardinality found in the Standard ITP course or any course that could reasonably be presumed to precede such a course. This additional closing clause is necessary as often ideas like field axioms (i.e., associative, commutative, distributive, identity and inverse) are taken for granted in the study of proofs in the Standard ITP course in that they are

often used though not defined or introduced in any formal way (Fagan & Melhuish, 2018). These topics are often introduced as early as pre-algebra and are widely available for undergraduate mathematics majors despite not being formally introduced in the Standard ITP course. Thus, the inclusion of these types of mathematical statements and others from calculus, and high school algebra, trigonometry, and geometry are part of the accepted statements of proofs for this research.

Acceptable Argumentation

To help clarify argumentation, I partition the topic into three sub-constructs coming out of the literature on argumentation in proof, namely logical structure (see Selden & Selden, 1995, 2003; Weber, 2008), line-by-line reasoning (see Alcock & Weber, 2005; Inglis & Alcock, 2012; Weber, 2008; Weber& Alcock, 2005), and argument type (see Healy & Hoyles, 2000; Inglis & Mejía-Ramos, 2009; Raman, 2002; Weber, 2010). In the following sections, each sub-construct is defined generally as it relates to this research.

Logical structure. Mathematical statements like theorems have an intrinsic logical structure. A basic assumption in this research is that a theorem's logical structure, no matter how convoluted, should in turn inform how a proof of that theorem is logically structured. Selden and Selden (1995) highlighted the interdependence of logical structures in define proof frameworks, a term used to describe the overall logical structure of a proof as implied by the logical structure of the statement it proves and the proof method used to accomplish said proof (cf. Weber, 2015). For example, in Figure 1, the logical structure of the theorem is that of a standard conditional statement, $P \rightarrow Q$ where P is "M is a compact set" and Q is "each infinite subset of M has a limit point." In

consistent fashion, what is assumed is in direct conjunction with what is allowed by the theorem, and what is concluded similarly coincides with the theorem. The logical structure along with what might be termed proof method (e.g., direct proof, proof by contradiction) makes up the proof framework.

Theorem: If M is a compact set, then each infinite subset of M has a limit point.

Proof:
Let M be a compact set and N be any infinite subset of M.
$\vdots$
Therefore, N has a limit point. $\square$

Figure 1. Proof framework.

An important aspect of the proof framework is the alignment of the logical structure of the proof and the proof method being used to produce the argument. Selden and Selden (1987) point out that a common mistake students make is to assume the consequent of a mathematical statement and arriving at a trivial conclusion. Moreover, in proving a theorem, to assume more or less than what is allowable based upon the theorem being proven is strictly at odds with aligning the logical structures of the argument and the statement being proved. This concept is known as weakening the theorem, which Selden and Selden (1987) characterized as, "When what is used is stronger than the hypothesis or when what is proved is weaker than the conclusion" (p 464-465). For defining proof, this means an argument must assume exactly that which the statement being proved allows and that the proof does not assume that consequent and arrive at a trivial solution.

Line-by-line reasoning. Mathematical statements have both meaning and function. It is the inferential chain which defines the line-by-line reasoning of a proof. As the line-by-line reasoning of a proof is an extension of the logical structure of the proof,

any assertion in a proof proceeds as a logical outcome of previous assertions, which in iterative fashion should trace back inferentially to the basic assumptions of the proof framework.

While admittedly overly simplistic, Figure 2 presents an example of how the proof framework begins and ends the inferential chain of line-by-line reasoning. The set of assumptions, A, granted by the antecedent of the initial statement starts the chain of inferences as data which combine with warrant X leads to the claim B. Once claimed, B becomes part of what is accepted. It can in turn be data itself, and when combined with a warrant Y leads to claim C. This cycle ultimately ends as C is now derived, and, therefore, is used as data all its own and when combined with the warrant Z leads to the logical conclusion D. The connective structure, or justification, in an implication between any previous assertion(s) and a new assertion is what is termed a warrant (see Toulmin, 1964; Weber & Alcock, 2005) which comes from the set of mathematical statements. Thus, for a proof to be logically consistent from line-to-line, the warrants must be accurate and appropriate for the claim they support.

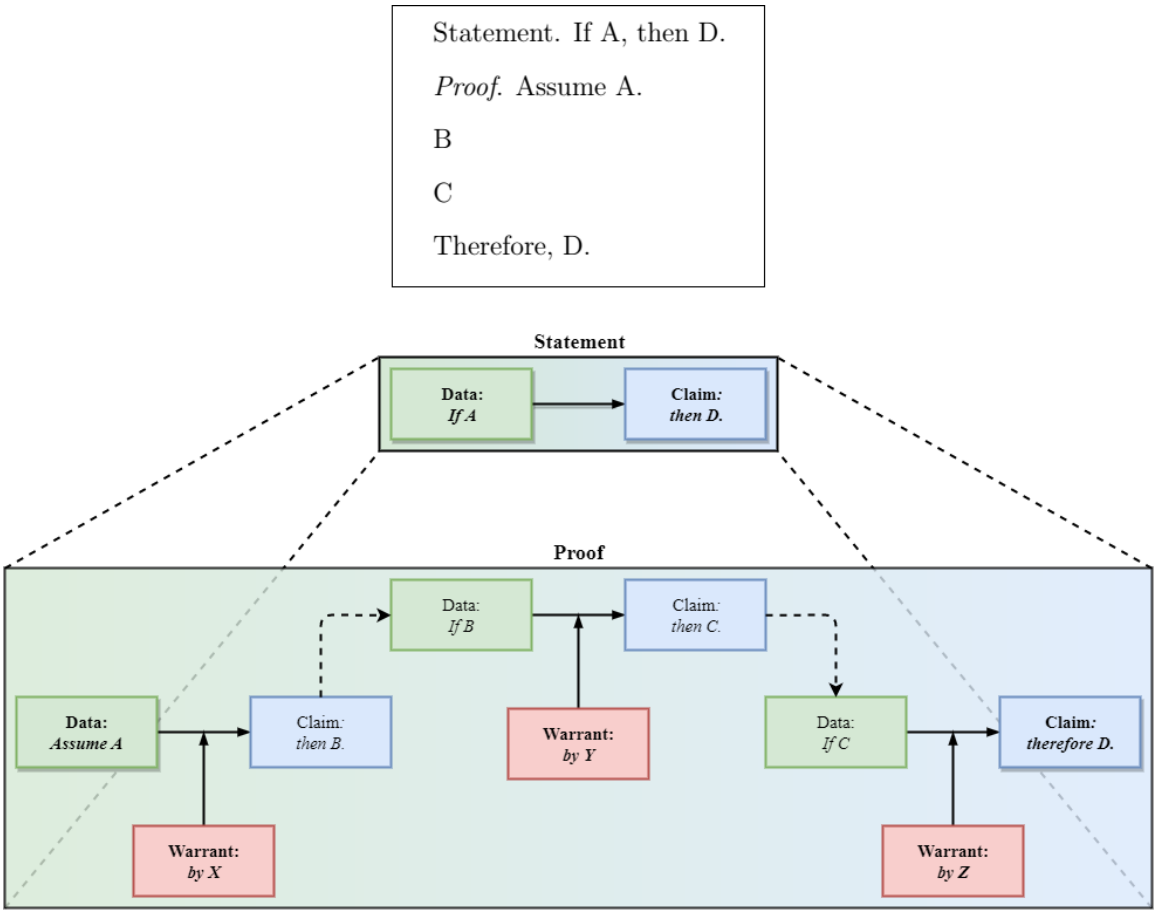


Figure 2. Line-by-line reasoning is built upon the logical structure and is supported by warranting. Adapted from Weber and Alcock (2005).

In terms of what warranting means for building an assessment of validating competencies, there are two distinct aspects I define to situate this research. The first deals with the necessary and sufficient level of explicit warranting in a proof: how often and for what types of inferential claims an explicit justification in terms of a warrant must be stated outright. This concept stems from professors' and teachers' grading habits in the classroom (e.g. Moore, 2016) and is a problematic topic as there seems to be no consensus on this idea generally. As it is most often defined by the society for which proofs are created, the notion of there being a necessary level of explicit warranting is an artificial construct of the didactic process. To some degree, it would seem incongruous to attempt to force such an arbitrary if not artificial norm on an instrument whose target

population includes students who will have experienced widely varying norms on this front. On the other hand, while this idea is contextual, by putting this notion in front of mathematicians who teach these courses and so often set these norms – both for the classroom and in practice for journals and publications generally – perhaps some consensus can be reached for the ITP level. Thus, while the professors in Moore's (2016) study may have considered stating warrants explicitly as an important factor for grading, this may not be the case universally (Alcock & Weber, 2005; Weber & Alcock, 2005). It is worth exploring from a mathematician standpoint to better understand if there exists any generality at all.

Warrants (whether implicit or explicit) must be appropriate to justify the connection between data and claim. For example, in Figure 3, a common problem for students learning modern algebra is the use of Lagrange's Theorem. As stated in Hazzan and Leron (1996), statement (A) overtly warrants using Lagrange's Theorem, which is the incorrect use of the theorem and therefore invalidates statement (A). On the other hand, statement (B) is invalid not because $\{a_n\}$ is not convergent, because it might be dependent upon the sequence. Instead, (B) is invalid because the implied warrant – *every bounded sequence is convergent* – is categorically false. For instance, the sequence $\{(-1)^n\}_{n\geq 1}$ is bounded such that $\forall n \geq 1, |(-1)^n| \leq 1$, but it does not converge. The warrant can be inferred from the fact that the statement is talking about the convergence of a sequence and the stated data, "it is a bounded sequence." So, while the antecedent *may* be true, if the warrant is false the statement itself is invalid. In either case, whether explicit or implicit warranting, this is the logical issue of line-by-line reasoning spoken of in this section.

(A) $\mathbb{Z}_3$ is a subgroup of $\mathbb{Z}_6$ by Lagrange's theorem, because 3 divides 6.

(B) The sequence $\{a_n\}$ is convergent, because it is a bounded sequence.

Figure 3. Implicit and explicit warrants.

Once again, because of the inconsistent manner in which warranting is treated in the curriculum (Fagan & Melhuish, 2018), the amount or level of explicit warranting within an argument will not serve in defining the norms about proof. Thus, a proof can be understood to either explicitly warrant all claims, or leave it to the reader to, as Alcock and Weber (2005) stated it, infer warrants. On the other hand, as the idea of the appropriateness of a warrant within a proof is a measure of the consistency of line-by-line reasoning, a proof can further be defined from the standpoint of justification as an argument where each warrant, implied or explicit, is accurate and appropriate for the claim it supports. Additionally, as a direct consequence of this understanding, in order for an argument to be valid both at the ITP level and the more general mathematical level, the norm is that computations must be carried out correctly (Selden & Selden, 2003; Weber, Inglis & Mejía-Ramos, 2014).

Finally, in relation to line-by-line reasoning, there is some debate amongst differing school of thought about the effect of irrelevant and extraneous statements on the validity of a proof (see Dawkins & Weber, 2016; Selden & Selden, 2003). The argument against extraneous statements is that they affect the clarity and flow of an argument and can often confuse the reader making it difficult to both validate and comprehend. Despite the difficulties that extraneous statements can cause, they do not a priori invalidate the argument by their very presence (Dawkins & Weber, 2016; Selden & Selden, 2003). Therefore, while they are undesirable in a well written proof, they can still exist as part of

the line-by-line reasoning of an argument deemed to be an acceptable proof. To avoid unnecessary convolution, I avoided needlessly including extraneous statements, however, as I used previously collected student work as inspiration in writing the instrument, extraneous statements may exist in these proofs. As such, I am careful in considering their effect on the line-by-line reasoning of the argument before I decided on whether to keep or remove said statements.

Argument type. By type of argument, I embrace the notion Weber (2010) referred to as the "Types of evidence contained in the argument included to convince the reader about the veracity of the theorem being proved" (p. 307). Thus, this discussion concerns the strength of the argument, whether it is example-based argumentation or general in nature. Stylianides (2007) stated that his definition,

> Describes a special class of arguments (those that qualify as proofs) without suggesting that other classes of arguments represent less valuable ways of knowing and doing mathematics. Indeed, there are many valuable ways of reaching valid conclusions (e.g. arguments by analogy) that may not be logically "tight" enough to meet the standard of proof. (p. 292)

There are many argument types which an individual may accept as valid, not all of which are valid for proving.

Generally, arguments break into two categories: deductive and inductive. Deductive arguments offer a guarantee that when the premise is true, the conclusion is also true, and necessitate the logical qualities mentioned previously (i.e., logical structure and line-by-line reasoning). Their results are logical and general, and follow from and extend theory. Conversely, inductive arguments use empirical evidence to establish cases

of theory, and while useful in building general results, inductive arguments are not general in nature.

A third and more elusive argument type is that of diagrammatic arguments. Diagrammatic arguments are not elusive because they are hard to identify or define, but rather because they present a conundrum in their acceptance and purpose. Indeed, Nelson's (1993) collection of *Proofs Without Words* is an interesting collection of arguments that in some case are canonical (e.g., proofs of Pythagorean's theorem) while others appear to be little more than interesting mathematical tidbits. Despite the moniker of proof, Nelson himself pointed out that these diagrams were not in fact proofs.

More often, diagrams act as heuristics in the undergraduate ITP course and have minimal sway in the establishing the veracity of a mathematical statement. For this research I again default to the classroom community of the Standard ITP course, and as Samkoff, Lai, and Weber (2012) pointed out, formal proofs at this level are based on sound mathematical statements, and deduction. They state, "The inferences within the proof are expected to be based on deductive logic, not the appearance of the diagram" (p. 50). Furthermore, deductive arguments are the primary form of acceptable proofs amongst mathematicians (Dawkins & Weber, 2016). Despite evidence that K-12 students and teachers may accept empirical arguments (Healy & Hoyles, 2000; Knuth, 2002), ITP students generally see empirical arguments as invalid (Weber, 2010). For this reason, the proofs presented in the validating measure are limited to this argument type: deductive arguments of general statements.

Acceptable Argument Representation

I take the position of Dawkins and Weber (2016) as they argued that proofs should use established symbolic conventions in their construction. As such, a proof is an argument which abides by these conventions as well. According to David and Zazkis (2017) one of the many topics discussed in the Standard ITP class is that of symbolic logical quantifiers (e.g., universal quantifier ∀, existential quantifier ∃) as well as other symbolic notation (e.g., ∧, ∼, ∈, ⊆, ≡, ∪). Furthermore, despite students' apprehension about proofs based upon algebraic computation (e.g., Knuth, 2002; Weber, 2010), once in an ITP class, it is reasonable to assume a student has spent time in their academic career learning the symbolic conventions of basic algebra, trigonometry, and calculus. For example, in Figure 4, whether faced with the highly symbolic proof as in Proof *A*, or the less symbolic more linguistic proof as in Proof *B*, students at the ITP level and above should be comfortable with either the heavily symbolic or more linguistically grounded conventions and base an evaluation upon the correct use of these conventions as well as the previously listed aspects of a proof. Thus, the proof norms about representation are that arguments should correctly[2] use the conventions of symbolic notation which are common[3] to the curriculum dealing with basic number systems/theory, sets, relations, functions, and cardinality found in the Standard ITP course or any course that could reasonably be presumed to precede such a course.

[2] This along with line-by-line reasoning implies that computations within proofs should be without mistake.
[3] Once again, common is defined by the texts, *Mathematics Proofs: A Transition to Advanced Mathematics* (Chartrand, Polimeni, & Zhang, 2013), *A Transition to Advanced Mathematics* (Smith, Eggen, & St. Andre, 2014), and *Book of Proof* (Hammack, 2013).

Claim:
If A, B, C and D are sets, then $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

Proof *A*:

$$\begin{aligned}
(x,y) \in A \times (B \cup C) \quad & iff \quad (x \in A) \wedge (y \in B \cup C) \\
& iff \quad (x \in A) \wedge [(y \in B) \vee (y \in C)] \\
& iff \quad [(x \in A) \wedge (y \in B)] \vee [(x \in A) \wedge (y \in C)] \\
& iff \quad [(x,y) \in A \times B] \vee [(x,y) \in A \times C] \\
& iff \quad (x,y) \in (A \times B) \cup (A \times C)
\end{aligned}$$

Therefore, $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

Proof *B*:
Case 1: Prove $A \times (B \cup C) \subseteq (A \times B) \cup (A \times C)$ by letting $(x,y) \in A \times (B \cup C)$, thus $x \in A$ and $y \in (B \cup C)$. From the latter we get that $y \in B$ or $y \in C$, hence $x \in A$ and $y \in B$ or $x \in A$ and $y \in C$. This implies that either $(x,y) \in A \times B$ or $(x,y) \in A \times C$, and therefore, $(x,y) \in (A \times B) \cup (A \times C)$ thus, $A \times (B \cup C) \subseteq (A \times B) \cup (A \times C)$.

Case 2: Prove $A \times (B \cup C) \supseteq (A \times B) \cup (A \times C)$ by letting $(x,y) \in (A \times B) \cup (A \times C)$, thus either $(x,y) \in A \times B$ or $(x,y) \in A \times C$. This implies that $x \in A$ and $y \in B$ or $x \in A$ and $y \in C$, hence we see that $x \in A$ and $y \in B$ or $y \in C$. Meaning that $x \in A$ and $y \in (B \cup C)$ thus, $(x,y) \in A \times (B \cup C)$. Therefore, $A \times (B \cup C) \supseteq (A \times B) \cup (A \times C)$.

Therefore, $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

Figure 4. Representation conventions, adapted from Smith, Eggen, & St. Andre (2014)

Conclusion – Proof Norms

The understanding of what defines the norms for proof in this research is extensive to be sure, but as mentioned previously, by clearly defining this construct, the task of creating prompts for the proposed assessment will be direct and less ambiguous in terms of what is valid and what is not. Thus, to be clear, the proofs in this research and for the proposed assessment are arguments that:

1. are built upon mathematical statements and conventions of symbolic notation which are common to the curriculum dealing with basic number systems/theory, sets, relations, functions, and cardinality found in the Standard ITP course or any course that could reasonably be presumed to precede such a course,
2. are deductive in nature and align structurally with the implied logical structure of the statement being proven,

3. incorporate valid logical connectives from line-to-line where each warrant, implied or explicit, is accurate and appropriate for the claim which it supports, and whose computations are correct,

With the norms of proof now set for this research, what is left is to define proof validation[4].

Limiting Scope

In defining proof, I introduced the idea of logical structure which in turn lead to discussing proof frameworks. A part of what is implied by a particular proof framework is a particular proof method (e.g., direct proof, proof by contradiction). While the norm for the Standard ITP course is to introduce students to a wide variety of proof methods (Fagan & Melhuish, 2018), for the purpose of the proposed validating instrument, I focus solely on direct proofs[5]. In the example in Figure 1, the proof framework is indicative of a direct proof, where for a conditional statement, $P \rightarrow Q$, the antecedent, P, is assumed true and the desired conclusion is that the consequent, Q, is also true[6]. While it would be interesting to use this assessment to explore students' ability to validate multiple proof types[7], for now, to assure the best possible instrument and reduce the number of variables in task creation, proof will be defined in terms of direct proofs only.

Finally, it should be noted for this study, I am most interested in general arguments. The effect of this is to exclude mathematical statements where a single

[4] The phrase proof validation seems inappropriate for the task it describes as the name implies the idea of innocent until proven guilty, in the sense that an argument is a proof until shown to be otherwise. While I dislike this way of thinking and feel that it could be detrimental to students' own thinking, I stick with it as it is the convention in the mathematics education literature.

[5] Admittedly, this is a coarse use of the term *direct proof* but should suffice and is clearly delineated enough for there to be no mistake about what is meant.

[6] Note that this proof method easily supports proofs which have multiple cases or are proofs of biconditional statements.

[7] This is certainly a possible source of future research, to expand this effort to other proof types.

example or counterexample suffices as a complete argument for or against the veracity of the statement in question. For example, Figure 5 presents an existential quandary; does such an integer exist? The proof then is a matter of showing that at least one such element from the integers exists. This is not the sort of statement – in this case a result to be proven – nor argument I wish to include as part of the proposed assessment[8].

Result to Prove: There exists an integer whose cube equals its square.

Proof: Since $1^3 = 1^2 = 1$, the integer 1 has the desired property.

Figure 5. Existential quandary adapted from Chartrand, Polimeni, and Zhang (2013)

Validating and Proof Validation

Generically, proof validation is the act of judgement or evaluation which leads the reader to identify the correctness of an argument. In similarly broad strokes, Selden and Selden (2003) called proof validation, "the reading of, and reflection on proofs to determine their correctness" (p. 5). Although both of these notions are vaguely accurate, neither gives a deeper account of what may be involved in the process of proof validation.

While the process of defining proof for this research is a normative procedure as what constitutes proof is defined by those who use it, validating is far less visible as it is a cognitive process (Selden & Selden, 2003) which takes into account one's conception of what is required of an argument to be a proof. In fact, the process of validating is so thoroughly cognitive in nature that it has brought about some novel methodological approaches to study how individuals validate (see Inglis & Alcock, 2012). Weber and Mejía-Ramos' (2011) investigated why and how mathematicians read published proofs

[8] Also an interesting possible avenue for further exploration.

and found that mathematicians may *zoom-in* and *zoom-out* in the process of validating proofs. This was a pair of processes that the authors conjectured occurred as mathematicians, "May (intuitively and implicitly) assign a probability p_i to his or her level of confidence that the i^{th} inference of the proof is correct" (p. 339). In zooming-in, the mathematician focuses on inferences that were problematic in nature, whereas in zooming-out they consider the overall structure of the proof. Similarly, Ko and Knuth (2013) found mathematics majors claimed to do two activities which were very similar to those from Weber and Mejía-Ramos' (2011) study; judging the arguments structure and evaluating the line-by-line reasoning.

An additional part of the cognitive process of validating proofs is checking the semantic content of the argument. Alcock and Weber (2005) supported this idea as they suggested that in order to validate proofs, the individual needs not only check the logical aspects of the proof, but also plausibility of the semantic content. As the semantic content represents the mathematical meaning held within an argument, checking it requires the individual to attend to inferred warrants, calculations, and the meaning of words, phrases, axioms, definitions, theorems, symbols, and quantifiers. Anything which belongs to the mathematics register – those objects which carry mathematical meaning – becomes a focal point for inspection during validation when checking the semantic content.

Based on this understanding, I assert that checking the logical structure, line-by-line reasoning, and the semantic content encompass the major aspects of proof validation (e.g., Alcock & Weber, 2005; Ko & Knuth, 2013; Selden & Selden, 2003; Weber & Alcock, 2005). From this standpoint proof validation involves:

1. Checking the alignment of the logical structure of an argument against the logical structure of the mathematical statement it claims to prove.

2. Assuring that each new assertion is supported by any combination of previous assertions.

3. Inferring, identifying, or creating sub-arguments that bridge the gaps that exists in the reasoning within a proof.

4. Scrutinizing the appropriateness of the warrants, implied or explicit, for each inference.

5. Making sure calculations are error free.

6. Checking that symbols and other representation are used consistently and correctly.

These six aspects directly link back to the definition of proof previously outlined for this research. Aspects 1 through 5 all deal with modes of argumentation as each, to some degree, deal with logical structure and line-by-line reasoning employed in an argument. Additionally, aspect 4 deals with the set of acceptable statements as it focuses on how mathematical statements are used in arguments. Lastly, aspect 6 deals with modes of argument representation as they both focus on the symbolic nature of an argument.

Validating in the Face of Comprehending and Constructing

Validating and comprehending. Mejía-Ramos et al. (2012), at a rudimentary level, define proof comprehension as the ability of an individual to understand a proof. Through their framing, Mejía-Ramos et al. (2012) delineate between understanding at a *local* and *holistic* level. Local comprehension focuses on: (1) the meaning of terms and statements, (2) the logical status of statements and proof framework, and (3) justification

of claims. On the other hand, holistic comprehension emphases: (1) summarizing via high-level ideas, (2) identifying the modular structure, (3) transferring the general ideas or methods to another context, and (4) illustrating with examples.

There is a possibility that validating and comprehension are related in some way, but at this point there is no evidence in the literature strong enough to link the two. Mejía-Ramos et al. (2012) outlined a set of possible aspects of comprehension, and admittedly – even prudently – there is overlap with their framework, and the framework presented here. For instance, the categories for local comprehension Mejía-Ramos et al. identify are shared with my framework, but I claim there are some important differences in how the comprehension assessment framework views these categories, and how I have set them up here.

For instance, while ideas about proof frameworks are obvious in the second category of the comprehension framework, Mejía-Ramos et al. frame them as an act of identifying the proof framework while not asking the individual to make any larger claim. For the work of validating, it is not simply enough to identify the proof framework: the reader needs to identify this structure, identify the analogous logical structure of the mathematical statement the argument is built upon, compare the two, and decide if the two align in a sufficient manner as to allow for the argument to prove the statement. Even considering the holist idea that Mejía-Ramos et al. outline as *summarizing via high-level ideas* does not ask the student to judge these ideas against the standard of the mathematical statement the argument is built upon.

Comprehension, as Mejía-Ramos et al. defined it, is most interested in having the individual identify particular aspects from the argument. But, validation as I have framed

it, goes beyond what comprehension is framed as, and asks the individual to: (1) identify particular aspects from the argument; (2) identify aspects about the initial mathematical statement being proven as well as the other inferred mathematical statements employed in the argument, (3) identify how those aspects affect aspects of the argument or the argument as a whole, (4) *evaluate* the appropriateness of each effect, and possibly (5) bridge logical gaps through the *creation* of new arguments. From Mejía-Ramos et al. (2012) framing, the cognitive load of comprehension, is mainly concerned with identifying features and possibly transferring those features to other settings, whereas validating builds upon this and adds the tasks of evaluating and creating. More importantly, at no point in Mejía-Ramos et al. (2012; 2017) are students confronted with an argument that is a non-proof. This aspect is not even considered in the comprehending framework or construction of their assessment but is a key aspect for validating as arguments can be invalid. Thus, while it is quite possible that validating is an evaluative extension of comprehension, not only is there no evidence to support this claim, there are striking differences in what each asks of the individual.

Validating and constructing. When compared to validating and comprehending, construction would appear to be a much larger set of cognitive processes than either of the other two. That being the case, the phrase *proof construction* seems to be an ill-defined but commonly used colloquialism in the mathematics education literature. In the microcosm, proof construction is writing an argument that proves a mathematical statement, though this does nothing to address some of the more complicated and nuanced processes involved in creating such an argument. In the macrocosm, proof construction encompasses a myriad of processes including conjecturing and generalizing,

semantic exploration, instantiating mathematical ideas and objects, applying strategic knowledge, intuitive logical exploration and syntactic manipulation, justifying and warranting, and perhaps even validating (see Fischbein, 1983; Pedemonte, 2007; Moore, 1994; Selden & Selden, 2003; Weber, 2001; Weber & Alcock, 2004).

Regarding validating being an explicit part of constructing, Selden and Selden (2003) posited, and I adopt the notion, that:

> One constructs a proof with an eye toward ultimately validating it and may often validate parts of it during the construction process. In fact, the final portion of a proof construction is likely to be validation of that proof. That is, each process, validation and proof construction, entails the other. (p. 6)

Thus, the claim is laid that there is a circular relation between construction and validation. At this point, this is little more than a claim as no study has explored the relationship between validating and constructing in any scale large enough to generalize this relationship beyond pocket cases. While studies like Powers, et al. (2010) supports this claim, as their research indicates that focused attention on validating activities can lead to increased proof-constructing ability, their study has limited generalizability due to the homogeny of their sample, as well as confounding issues due to unforeseeable interference in their study. While I adopt the idea that validating is part of the process of constructing proofs, I wish to be careful as more research is needed to establish this claim.

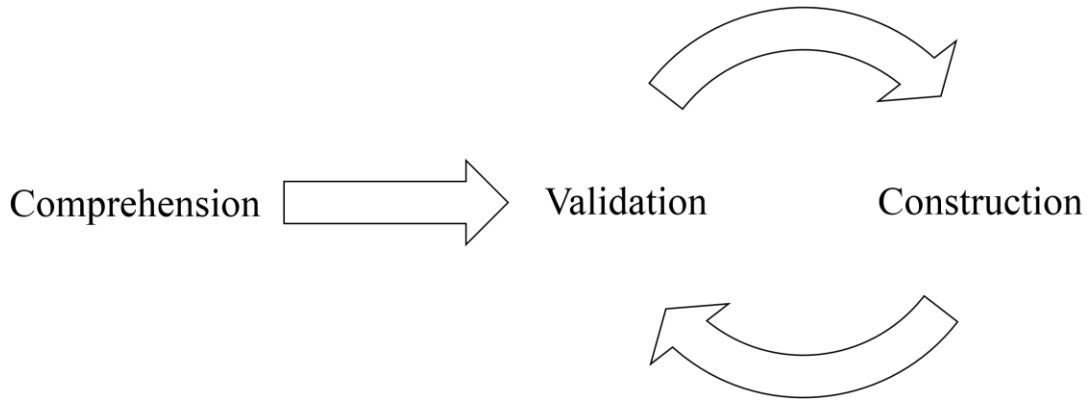


Figure 6. A possible relationship between comprehension, validation and construction.

Conclusion. For this research, I adopt something of a hierarchical notion of comprehension, validation, and construction in terms of proof (see Figure 6). Comprehension entails being able to identify and transfer notions from within an argument. Validation builds upon comprehension in that it too requires an individual to identify and transfer, but also evaluate and create ideas in the process. Construction is the top-level structure which entails validating, and by extension comprehending. As validation requires the creation of sub-proofs, there may be instances where validating is the top-level structure with constructing existing as a supporting role for validating. More research in this area is needed to better understand how these complicated and underdefined constructs work. This study is one such undertaking.

Creating an Objective Instrument to Capture a Subjective Activity

The literature, and my theoretical orientation, acknowledges validating as a subjective activity that is dependent on the context and mathematician validating. At face value, this may reflect an inability to declare an argument as valid or invalid. However, I contend there are a set of validity issues too egregious that they will be accepted as universal within the ITP setting. From a design perspective, a consensus process was needed to identify arguments that are universally deemed valid or invalid to construct the test

around. This consensus process – outlined in detail in the methodology for Phase 1 – allows for the assessment proposed in this study to meaningfully ask students in the ITP setting to make binary validity judgements despite the non-binary nature of validity in mathematics more generally.

Assessment Framing - Common Validity Issues[9]

This study leverages the idea of *issues in proof writing* which focuses on the prevalent validity issues amongst undergraduate mathematics major's own written arguments (Hazzan & Leron, 1996; Selden & Selden, 1987, 2003). Though these issues arise in the written arguments of students, Selden and Selden (1987) frame them more as fundamental issues with student's mathematical reasoning. Thus, the understanding that students have – perhaps more aptly the lack of understanding – is more prevalent than simply in constructing arguments but could conceivably extend to reading arguments and their ability to validate what they are reading. It is for this reason that these types of validity issues are important in the context of this study.

The validity issues considered for this study fall into one of six categories as presented in Table 2. Taken as a whole, this set of issues represent what I refer to as the Common Validity Issue (CVI) framework. This framework represents the basis upon which argument creation occurred for this study. It is important to note that the categorizations for the CVI framework came from observations in studies whose focus, with few exceptions (e.g., Selden & Selden, 2003), were not necessarily that of the ITP classroom or its students. Rather these studies focused on undergraduate algebra (Hassan

[9] Extending the theoretical framing to a usable analytic framing was a large portion of this study but not the main goal. This begs the question of whether this is a result or a part of the framing itself. To remove this ambiguity, I make note that what is to follow is presented here as part of the discussion on the creation of the assessment itself and not, in circular manner, a discussion of the framing of the first part of this study.

& Leron, 1996; Selden & Selden, 1987) and real analysis (Alcock & Weber, 2005; Weber, 2001), with Alcock and Weber's (2005) work being that of a theoretical contribution encompassing advanced undergraduate mathematics generally. This does nothing to lessen the validity of the CVI framing. In fact, it is very safe to assume that the issues students have in more advanced courses, like algebra and real analysis, either began while they were first being introduced to proofs in an ITP course or are common issues that students in such a course might have.

Table 2

Common Validity Issues

Issue (Abbr.)	**Definition**
Assuming the Conclusion (AC)	An argument assumes the consequent (conclusion) of the proposition it is claiming to prove and attempts to show that the antecedent is a direct consequence.
Circular Reasoning (CR)	An argument assumes the consequent (or antecedent) of the statement it is claiming to prove and comes to a trivial conclusion, namely the consequent (or antecedent) once again. Within an argument a claim is made and used to argue to trivial ends, the claim itself. ($P \to Q \to \cdots \to P$). An argument uses the proposition it is trying to prove.
Logical Gap (LG)	An argument omits a portion of reasoning; the argument has a hole. This could be thought of as a lack of a data → warranting → claim turn, or any individual portion of one where such would seem prudent.
Misuse of Notation (MN)	Within an argument, proper notation or variable naming conventions are not adhered to, or notation and variable naming conventions are used inconsistently.
Warranting (W)	Within an argument, an error in justification is made either explicitly or implicitly. This can take the form of an incorrect explicit warrant, or an incorrect implicit warrant which may emerge as an arithmetic or computational error.
Weakening the Theorem (WT)	An argument proves less than what is implied by the statement being proven or begins by assuming more than is permissible.

There might be some question as to the exhaustive nature of this framework from an ITP setting. It is possible that by the time students reach algebra or analysis, there are fundamental issues in their understanding about the validity of arguments that are

overcome, which are still prevalent at the ITP level. To ensure this was not the case, when testing the framework in the expert phase of this study, the door was left open for new validity issues to be introduced by mathematicians from their observations as instructors in the ITP setting. While many suggestions arose from both the mathematician survey and the two focus groups that followed, none were novel contributions or contributions which could not be encompassed by an existing categorization. Additionally, keep in mind this framing is looking at *common* validity issues, and is not completely exhaustive. It is, therefore, safe to assume that the CVI frame is stable for the sake of this work.

In the remainder of this section, I discuss each of the six categories of the CVI framework and link them back to the theoretical framing of *proof* from earlier in the chapter. It should be noted that in some form or another, each of these categorizations interacts with the framing dealing with both the set of acceptable statements and modes or argument representation. What is left is the task of organizing each issue categorization according to the concept of acceptable argumentation, specifically the ideas of logical structure and line-by-line reasoning. For additional commentary on the methods for creating this framework, as well as examples, please refer to the methodology and results chapters. The former chapter explicates the process by which this framework was created, and the later chapter presents the framework as a result of the first half of this study and how it was used in the second half: the creation of the assessment.

Assuming the conclusion (AC). Selden and Selden (1987) outlined the common pitfall of beginning an argument with the conclusion of the proposition or statement to be

proven and arguing to the ends of an "obvious truth" (p. 460). While this works in the case of something like *proof by contradiction* – though the obvious truth in this case is a contradiction and is therefore at odds with the truth – this was not the observation or case to which Selden and Selden (1987) were referring. Instead, assuming the conclusion is a validity issue which proves the converse of a statement is always equivalent to proving the original statement itself; or, more simply, a statement and its converse are always equivalent. In terms of writing, when assuming the conclusion, the author of an argument constructs a logical connective by beginning with the consequent of the statement to be proven and then shows the antecedent is a direct consequence.

In terms of validity, this issue is apparent in all statements where the converse is not equivalent to the original statement. For the reader, identifying this issue in a written argument requires a balance of understanding the logical structure of the statement and what that implies about the logical structure of the argument. This interplay was later commented on by Selden and Selden (2003) as a knowledge about the first-level proof framework, and as such this type of validity issue is linked to the overall logical structure of the argument.

Circular reasoning (CR). This issue is exactly as the name implies, as circular arguments reason from an initial condition back to said initial condition (Selden & Selden, 1987). This can take the form of an entire argument being circular, some portion of the argument circling back on itself, or an argument which uses the proposition it is proving as a mean to prove said proposition. Selden and Selden (1987) point out this sometimes occurs in blatant fashion but is seen in much less obvious ways too. For

instance, an argument can sometimes be circular as it argues for some equivalent other version of the initial conditions.

Another common instantiation of circular reasoning, which in some respects feels like assuming the conclusion, is when an argument begins with the conclusion of the proposition being proven and manages to meander back to itself (Selden & Selden, 1987). I differentiate this from the former categorization as assuming the conclusion rests on the specious understanding that a statement and its converse are equivalent, whereas in the situation of circular reasoning, the error rests in the understanding that the conclusion warrants itself. Since circular reasoning can encompass an entire argument from beginning to end or can be found within a single statement in an argument, this validity issue relates to both the logical structure and line-by-line reasoning of an argument.

Warranting (W). I outlined warranting with considerable depth in the theoretical framing for this study. For more on warranting from a theoretical perspective, please refer to the earlier sections on *acceptable statements* and *line-by-line reasoning*. By extension, such issues hereafter identified as warranting are linked to the framing concerning line-by-line reasoning.

From a validity issue standpoint, the character of warranting in this study is multifarious and enjoins not just problematic linguistic inferences (see Figure 3), but also issues with arithmetic and calculations. The former classification of warranting issues I covered, but the latter needs some justification.

The reason for these latter classifications, arithmetic and calculation errors, is due to their relationship to fundamental divides in understanding of sometimes basic mathematical concepts. These concepts are often not the main goal of proof-based,

advanced undergraduate courses like ITP, modern algebra, real analysis or topology, but are the target content of classes like high school algebra, geometry, trigonometry, and calculus. For example, learning to square a binomial expression is certainly not a part of the ITP curriculum (Fagan and Melhuish, 2018), but at times is requisite knowledge for the topics these classes cover. Squaring a binomial, or even raising a binomial to some higher power, follows a set pattern governed by the definition of squaring; specifically, multiplying a factor by itself. Any miscalculation of squaring then is based upon a misuse of said definition – whether intended or otherwise – and can therefore be best understood in terms of a mis-warranting, or more simply a warranting issue. For this reason, arithmetic and calculation errors do not have their own classification but serve as a subclassification of warranting.

Logical gap (LG). Issues surrounding logical gaps most exemplify the relationship between data and explicit warrants. In proof writing, the issue of logical gap arises from the need for expediency in the communicative efforts of proving. Alcock and Weber (2005) point out this limitation noting if mathematicians or students explicitly stated every data and warrant for a proof, even a single proof would be unreasonably protracted. This implies there is a level of implicit proving in written arguments which is allowable but walks the line between acceptable and unacceptable. This is an important issue for student and mathematician alike to understand

Logical gaps differ from warranting because warrants and claims are missing in this categorization, whereas the category of warranting only focuses on the tacit and overt justifications in an argument. Selden and Selden (1987) spoke of holes, or logical gaps, where logical deduction does not follow directly from any previous part of the argument,

and for consistency, a sub-argument justifies the connection. For instance, a characterization of a logical gap could be seen if in Figure 2 the data, warrant and claim, "B implies C because Y" were missing. The argument becomes something more trivial as an entire chunk of logic is missing where the data B, along with the warrant Y, connects claim A with claim C, a seemingly crucial development for A to imply D as claimed. This sort of validity issues, much like warranting, is an issue of line-by-line reasoning where the line-by-line consistency breaks down as logical gaps become more prevalent.

Misuse of notation (MN). Due to the important role notation plays in communicating mathematics, there are numerous ways its misuse could invalidate an argument. Selden and Selden (1987) identify no less than six different possible incarnations of notational misuse[10] which cause issues in proof writing and are ways in which arguments can become invalid. For this reason, it is possible an assessment could be constructed to look at this singular validity issue; this is not the focus of this study. I know there are many types of notational misuse but also recognize in my attempt to make a broad assessment of validating abilities at the ITP level, only a few, or even one example, makes it into the assessment itself.

In broad terms, notational misuse is when proper notation or variable naming conventions are not adhered to, or when notation and variable naming conventions are used inconsistently. This general understanding encapsulates the six ideas presented by Selden and Selden (1987) accepting that each is relevant but gives me the option of choosing a single or pair of notational issues to be representative in assessing students'

[10] Selden and Selden (1987) identified the following six notational issues: (1) names confer existence; (2) apparent differences are real; (3) element set interchanged; (4) overextended symbols; (5) national inflexibility; and (6) using information out of context.

ability to identify such issues. For this purpose, I let the mathematicians decide which issues are most problematic.

Weakening the Theorem (WT). This categorization is one of the more straightforward classes in the CVI framework. Weakening the theorem is akin to running a race and stopping before the finish line, or showing up to a triathlon without a bike or swimmers cap. In the former case the goal was not accomplished, and in the latter the wrong basic assumptions were made about the race. Selden and Selden (1987) outlined weakening the theorem as an argument which proves less than what is implied by the statement being proven or begins by assuming more than is permissible. In both cases this occurs as a structural issue where the first-level proof framework is violated by either coming short of the intended summation or assuming too much with regards to the statement being proven.

Valid. Though not a part of the CVI framing, valid items are an important aspect of assessing students' ability to validate arguments. I defined valid items throughout the mathematician survey and throughout the process of creating the assessment as items which do not include the above CVI framing issues. If there are no valid arguments amongst a set of invalid arguments, then only some of a student's ability is tested. Additionally, there is the possibility students learn to recognize there are no valid arguments if nothing obvious presents itself. It was therefore important to present mathematicians with a set of arguments with none of the CVI issues for these reasons, and to test the CVI framing to see if another, unforeseen, issue arose in any of these seemingly valid arguments. Therefore, these valid items did indirectly test the CVI framing so no other issues arose that were not previously presented.

Proof Comprehension

Though this assessment tests a student's ability to validate arguments, it is possible that in attempting to validate arguments, students might work in a mindset of simply trying to understand the arguments. This understanding might act as a replacement construct for some students and be the main reason for judging each argument. It is important for the means of understanding students' processes that proof comprehension is a consideration throughout the analysis of students' interactions with the assessment.

Proof comprehension (Mejía-Ramos, et al., 2012) is broken down into two main distinctions: local versus holistic understanding. Local understanding deals mostly on the line-by-line understanding where terms and individual statements are the main focus. Here understanding focuses on ideas concerning; (1) meaning of terms and statement, (2) logical status of statements and proof framework, and (3) justification of claims (Mejía-Ramos, et al., 2012). It is important these three are referenced in terms of understanding only, as all three of them qualitatively feel similar to ideas concerning validity. Holistic proof comprehension is more concerned with understanding the proof as a whole or complete argument. This type of understanding focuses on (1) summarizing via high-level ideas, (2) identifying modular structure, (3) transferring general ideas or methods to another context, and (4) illustrating with examples (Mejía-Ramos, et al., 2012).

Assessments

An educational assessment is a formalized tool for observing, documenting, and quantifying a particular phenomenon (see, Pellegrino, Chudowsky, & Glaser, 2001), but

so rarely is an assessment a straightforward measure of said phenomenon. To this point Mislevy, Steinberg and Almond (2003) said that,

> In assessment, the data are the particular things students say, do, or create in a handful of particular situations, such as essays, diagrams, marks on answer sheets, oral presentations, and utterances in a conversation. Usually our interest lies not so much in these particulars but in the clues they hold about what students know or can do as cast in more general terms. (p. 9)

The implication here is that the data collected from an assessment means more than the individual answers and says more about the individual's knowledge or ability than the individual tasks.

Test theory. The set of responses which are gathered to evaluate the effectiveness of an assessment must be analyzed to determine the veracity of the assessments capability to meaningfully make inferences about the test taker on a given subject. Classical test theory (CTT) and item response theory (IRT) are the two main forms of analysis which inform this evaluative effort for assessment creation.

Where CTT assumes a simple linear relationship between tests score and measurement of tacit trait, IRT assumes a more dynamic two-parameter logistic relationship (see Figure 7) for each item of an assessment. Thus, IRT is a more nuanced approach to assessment analysis which looks at the individual items to infer the measurement capability of the entire test. With the rise of modern computing in the last 40-years, IRT became the more common approach as it is often considered the superior of the two approaches (Embretson & Reise, 2000). Thus, while CTT would be a simpler

approach using linear modeling, IRT is preferred assuming the amount of data needed for analysis can be generated.

$$P(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

Figure 7. IRT employs a two-parameter logistic model where for any dichotomous item, i, the probability of a correct response, $P(\theta)$, based upon the ability, θ, defined by the item's discrimination (a_i) and difficulty (b_i).

In practice, when analyzing an assessment using IRT, characteristic curves are generated for each dichotomous item based upon the item's difficulty and discrimination. These characteristic curves are generated using the probability function in Figure 7 and represent the likelihood that an individual with ability θ will get the correct answer for that item. Baker (2001) defines difficulty as a location index for each item on the ability spectrum, where items with high difficult – the green curve in Figure 8 – are associated with high ability and low difficulty – the blue curve in Figure 8 – with low ability. The location index of difficulty is defined in terms of ability where the probability of getting a correct response is .5 for any one item (Baker, 2001).

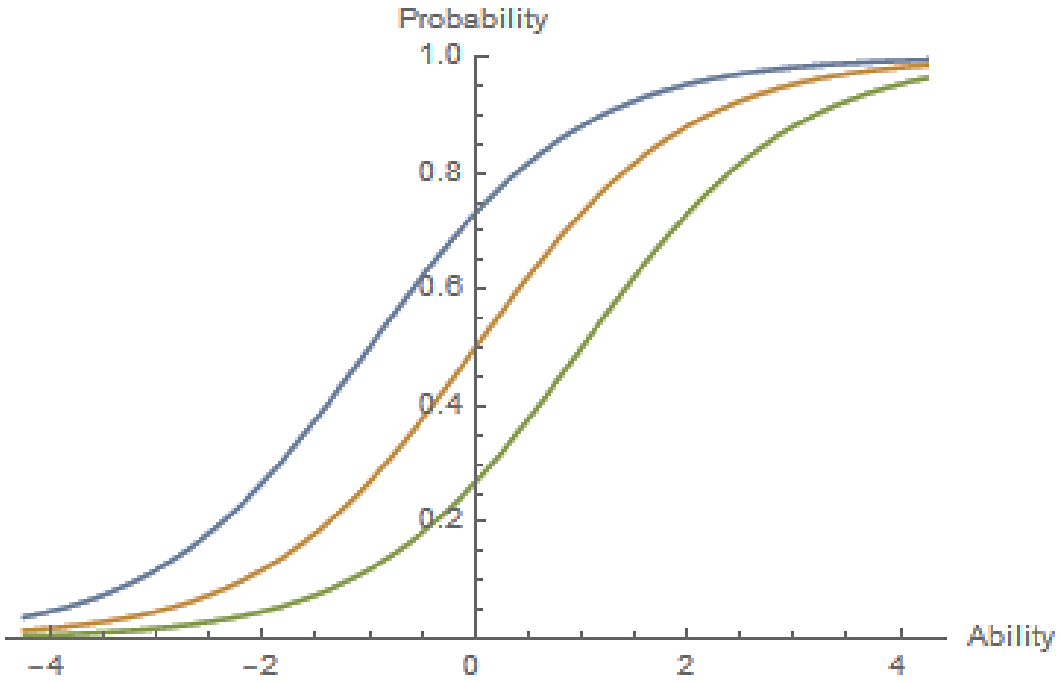


Figure 8. Three characteristic curves with identical discriminations, but differing difficulties, the blue being the least difficult and the green being the most difficult. Adapted from Baker's (2001) The Basics of Items Response Theory.

Discrimination is a measure of an item's capability to distinguish between individuals with varying ability. Items with higher discriminations – the green curve in Figure 9 – are better at distinguishing between individuals with small differences in ability, whereas items with low discrimination – the blue curve in Figure 9 – do not indicate as much information about the difference in ability between individuals. Discrimination is related to the slope of the characteristic curve at $\theta = b$ for an item and as the item discriminates more, the slope is steeper (Baker, 2001). As a note, items can have a negative discrimination just as items can have negative difficulty. While the latter case is acceptable, as 0 is considered the mean ability, the former case of negative discriminations is unwanted and infers that something went wrong with that item. Such items need to be reviewed and adjustments need to be made to it or the item needs to be thrown out altogether.

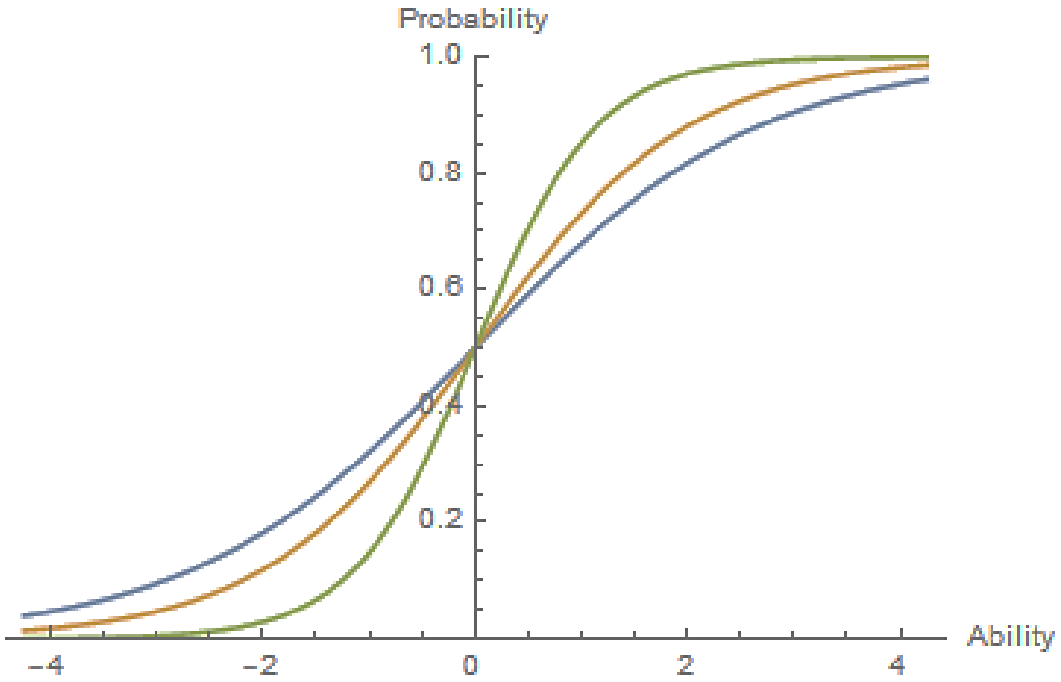


Figure 9. Three characteristic curves with identical difficulties, but differing discrimination, the blue being the least discriminatory and the green being the most discriminatory. Adapted from Baker's (2001) The Basics of Items Response Theory.

While it is true that IRT is considered the superior method of approach for analyzing assessments, the CTT measure of Cronbach's alpha is considered the standard score for test reliability amongst educational assessment (see Carlson et al., 2010; Hestenes et al.,1992; Melhuish, 2015). Values for Cronbach's alpha range from 0 to 1, and for an assessment like the one proposed for this study – a non-high-stakes assessment – an alpha score between 0.6 and 0.9 is acceptable (George and Mallery, 2003; Streiner, 2003). While scores outside the range from 0.6 to 0.9 are possible, scores below 0.6 are less desirable though between 0.5 and 0.6 are still somewhat acceptable and obviously score higher than 0.9 are considered outstanding.

Conclusion. Taking IRT and CTT into account as separate but important analysis tools in their own rights, I employ IRT as the main process by which the assessment is analyzed but use the CTT measure of Cronbach's alpha as a means of reporting the

assessment's reliability. I take the 0.6 baseline benchmark of Cronbach's alpha as the aim of this study. Additionally, the goal for the IRT analysis is having a diverse set of items which measure a range of abilities and internally differentiate between these abilities. The implication for the items of this assessment are that they should vary in their difficulty and have consistently high, positive discriminations. For the purposes of this study I define high discrimination as being greater than or equal to 0.5 as suggested by Baker (2001).

Reliability and Validity

Reliability and validity focus on two different but important aspects of an assessment. Reliability emphasizes an assessment's ability to perform consistently under repeated uses whereas validity underscores whether the assessment actually measures the construct in question. Each is an important aspect to consider in the creation of an assessment, for an instrument which is unreliable or invalid fails the measure of its creation: to accurately and consistently measure a specific tacit phenomenon.

IV. Methodology

The process for building a large scale student assessment of mathematical competencies for use at the university level has been outlined in the development of three distinct assessment (Carlson, Oehrtman, Engelke, 2010; Mejía-Ramos et al., 2018; Melhuish, 2015), all of which borrow from the work of Hestenes and colleagues on the Force Concept Inventory and Mechanics Baseline Test (Hestenes, Wells, Swackhammer, 1992; Hestenes, & Wells, 1992). The process that Hestenes and colleagues followed was further explicated in Lindell, Peak, and Foster's (2007) meta-analysis where they identify nine steps used in the design of instruments:

1. Identifying purpose
2. Determine the concept domain
3. Prepare test specifications
4. Construct initial pool of items
5. Have items reviewed – revise as necessary
6. Hold preliminary field testing of items – revise as necessary
7. Field test on large sample representation of the examinee population
8. Determine statistical properties of item scores – eliminate inappropriate items
9. Design and conduct reliability and validity studies (p. 15)

While there is some question about the ordering placed upon the construction of an assessment (see, Melhuish, 2015), Lindell et al. (2007) does present a sufficient blueprint with which to guide the work. As a clear purpose already exists, and the work of defining the concept domains already commenced, what follows is a discussion of the procedures observed in creating a multiple-choice instrument.

As with the work of Mejía-Ramos et al. (2017), the assessment of students' proof validation ability does not strictly fit the model of being a concept inventory. While concept inventories explore students' understanding of a broad set of concepts, the assessment aimed for in this study is, as yet, focused only on students' ability to validate

a subset of proofs – what I coarsely defined as direct proofs[11] – from a singular mathematical context ITP. While it is conceivable this study will lead to an assessment focused on validating many types of proofs from a variety of mathematical contexts (e.g., ITP, Algebra, Analysis, and Topology), it is not so at this point. Despite this difference, the mapping supplied by these concept inventories are still the most comprehensive and sensible solution as the long-term goal beyond this work would be just such an instrument. Again, as with the proof comprehension tests (Mejía-Ramos et al., 2017), this initial work will be critical in determining if a more comprehensive concept inventory is possible or even useful.

Identifying and Testing the Analytic Framework

It was vital to identify a robust analytic framing to guide the creation of the proposed instrument. The process taken to identify and test this framework, that of determining the concept domain (see Step 2 of Lindell, Peak, & Foster, 2007), was like what Messick (1995) referred to as a *Domain Analysis*. This effort, as adapted by Melhuish (2015), is a triangulation of knowledge. Adapted further for this study, I employed both existing literature, and experts' understanding to build a cohesive framing on which to build an instrument. In the end, this domain analysis is the basis for the argument for content validity of the assessment. The domain analysis began by thoroughly probing the literature followed by consulting mathematicians.

Literature. The process of identifying a conceptual framing started by plumbing literature focusing on commonly held beliefs or actions of students related to constructing proofs. This literature lead to the identification of possible domains for the *Common*

[11] For more on this definition see the write up in the theoretical background.

Validity Issues (CVI) framework. This framework focuses on issues in students' written proofs with the belief that these validity issues in written proofs can and do represent the understanding students have about argument validity whether they themselves are the author or they are reading a novel argument. The issues which arose from students' written proofs are genuine reflections of the cognitive structures which are pervasive in their own understanding of what is requisite for an argument to be valid. From the literature, general categorizations related to validity issues were identified which categorizations represented possible assessment domains.

The process for conducting a comprehensive exploration of literature focused on common validity issues started with searches employing both Google Scholar and ERIC of literature related to errors and misconception in student written proof construction. When pertinent literature was found, I employed Google Scholar to explore the set of more recent literature where each piece of literature from the original search was cited. Additionally, for each piece of literature from the searches, I explored the citations listed to gather more appropriate literature. In this way, I cast a web from each found piece forward and backward to be thorough in my inclusion of research studies which were germane to the topics at hand. This was useful, if not necessary, as the list of related literature is brief.

Experts. Research active mathematicians were consulted as a final step in identifying and testing the conceptual framework. Their efforts amounted to validating the conceptual framing and the analytic framing for the assessment, namely the CVI framework. This was accomplished in a two-step process of surveys and focus groups as outlined in the following section.

Creating the Assessment

The creation of the assessment itself followed a four-phase process used and outlined by Carlson, Oehrtman and Engelke (2010) and later adapted by Melhuish (2015). As phases 3 and 4 deal with refining and validating the multiple-choice assessment, I will omit any discussion on these and focus only on Phases 1 and 2. This choice is due to the scope of this work. Follow-up efforts will be needed to complete the assessment including working through refining and validating the assessment.

The steps of the process were described by Melhuish (2015), and while an open-ended survey is part of the phase 1 described here, much more effort has been put into this effort to build a set of distractors. To clarify the conversation, I use the term testlet to denote a theorem, its argument, and all the open-ended or multiple-choice questions dedicated to that theorem/argument pair. I define the terms stem, key, and distractor as they are the major parts of any multiple-choice question: the stem is a question, the key is the correct choice, and the distractors are the set of plausible but incorrect choices (Haladyna, 2004).

Phase 1 – Open-Ended Survey Development and Analysis

The first phase of creating a closed-form multiple-choice assessment was constructing an open-ended free-response survey. This process involved multiple steps, the first, and most important of which, was validating the framework and arguments, which are the focus for each item of the open-ended student survey. This validation process involved creating/collecting arguments and expert (i.e., active mathematicians) endorsement of said arguments in terms of validity. The process breaks into two stages; (1) item creation and mathematician survey; and (2) mathematician focus groups. After

completing the validation of the framework and arguments, next I constructed and analyzed the open-ended survey which process gave me a set of possible distractors for the closed assessment.

Item creation and mathematician survey. Qualtrics, an internet survey system, was the main resource for collecting and obtaining a large sampling of mathematicians' data. The survey consisted of 30 arguments to 22 different propositions considered germane in the ITP setting (David & Zazkis, 2017). The arguments were a collection of: (1) altered and unaltered student work collected from a pair of ITP courses offered at a large public university in the southern United States; (2) work collected from the internet site Mathematics Stack Exchange; or (3) altered and unaltered versions of proofs from the texts listed in Table 3, the three most common ITP texts in the US (David & Zazkis, 2017). Arguments were selected based upon their fit within the greater framing of the CVI framework and altered so each argument at most includes a single validity issue.

Table 3

Introduction to Proof Textbooks

Title	Publisher	Author(s)	Year
Mathematical Proofs: A Transition to Advanced Mathematics	Pearson Education	Chartrand, Polimeni, & Zhang	2013
A Transition to Advanced Mathematics	Brooks/Cole	Smith, Eggen, & St. Andre	2011
Book of Proof	Richard Hammack	Hammack	2013

For each argument, the participants of the mathematician survey were first asked (see Figure 10), "Is the argument for the included proposition a valid proof?" and given the binary option of "Yes - valid," or "No - invalid." Participants were initially warned against grading the proofs as though they were student proofs, but to instead answer for

themselves the question, “Does this argument actually prove the proposition in a way that I feel is appropriate, based upon what I believe is requisite for an argument to be valid?” In this way, it was left to the participant to infer what they felt was requisite for an argument to be a valid proof while dually attempting to move them away from practices that more closely approximated what Moore (2016) was looking at when he explored mathematics instructors’ grading habits.

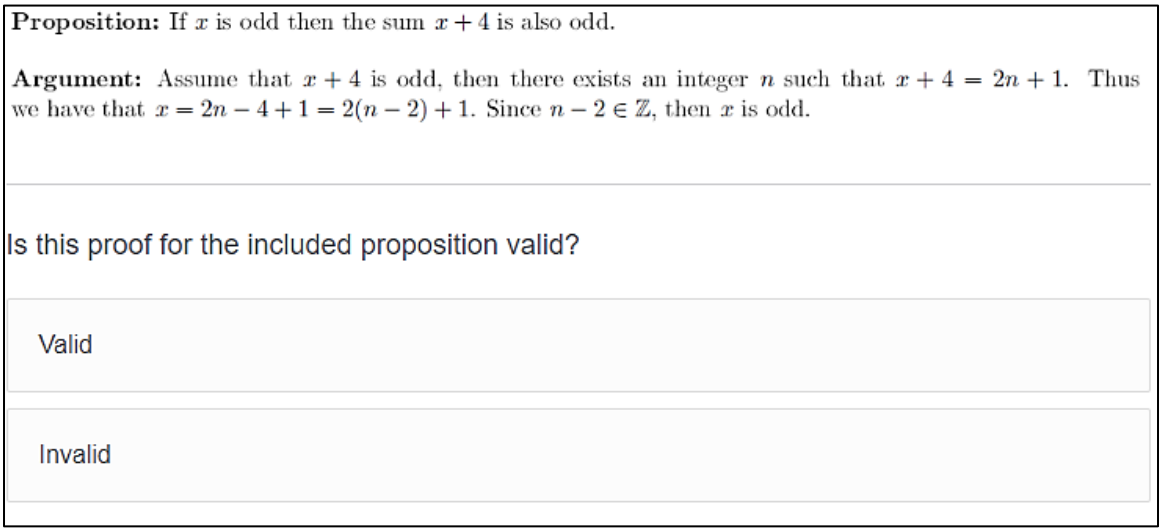


Figure 10. First question asked for each argument

If the proof was initially coded as invalid and the participant disagreed (i.e., they chose “valid” as their response), the participant was presented (see Figure 11) with the proposed validity issue and asked how the presence of said possible issue affected their initial response, and then given the chance to change their minds about the validity of the argument. If the participant did not change their mind, they were asked to share why they felt the possible issue did not invalidate the argument. Additionally, for each argument that was initially coded as invalid, if the participant agreed that it was in fact an invalid argument, they were also presented with the possible issue and asked if it was the reason they choose invalid (see Figure 12). If it was not the reason, participants were asked to

state why they thought the argument was invalid. For all arguments which were initially coded as valid, if the participant disagreed and chose invalid, they were asked to justify their views by stating why they thought the argument was invalid.

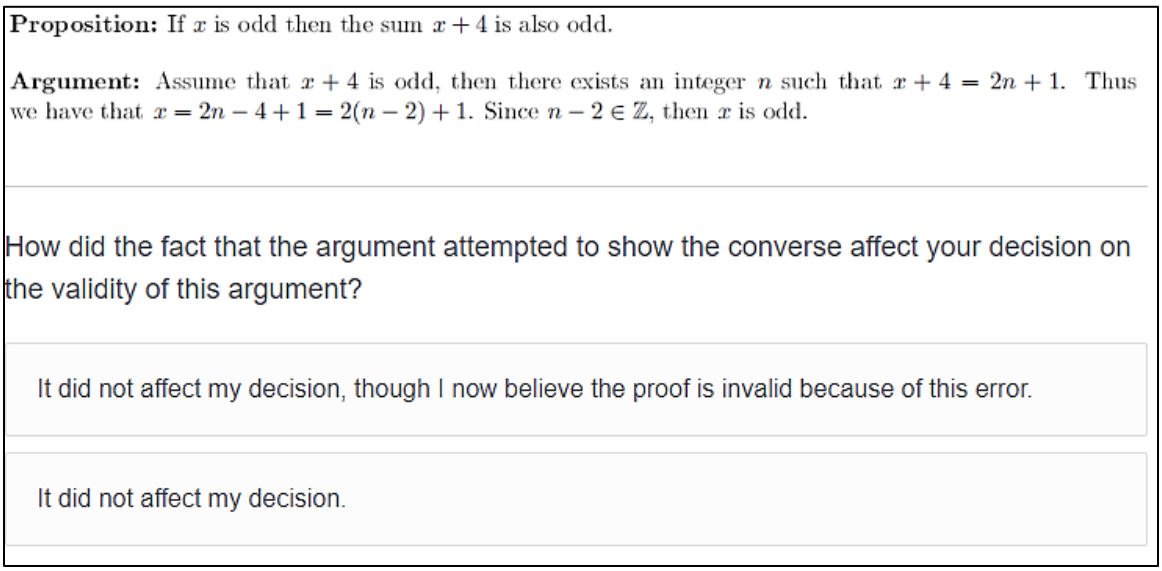


Figure 11. If the mathematician chose valid for an argument coded as invalid, they were asked how the presence of a validity issue affected their decision.

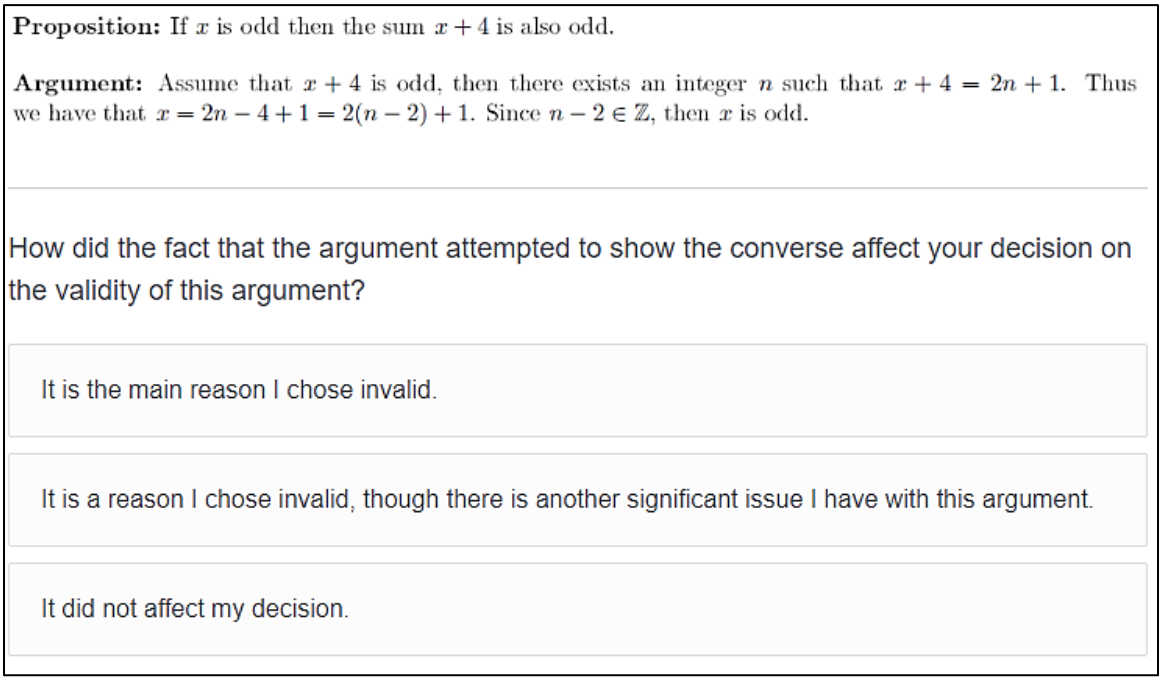


Figure 12. If the mathematician chose invalid for an argument coded as invalid, they were asked how the presence of a validity issue affected their decision.

The arguments themselves were clustered into one of seven groupings based upon their initial validity coding and issue. Participants were then randomly presented with an argument from each cluster to ensure they saw an argument whose validity issue came from each area of the framework. They were also randomly presented with an argument which was initially considered to be valid. No participant saw the same argument twice, and for propositions which had multiple arguments in the survey, no mathematicians saw more than one argument per proposition. In total, 1528 survey invitations were distributed via email to research-active mathematicians across the United States, of which 228 submitted responses to the survey[12]. Of the 228 participants, 178 completed all 7 argument sets with which they were presented; all others completed no less than 2 argument sets.

Along with the arguments, mathematicians were asked to fill out some basic demographic information concerning their backgrounds in mathematics, their experience teaching ITP-like courses, as well as a semi-open question focusing on what necessarily invalidates an argument. The demographic information included: (1) their current institution; (2) the number of years at said institution; (3) their rank (e.g., lecturer, assistant-, associate-, full-professor) in their department; (4) the highest level of degree they had obtained; (5) the institution they received said degree; (6) how many times they taught an ITP course; and (7) their area of expertise within mathematics (e.g., algebra, calculus/analysis, etc.).

For the analysis of this mathematician survey, all free responses were analyzed using thematic analysis (Braun & Clarke, 2006). The analysis began with open coding of

[12] This represents a 14.9% response rate.

the free responses for each of the 30 arguments independently and categorizing responses relative to each argument in terms of their appropriateness. All nonsensical free-responses led to a cycle of analysis of the quantitative data supplied by the author of said free-response to ensure the author was not supplying false or unintended data. This sort of data was omitted from further analysis. Following open coding, themes were identified, categorized and condensed for each argument. No cross-argument analysis occurred as the questions for this study do not focus on how responses to one type of validity issue are correlated to responses to other validity issues.

Each item was quantitatively explored to see if a consensus was reached on the point of validity for the arguments. For each invalid argument, three simple statistics were calculated (see Table 4): (1) initial percent invalid (IPI), which measures the percentage of respondents who initially thought the argument was invalid; (2) final percent invalid (FPI), which represents the number of individuals who either initially thought the argument was invalid or changed their response to invalid after being confronted with a possible flaw; and (3) percent agree (PA) which takes into account both the number of respondents who chose invalid – initially or after they were confronted with a possible flaw – and their reason for choosing "No-Invalid" or changing their minds. This last statistic, PA, was important as it said something about how the CVI framework was performing. As each argument had a singular intended validity issue, PA said whether the intended validity issue was the main reason, or even a reason at all that respondents chose invalid in the end. For valid proofs, the only calculation was akin to PA, valid percent agree (VPA). This was a straightforward measure of the percent of individual respondents who thought the argument was valid.

Table 4

Quantitative measures of the mathematician survey

Coded	Statistic	Definition	Calculation
Invalid	Initial percent invalid (IPI)	Percentage of respondents that initially thought the argument was invalid.	$\frac{\#\ of\ "No-Invalid"\ responses}{total\ \#\ of\ responses}$
	Final percent invalid (FPI)	Percentage of respondents that, both initially and in the end thought the argument was invalid.	$\frac{\#\ of\ "No-Invalid"\ responses + \#\ of\ changed\ responses}{total\ \#\ of\ responses}$
	Percent agree (PA)	Percentage of respondents whose reason for choosing invalid – both initially and after seeing the proposed flaw – agreed with the intended validity issue for that argument.	$\frac{\#\ of\ "No-Invalid"\ reason\ agreed + \#\ of\ changed\ responses}{total\ \#\ of\ responses}$
Valid	Valid percent agree (VPA)	Percent of individual respondents who thought the argument was valid.	$\frac{\#\ of\ "Yes-Valid"\ responses}{totaly\ \#\ of\ responses}$

Mathematician focus groups. After the mathematician survey completed, a series of two audio recorded focus groups were conducted. These focus group were aimed at extending the work started in the survey; that of further clarifying the idea of validity. Focus groups were employed as part of this research as they are considered appropriate in their ability to "elicit people's understandings, opinions and views, or to explore how these are advanced, elaborated and negotiated in a social context" (Wilkinson, 1998, p. 187). As proofs are generally accepted as a form of social communication defined by specific contexts (e.g., Stylianides, 2007) the use of focus groups to determine the opinions and views of mathematicians about what is requisite for a proof to be valid is apt according to Wilkinson's (1998) view. Because the goal of this research is to zero in on features common to valid proofs, it is important to identify areas

of agreement for mathematicians on these features. Wilkinson (1998) argues that due to the interactive nature of focus groups, they lend themselves to clarifying these areas of agreement and disagreement far better than individual interviews.

Two groups of known survey participants were gathered. In order to overcome the possibility of what Smithson (2000) referred to as dominant voices, the participants for each focus group were selected based upon their experience and establishment in their individual fields within mathematics. By design and selection, all participants had taught advanced-level undergraduate mathematics at least three times in their careers, with most having established a record of at least a decade of teaching these types of courses.

The first focus group consisted of four participants (see Table 5), all of whom were established associate and full professors at a large university in the southern United States. These individuals had a variety of ethnic backgrounds as well as mathematical backgrounds. From a mathematics standpoint, the participants foci were bifurcation theory and differential equations, geometric and algebraic topology, group theory, and topology.

Table 5

Focus group 1 participants

Participant	Title	Mathematical Expertise	Years Teaching
Tyler	Professor	Differential equations	29
Jeremy	Associate Professor	Topology	29
Jon	Professor	Group theory	20
James	Professor	Topology	40+

The second focus group was a larger group of seven participants from a different university in the southern United States (see Table 6). The level of teaching expertise varied more in this group. Much like the first group, these individuals had a variety of ethnic backgrounds as well as mathematical backgrounds. Their mathematical foci included mathematics education, functional analysis, non-associative algebra, geometric group theory and analytic number theory.

Table 6

Focus group 2 participants

Participant	Title	Mathematical Expertise	Years Teaching
Trevor	Associate Professor	Mathematics education	13
Taylor	Associate Professor	Analytic number theory	16
Travis	Associate Professor	Mathematics education	32
Justin	Assistant Professor	Geometric group theory	8
Joseph	Assistant Professor	Functional analysis	12
Beth	Assistant Professor	Mathematics education	4
Tim	Professor	Non-associative algebra	39

Prior to the focus groups themselves, a series of general questions and prompts were organized into a slideshow to homogenize the two groups and to organize the content of the discussions. Because there was a space of nine months between the focus groups, I went back to the audio recording of the first focus group and refamiliarized myself with the questions and the content in order to help the two groups cover the same overall topics. Though this refresh of ideas was conducted prior to the second focus

group, no deep analysis of the first focus group was conducted prior to the second focus group.

Each focus group was audio recorded and lasted around an hour. The audio recordings were transcribed and underwent an analysis which focused on communal consensus. This process involved multiple cycles of reading coding and analysis. I started first by reading each transcription through, chunking the conversations according to the prevalent topic with respect to the CVI framing. Next I took the chunked portions of each transcript and compared them categorically. This process consisted of thoroughly reading all chunks of dialogue which were from a singular CVI categorization. I noted portions of agreement and disagreement which the participating mathematicians shared about each categorization and the effect the mathematicians thought each had on validity. For this research, the focus groups were a consensus gathering, and thus the analysis needed to allow a focus on what consensus was gained through communal discourse. Despite the focus on consensus, it was also important to hear and recognize dissenting opinions. Voices of dissent were also captured in the analysis process to give the broadest overall understanding with respect to the mathematicians' thoughts about validity. Even though these dissenting voices are part of the analysis, in the end the focus remained on the communal consensus as the driving force for the analysis.

From this focus on communal consensus, the unit of analysis was the group itself rather than any distinct individual within the group. In the analysis phase, it is important to take into account the possibility of a dominant voice being representative as the group voice (Smithson, 2000). In the words of Smithson (2000), it is important that "The analytic focus is not on what individuals say in a group context but on the discourses

which are constructed within this group context" (p. 110). To overcome this issue, in the analysis phase I carefully selected and interpreted important series of discourse as a constructed understanding rather than as a stagnant artifact stated by an individual. Therefore, the reported data, though at times disjointed from the original transcription, are what I felt represented of not only what the group consensus was, but also how the group built said consensus.

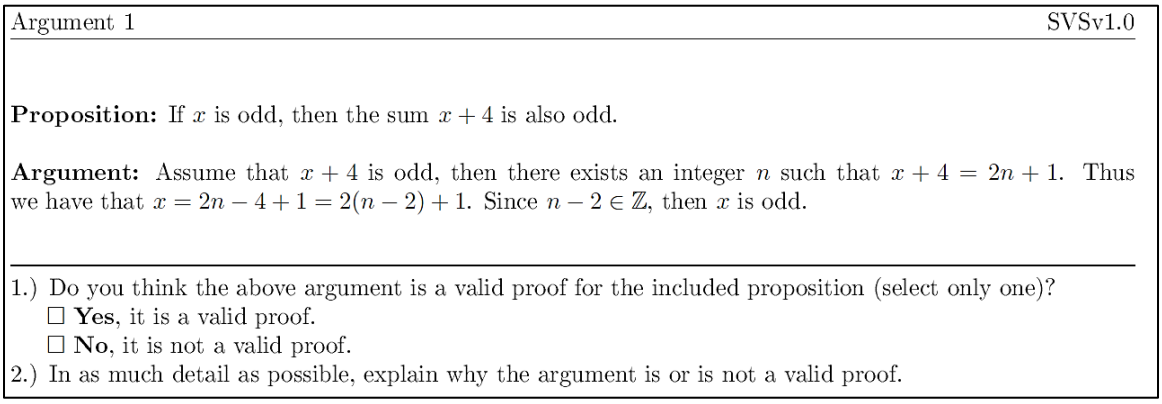
Open-ended student survey. Once the data from the mathematician survey were analyzed after the completion of the first focus group, an open-ended student survey was constructed. The propositions and arguments for this survey came from the mathematician survey and selected based upon the arguments PA or VPA. The twelve (12) arguments – ten invalid and two valid arguments – whose PA or VPA were at or above 90% were included in the open-ended student survey.

The open-ended student survey was a paper and pencil survey distributed to 69 students enrolled in 12 sections of 9 different advanced undergraduate mathematics courses at a large university in the southern United States. All students taking part in the open-ended survey were working on their bachelor's degree. Surveys were distributed in these 12 sections with prior permission from the class's instructor. The instructors could choose to give extra credit for completing the survey but needed to offer an alternative for students who opted out of participating in the survey.

The survey presented students with 12 proposition and argument pairs, asking them first if they thought the argument was a valid proof, and then second to explain their thinking (see Figure 13). Additionally, students were asked for mathematical demographic information including: (1) what advanced undergraduate mathematics

classes they were currently enrolled in; (2) what advanced undergraduate mathematics courses they had previously taken; (3) what degree they were pursing – done to pull out graduate students taking leveling courses from the analysis – and (4) what their degree concentration or major was. Just as with the mathematicians, students were advised against grading the proofs as though they were student proofs – in case this was a practice they were familiar with – but were to answer for themselves the question, "Does this argument actually prove the proposition in a way that I feel is appropriate, based upon what I believe is requisite for an argument to be valid?"

The goal of the open-ended survey was not necessarily to act as a pilot for the actual assessment. Rather, the goal was to gather information about students' reasoning about validity. This information was used as a basis for constructing distractors in what became the semi-closed assessment later. As this was the case, only rudimentary analysis took place on the survey with a major focus placed upon the open responses.


Argument 1 SVSv1.0

Proposition: If x is odd, then the sum $x + 4$ is also odd.

Argument: Assume that $x + 4$ is odd, then there exists an integer n such that $x + 4 = 2n + 1$. Thus we have that $x = 2n - 4 + 1 = 2(n - 2) + 1$. Since $n - 2 \in \mathbb{Z}$, then x is odd.

1.) Do you think the above argument is a valid proof for the included proposition (select only one)?
☐ **Yes**, it is a valid proof.
☐ **No**, it is not a valid proof.
2.) In as much detail as possible, explain why the argument is or is not a valid proof.

Figure 13. The open-ended student survey consisted of 12 proposition and argument pairs like this one where students assessed the validity of the argument and explained their thinking.

The open responses were analyzed using thematic analysis (Braun & Clarke, 2006). This analysis began with open coding of the open responses for each of the 12 arguments independently and categorizing responses relative to each argument in terms of their appropriateness. All nonsensical open-responses led to a cycle of analysis of the

respondent's closed-question responses and other open-question responses to ensure the respondent was not supplying inconsistent data. Inconsistent data was omitted from further analysis. Following open coding, a set of counts were constructed for each code. From these counts and through a review of the individual codes, themes were identified, categorized and condensed for each argument. From these condensed categorizations, along with the overall counts, distractors for the semi-close assessment were selected and constructed.

Phase 2 - Semi-Closed Assessment Pilot

The end product for this study is a closed assessment, but for the pilot I left the ability for open data collection to refine the distractors. In order to construct the semi-closed assessment pilot (SCAP), simple frequencies were computed for each of the 12 arguments to determine which items students struggled with most. This information, along with any non-mathematical comments made about the arguments (e.g., I am not sure, I am just guessing, I do not remember/know this concept, etc.) identified candidates to cut or adjust before including arguments in the SCAP. Before an argument was cut, its representative CVI category was considered. None of the categories which had a single representative argument in the open-ended survey were cut, though appropriate small adjustments were made to clarify content not directly related to validation.

The SCAP included 11 arguments to 10 propositions[13], each student seeing 8 arguments (see Table 7). Of the 11 arguments, two were valid with all others containing a validity issue from the CVI framework. Only one argument contained more than one

[13] The proposition for *Arg4 long* concerned showing that a specific relation, R was an equivalence relation, thus requiring an argument that shows reflexivity, symmetry, and transitivity. The proposition for *Arg4 short* only required that the same relation R be transitive thus shortening the accompanying argument.

validity issue – *Arg6* had a pair of validity issues. In total, there were five anchored arguments which all students saw. These anchor items included both valid arguments and three other invalid arguments. For each of the randomized pairs, Random (1) – Random (3), students only saw one argument from each grouping. For instance, if a student was randomly assigned *Arg4 short*, they would not also see *Arg4 long*, and so on.

Table 7

SCAP arguments and validity issue from the CVI

Item	Validity	Validity Issue	Anchor/Random (Grouping)
Arg1	Invalid	CR	Anchor (1)
Arg2	Invalid	AC	Anchor (1)
Arg3	Valid	N/A	Anchor (1)
Arg4 short	Invalid	MN	Random (1)
Arg4 long	Invalid	MN	Random (1)
Arg5	Valid	N/A	Anchor (2)
Arg6	Invalid	LG & W	Anchor (2)
Arg7	Invalid	WT	Random (2)
Arg8	Invalid	WT	Random (2)
Arg9	Invalid	W	Random (3)
Arg10	Invalid	W	Random (3)

The structure of the SCAP was a nontrivial task to tackle as asking if an argument is valid or invalid is akin to a coin toss and would not reflect what students actually know. To overcome this, each proposition and argument pair asked a set of questions for students to answer which set I refer to as a testlet. Each testlet utilized an intricate

structure (see Figure 14) affording students the opportunity to validate and say why they made a particular validity choice without trivializing the assessment in the process. Additionally, the implemented structure tests the hypothesis that students can be led to correctly validate an argument through external prompts which lead to personal reflection and reconsideration (see Selden & Selden, 2003). The testlets focusing on each argument and proposition pair were presented to students in a semi-random order based upon their group. Testlets for the grouping Anchor (1) were presented in a random order followed by a Random (1) testlet then Anchor (2) testlets were presented in a random order and then the Random (2) and (3) testlets. In this way, each student took one of eight possible test forms.

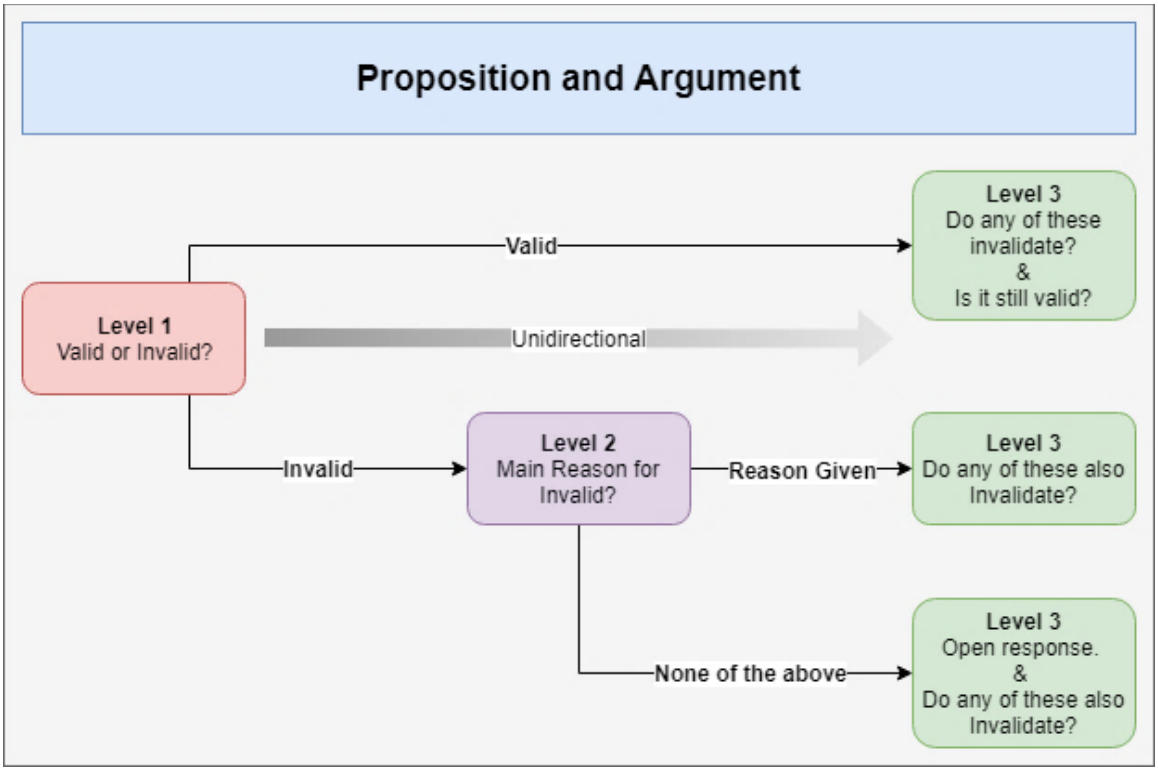


Figure 14. The structure of each testlet was designed to make the validating task non-trivial and to allow students to change their mind about the validity of an argument.

Each testlet had a structure identical to the one in Figure 14 where students were led from Level 1 to Level 3 without the ability to go back and change their answers.

Level 1 consisted of a single question: is the argument a valid proof for the included proposition (see. Figure 15). From there, students advanced depending upon their response. If students chose invalid at Level 1, they were presented with a Level 2 question, and then a Level 3 question. If at Level 1 they chose valid, they would proceed directly to Level 3 questions.

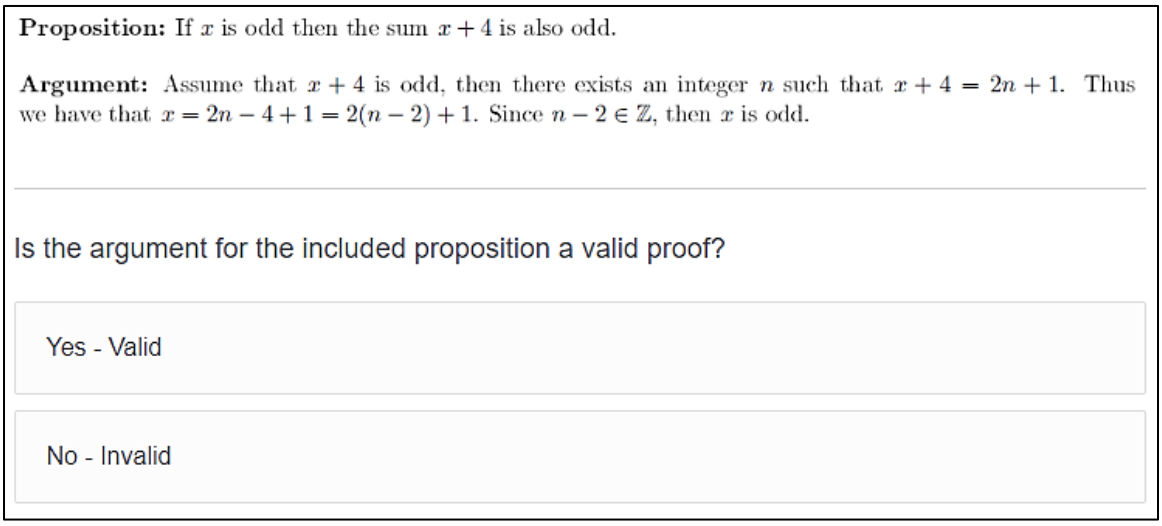


Figure 15. Students were first asked Level 1 questions which asked if the argument was a valid proof.

The key and distractors for each testlet were consistent throughout the testlet at Level 2 and Level 3. This means for Level 2 and Level 3, students were always presented with the same set of options in Level 2 and Level 3 regardless of their response in Level 1. It also meant every student who was presented with a particular testlet, their set of possible answers in Level 2 and prompts in Level 3 were the same as every other student, thus allowing for sensible analysis afterward.

The testlets for *Arg2* and *Arg6* had two keys (i.e., two correct answers). In the case of *Arg2* (see Figure 15) the idea of assuming the conclusion meant the argument was for the converse of the proposition to be proven. To combat the possibility students would learn the test, the idea was broken into the effectual meanings of *beginning with*

the consequent and *proving the antecedent*, instead of directly presenting students with *the argument was for the converse*.

However, Arg6 was a novel instance where two separate and distinct validity issues arose, in which a logical gap (LG) led to a warranting (W) issues (see Figure 16). Instead of removing one or both validity issues after the mathematician survey, the dual issues persisted to see how students handled the dual cases. Because the testlet for *Arg2* is always prior to the testlet for *Arg6*, students were already confronted with a testlet with two keys making it a smaller assumption that students would willingly signify that two validity issues existed within a singular argument.

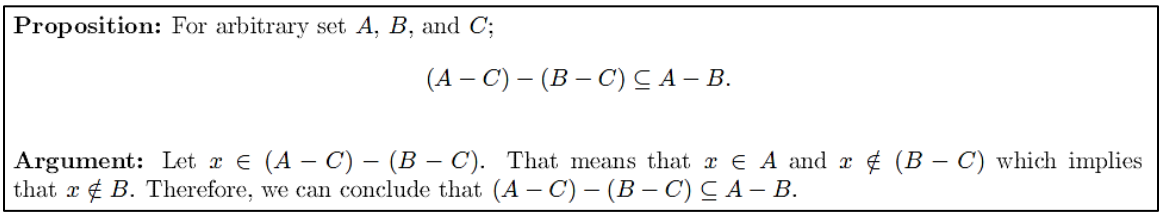


Figure 16. The argument skips the assertion (LG) that $x \in (A - C)$ which leads to the dual understanding that $x \in A$ and $x \notin C$. The latter is important for justifying why $x \notin B$ (W).

Level 2 consisted of a single-answer, multiple-choice question asking students for the main reason they chose invalid in Level 1 (see Figure 17). For each testlet, Level 2 questions included a set of 3-5 distractors compiled from Phase 1, with the final option being "none of the above." This last option was an important affordance as the SCAP was intended as a pilot, so it was entirely possible that new understandings could emerge throughout the piloting process (see Melhuish, 2015). For arguments which were valid proofs, there was no key in Level 2, thus all options were distractors.

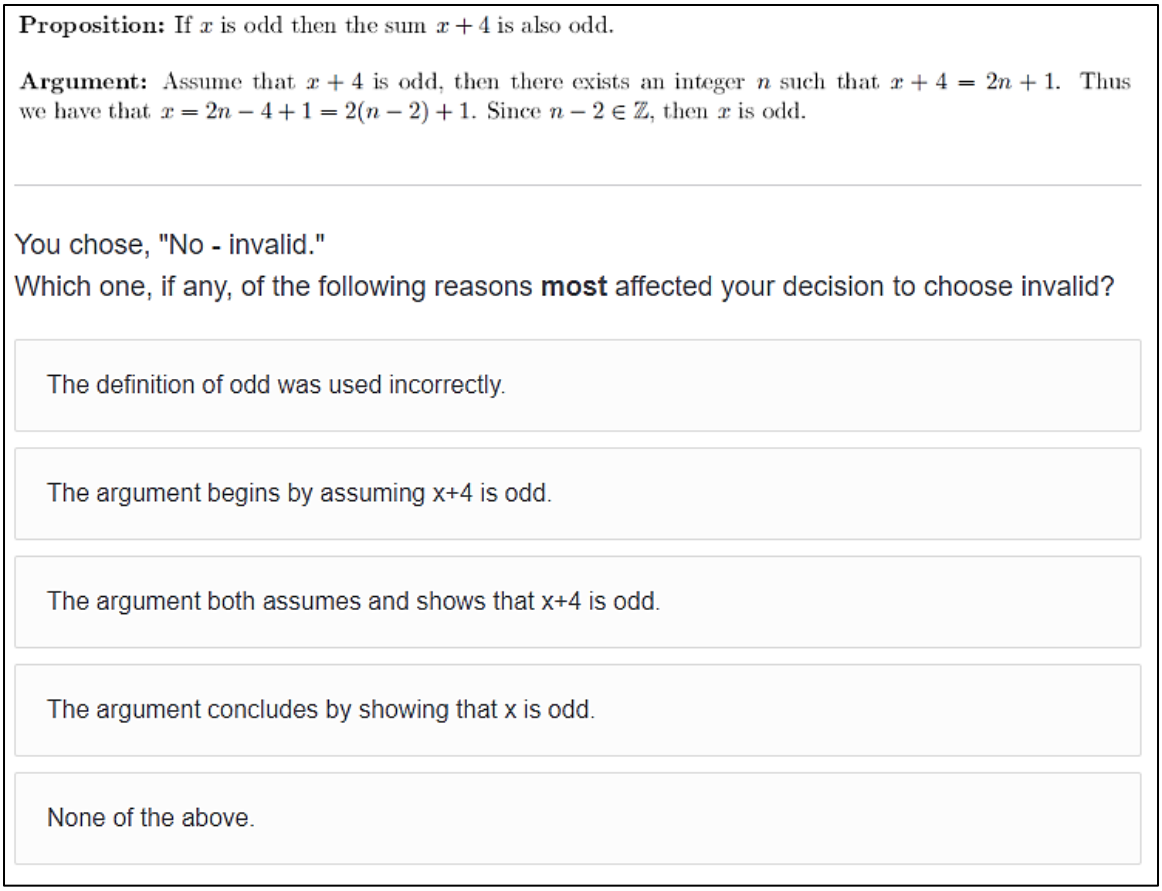


Figure 17. A Level 2 question was asked in the case a student selected "No-Invalid" for the Level 1 question. This question was a single-answer, multiple-choice question including none of the above to allow additional input from students on possible reasons for invalidity.

Level 3 included a matrix style question (see Figure 18) and in two cases had an additional question depending upon which Level 3 block the student was filtered into from prior questions (see Figure 19). The example in Figure 18 is the case where the student selected valid for the Level 1 question and then Level 3 included a dichotomous follow-up question asking if the student still thought that the argument was valid. In the case where students initially selected invalid to the Level 1 question and in Level 2 selected "none of the above," students were given the open prompt, "In as much detail as possible, please explain why you feel the argument is not valid." This open prompt afforded new understanding into student thinking, but also gave students the opportunity to change their minds and claim the argument was actually valid. If a student gave a

specific reason at Level 2 other than "none of the above," they were only given a matrix question in Level 3 which did *not* including the option they selected as the main reason the argument was invalid in Level 2.

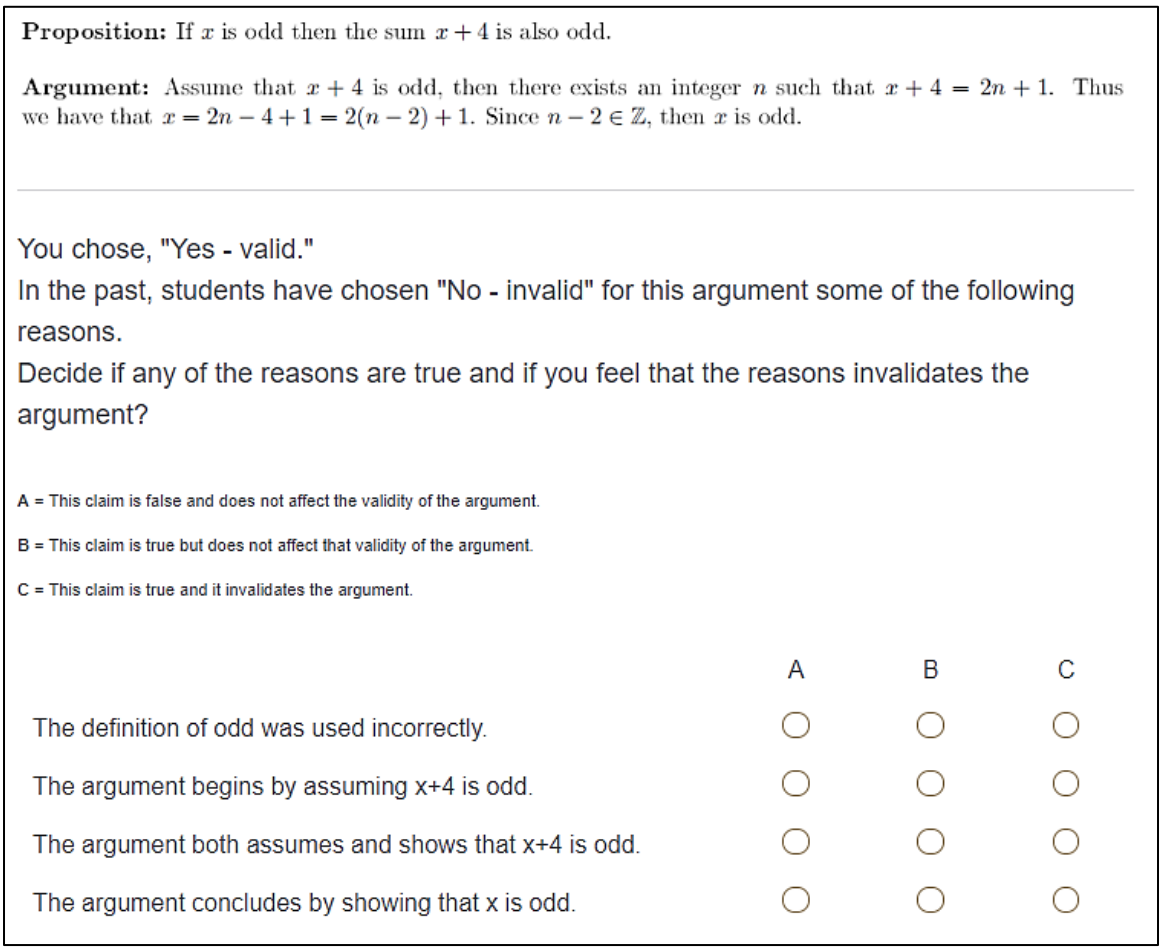


Figure 18. Level 3 always presented students with the opportunity to "grade" the set of distractors where (A) meant they thought the claim was false and had no bearing on validity, (B) meant the claim was true but had no bearing on validity, and (C) meant the claim was true and necessarily invalidated the argument.

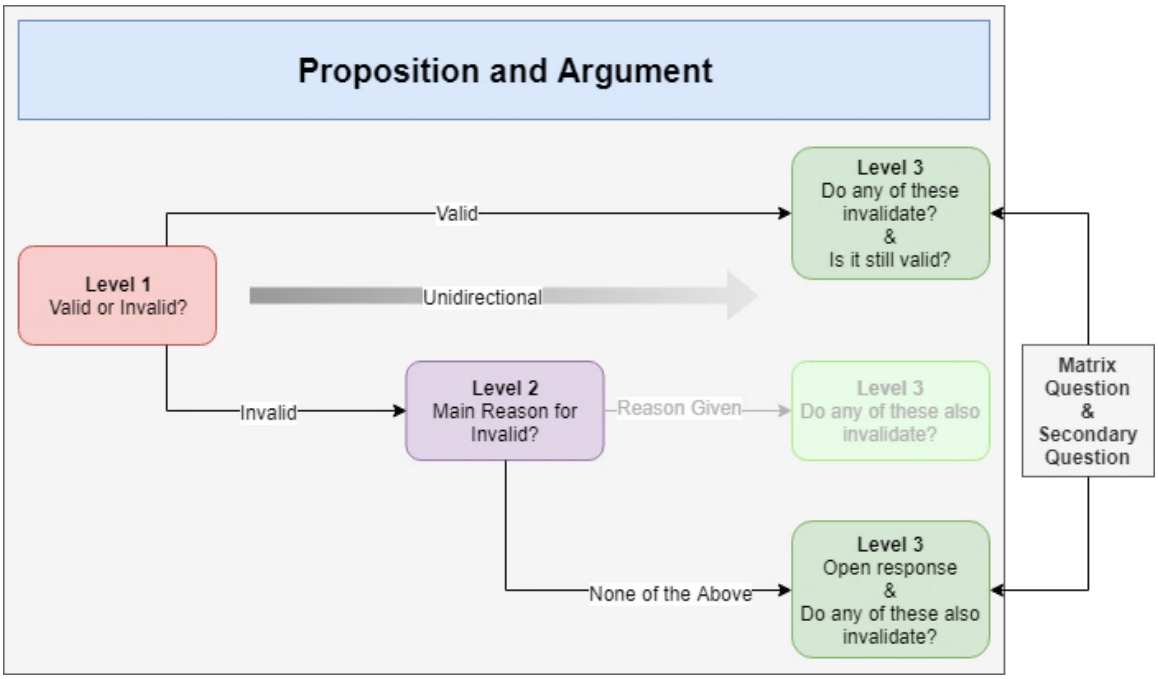


Figure 19. In two of the three Level 3 blocks, students were asked both a matrix style question and one other question. For those in the "valid" Level 3 they were asked if they still thought the argument was valid. For those in the "none of the above" Level 3 they were asked to give the main reason the argument was invalid.

Beyond the validity testlets, students were asked questions concerning their mathematical backgrounds. This information included the number of proof-based mathematics courses previously taken and currently enrolled in at the time of the assessment. Students were then asked if they had taken a class in number theory, analysis/real analysis, abstract algebra/group theory, topology, or an ITP course[14]. If a student answered zero to both of the first two question (i.e., they had not taken nor were they currently enrolled in a proof-based course) and they said that they had not taken any of the specifically listed courses, they were taken to the end of survey and thanked for their time. Information was also collected concerning students' GPA, the degree they

[14] The acronym ITP was not used, instead students were asked, "Have you taken or are you currently enrolled in a course where, beyond the mathematical content itself, the main point of the class was to introduce you to mathematical proofs? This course may have focused on things like logic; proof techniques and types (e.g., direct proof, proof by induction, etc.); and mathematical content including sets, relations, functions, and cardinality of sets."

were working toward, and their major and minor concentrations. For students working toward their bachelor's degree, they were given the opportunity to opt into a single, 1-hour interview about the assessment. Finally, students were asked to estimate the time it took them to complete the assessment and thanked for their time.

The target audience for the SCAP was any undergraduate student who had previously taken or was currently enrolled in an ITP at the time of the assessment. Students who indicated they had not taken an ITP course, but had taken some proof-based undergraduate course were included in the data collection.

At the beginning of November 2018, instructors were contacted via email from universities across the U.S. Universities were selected for participation based upon their inclusion in the AMS, *Directory of Institution in the Mathematical Sciences*[15]. Instructors were given a brief description of the assessment and asked if they were willing to have their students participate in data collection. In exchange for instructors taking part, I offered to give a detailed breakdown of how their classes faired on the assessment[16] against the national average. They were additionally told if they offered extra-credit for participation, they would need to offer an alternative form of extra credit so students who did not want to participate would have an opportunity to earn extra credit.

Included in the recruitment email was a PDF printout for their students which contained the same basic information given to the instructor as well as a class code which would sort their students' data according to their instructor. Instructors were told not to

[15] http://www.ams.org/profession/dirinst/dirinst-index.html
[16] Instructors were informed the data they received would not include individual data but would give them an overall picture of their class's performance.

use the class code if they wanted to view the assessment themselves, but instead use the code *TEST* so their data could later be removed from analysis.

Analysis of SCAP. All told, 186 students from 23 universities across the U.S. took the SCAP. The first step in analyzing the data from the SCAP was to look at the open responses for the invalid arguments and compare them against the key for each testlet. This process was done by simple comparison of the written responses and the actual reason for which the arguments were invalid. Open responses which matched the key were given a score of 1 and all other open responses were given a score of 0. Similarly, for the valid arguments, the open responses were analyzed, looking for a change of mind where they initially thought the argument was valid despite the fact that they had chosen invalid at Level 1. Open responses which changed back to valid were given a score of 1 and all other open responses were given a score of 0.

After scoring the open responses, the next step was to come up with a sensible way to score each testlet (see Table 8). For the testlets with valid arguments (i.e., *Arg3* and *Arg5*) responses were split into three cases; (1) correct, (2) corrected, and (3) incorrect. Responses scored as correct answered Level 1 and Level 3 correctly and were given a score of one (1). This means they correctly identified the argument was valid initially, did *not* indicated a C – meaning they did not think the prompt invalidated the argument – for any of the prompts to the matrix question, and did not change their mind on the validity in the final question. Valid testlets scored as corrected initially identified arguments as invalid in Level 1, but in Level 2 selected "none of the above," and in the open prompt in Level 3 suggested that they changed their mind while not selecting C for

any of the matrix prompts. Valid corrected testlets were given a score of zero (0). All other responses to valid testlets were given a score of negative one (-1).

Table 8

Scoring for the SCAP

Testlet Validity	Score		Meaning	Practical Meaning
	Original	LTM		
Valid	1	1	Correct	• Level 1 – Valid • Level 3 – No "C" in matrix and did *not* change mind.
	0	0	Corrected	• Level 1 – Invalid • Level 3 – No "C" in matrix and in the open prompt stated that they changed their mind.
	-1	0	Incorrect	• All other responses
Invalid	1	1	Correct	• Level 1 – Invalid • Level 2 – Identify Key • Level 3 – Correctly scored matrix prompts. *or* • Level 1 – Invalid • Level 2 – None of the above • Level 3 – State equivalent reason to the key, and correctly scored matric prompts.
	0	0	Corrected	• Level 1 – Valid • Level 3 – Correctly scored matrix prompts, and change to invalid.
	-1	0	Incorrect	• All other responses

The testlets with invalid arguments (i.e., *Arg1*, *Arg2*, *Arg4 short* and *long*, and *Arg6–Arg10*) responses were similarly split into three cases; (1) correct, (2) corrected, and (3) incorrect. Reponses scored as a correct did one of two things. If they correctly answered the Level 1 question as invalid, identified the key in Level 2, and then correctly sorted the prompts for the matrix in Level 3 as not invalidating the argument, except in

the case of *Arg2* and *Arg6* where they needed to correctly identify the second key as a C, then it was marked correct. The other correct scoring response was if they correctly answered the Level 1 question as invalid, but at Level 2 selected "none of the above," but in Level 3 their open answer was equivalent to the correct answer and they correctly sorted the prompts for the matrix in Level 3 as not invalidating the argument except in the case of *Arg2* and *Arg6* as previously mentioned. Responses scored as correct were given a numeric score of one (1).

Responses scored as "corrected" started by stating the argument was valid in Level 1, but in Level 3 correctly identified the key as C and sorted the other prompts for the matrix in Level 3 as not invalidating the argument except in the case of *Arg2* and *Arg6*, and then changed their validity evaluation to invalid in the final Level 3 question. These individuals were those who were led to correctly validate the argument as was suggested by Selden and Selden (2003). These responses were given a numeric score of zero (0). All other responses that were not correct or corrected for invalid testlets were given a score of incorrect with a numeric score of negative one (-1). This way of scoring meant that the overall range for total scores was from negative eight to eight, $[-8,8]$.

Once scoring was completed, to get a general understanding of reliability, the CTT measure of tau-equivalent reliability estimate, known commonly as Cronbach's Alpha, was calculated for the assessment. As Streiner (2003) pointed out, a large Alpha is always better, but as this assessment is not a high-stakes test, it could be better described as a research tool in the early stages. As such, Streiner (2003) suggests a score between 0.7 and 0.9 is acceptable. This does not completely discount scores below 0.7, as George

and Mallery (2003) suggest scores between 0.6 and 0.7 are still acceptable. This was just a first step into understanding how the assessment was performing and was done without considering the anchored aspect of the assessment.

Next the set of demographic information was used to see if test scores were correlated with known research about undergraduate students in proof-based mathematics courses. Ideally, there should be a positive correlation between GPA and assessment scores, as well as a positive correlation between the number of courses taken or enrolled in and assessment scores. Students who were pure mathematics majors should also have outperformed all other students with respect to final scores. Three tests were performed to test these hypotheses; (1) two sample t-test were performed on dichotomous data, (2) for discrete and categorical data[17] a one-way ANOVA was calculated, and (3) the bivariate Pearson' *r* correlation was calculated for continuous data including GPA. For establishing correlation for items like GPA, Pearson's *r* is both adequate and the typical approach for such instruments (Carlson, Oehrtman, & Engelke, 2010; Stone et al., 2003).

Small modifications were made to the scores in order to analyze the data when using the LTM package in the statistical software *R*. As discussed in the previous chapter, LTM is an IRT approach to quantitative data analysis which requires dichotomous non-negative data, typically 1's and 0's (Baker, 2001). To account for this, the scoring needed condensing and made non-negative for all data. To accomplish this, all scores which were not one (1) would be given a score of zero (0) (see Table 8). This transformation was only done to calculate the characteristic curves but was not a part of the CTT analysis, thus the original scoring has meaning going forward.

[17] For categorical data with more than 3 categories the bivariate Pearson correlation was calculated.

Additional analysis occurred at this level to take into account that the assessment was an anchored assessment. This was done in order to understand which forms (see Table 9) and testlets were preforming optimally and could be kept for subsequent future assessing. Once again, the set of condensed, dichotomous, non-negative data was used for this analysis which was done in *R* using the *equateIRT* package in conjunction with the LTM package. The eight forms were compared, and problematic items were closely analyzed for testing errors in Quatrics as well as coding and scoring errors in SPSS and *R* to ensure that the analysis was an accurate reflection of the data.

Table 9

The set of forms without anchored testlets[18]

Form	Randomized Testlets		
	1	2	3
F1	*Arg4 short*	*Arg7*	*Arg9*
F2	*Arg4 short*	*Arg7*	*Arg10*
F3	*Arg4 short*	*Arg7*	*Arg9*
F4	*Arg4 short*	*Arg8*	*Arg10*
F5	*Arg4 long*	*Arg8*	*Arg9*
F6	*Arg4 long*	*Arg7*	*Arg10*
F7	*Arg4 long*	*Arg8*	*Arg9*
F8	*Arg4 long*	*Arg8*	*Arg10*

Student interviews. To complete the data collection for phase 2, a series of one-hour student interviews proceeded the pilot run. The aim of these interviews were three fold: (1) to ascertain what students were attending to when they were taking the assessment (i.e., were they actually validating or were they doing something else), (2) to determine if there were items, distractors, concepts or other ideas which were worded in problematic ways or incomprehensible due to a lack of explicit explanation or lack of

[18] Anchored testlets were ones which included *Arg1-Arg3, Arg5* and *Arg6*. All forms had these five items which items were a part of the anchored analysis be were withheld from the table for brevity.

students' prior knowledge, and (3) to determine if students' scores could be accounted for by their ability to validate as opposed to other possible explanations.

Recruitment emails were sent to all students who had opted into participating in interviews. Once affirmative replies were received, a list of possible candidates for interviews was compiled and the assessment data for each interview candidate was pulled and analyzed. Candidates were selected for interview based upon the breadth of total test score, mathematical backgrounds, and anomalous responses on the assessment (see Table 10). A total of six students were asked to participate after reviewing potential candidates. Return emails with consent forms were sent to all six participants as well as limited access to their assessment with a request to review the assessment prior to the interview.

Table 10

Interview participants from the SCAP

Name	Score	ITP		Major	Year
		Taken	Enrolled		
Al	8	x		Pure	Sophomore
Brent	1	x		Pure	Senior
Christopher	2		x	Applied	Freshman
Gerald	6	x		Pure	Senior
John	6		x	Pure	Sophomore
Shannon	-2	x		Pure	Senior

The interviews themselves took place via ZOOM, an online video conferencing service. The interviews lasted one hour and were audio recorded. Each interview proceeded in identical fashion, with the participant's survey screened shared and each portion of the mathematical content reviewed with the student. For each proposition and argument pair the participant was asked what it was they thought about or attended to while they were validating each of the eight arguments. While they were discussing this, they were not being shown the main reason why they had selected the particular validity

judgement they made so their decisions about their process of validating would be as genuine as possible, as though they were taking the assessment without seeing the reasons the argument might be invalid. Questions were asked about the clarity and effect of individual keys and distractors for each testlet, as well as a set of questions probing concepts or language which might have been problematic (e.g., concepts like equivalence relations, Cartesian cross-products, and symmetric difference, or language like the phrase "double inclusion").

As with the data from the focus groups, the interviews were analyzed using thematic analysis (Braun & Clarke, 2006). The interviews were open coded to determine the process each student took as they attempted to validate each argument. The processes were then analyzed for each student to help determine what the students' individual actions were for validating, with the question of whether students were in fact validating or undertaking some other paradigm. The possibility students were doing something other than validating was left open while coding. The codes themselves helped determine the quality of their process. The two most likely student processes were students validating, or students instead grappling with comprehension as a basis for validating cues. Because of this, codes were resolved using two different frameworks. First was the CVI framework presented in this study and the other was the Proof Comprehension framework from Mejía-Ramos et al. (2012). Though these two were the most likely, the possibility for other processes was left open.

The set of codes represented what I termed the *process* (see Table 11) by which students attempted to validate each argument. I defined the process to be the moment-by-moment approach students took in verbalizing how they viewed the argument. The

process focused on how they spoke about the argument. Were they restating each line of the argument and making comments about the veracity of each line? If so, did they also make mention of the implications between each line of the argument? Did they spend any time considering the proposition itself and how the proposition might affect the argument which was to follow? What part or parts of the argument did they identify as being meaningful in terms of what they were searching for in the argument? All of these questions informed the coding of the process.

Table 11

Set of processes from student interviews

Process	Definition	Indicative Action
Intuition driven checking	This process was informed by the students' intuition about how an argument might be formed based upon the proposition for the testlet.	Validating
Line-by-line checking	This process consisted of a student reading a statement from an argument and then making a validation judgement about the statement. This process might continue through multiple lines of the argument and could also include making judgments about whether each line or statement logically is implied by previous statements. It includes checking algebraic manipulations, notation and parameters, justifications or warrants, and gaps in logic.	Validating
Checking proof frameworks	This process involves checking the overall logical consistency of the proof framework and may answer one of these questions: Does the proof begin and end as it should? Do the basic assumptions of the argument match with what is permissible based upon the proposition? This includes checking if the argument assumes the conclusion or weakens the theorem in any way.	Validating
Probing for local understanding	This process involved the student trying to resolve an initial lack of understanding concerning ideas dealing with; (1) meaning of terms and statement,	Comprehending

	(2) logical status of statements and proof framework, and (3) justification of claims.[19]	
Probing for holistic understanding	This process involved the student trying to resolve an initial lack of understanding of ideas dealing with: (1) summarizing via high-level ideas, (2) identifying modular structure, (3) transferring general ideas or methods to another context, and (4) illustrating with examples.[20]	Comprehending

After resolving these processes with the aforementioned frameworks, the processes were used to determine what I termed the *action* (see Table 11) of each validating session. The action is representative of the quality of the process the students undertook with regards to the frameworks with which they were reconciled, whether they were genuinely validating, comprehending, some combination of the two, or possibly some other construct. Students took a validating action if their responses fully aligned with processes paralleled with the CVI framework. This means they explicitly checked for the types of validity issues that mathematicians determined were common in the ITP setting. Conversely, students took a comprehending action if their responses most aligned with ideas from the Proof Comprehension framework. In these situations, a student's ability to understand the argument was tantamount to the validity of the argument, meaning that a lack of understanding often was enough for invalidating an argument.

As humans are apt to not completely codify into a singular group, crossovers in actions were anticipated and accounted for as an action, which meant not all interview sessions were wholly validating or wholly comprehending. When this occurred, the quantifiers of minimal, some, and considerable were used to clarify the amount of comprehension which was made explicit. Minimal meant the student was involved in

[19] Mejia-Ramos, et al. (2012)
[20] Mejia-Ramos, et al. (2012)

explicit acts of comprehension only on rare occasions and typically to the ends of understanding in a local sense. Some meant the student was involved in explicit acts of comprehension, with most attempts being local in nature but occasionally holistic in nature. Considerably meant the student was regularly engaged in comprehension attempts while validating and, in most cases, they were holistic in nature, though local comprehension may have also occurred.

Conclusion

Creating a semi-open assessment is not my ultimate goal but is certainly the end goal for this study. Each step of this process took careful consideration and attention to detail in order to create a meaningful assessment. Starting with creating and validating the framework and items for the assessment and continuing on to the creation and collection of distractors for each testlet. Each proposition and argument pair were selected for inclusion in the SCAP because it met a series of criteria, including their inclusion in what might be deemed the typical ITP course, and its alignment with both the CVI framing and mathematicians' own evaluations of the items. While this is not the final iteration, the careful efforts to construct a meaningful instrument through the assessment's leveled structure and follow-up interviews with students will be considered in the future. It is in these carefully planned and considered steps that the final assessment gains strength in terms of validity and meaning.

V. Results

The main goal of this study is to create an assessment of students' ability to validate arguments at the ITP level. The following results are an argument that through the process of this study such an instrument has been created, though admittedly, refinement can and should be made. The argument this assessment created, at least in a first iteration, takes on the following nature; (1) results from building and validating the framework, which acted as the undergirding for the assessment's item creation, which results amounts to internal or content validity, (2) results from the open pilot, which created distractors for the SCAP, again more content validity, (3) results from the statistical analysis of the SCAP itself, which results build the case for test reliability, and (4) the results from the student interviews, which results state that while taking the assessment, students were actually validating arguments as opposed to other possible activities. Therefore, the assessment measuring the desired construct.

Framework for Assessment Construction and Item Selection

Here is an in-depth exploration of each of the individual categorizations from the CVI framework from the mathematicians' perspective. The mathematicians' survey is the main resource for determining the consistency of the categories of the CVI framework and for item selection. In each categorization, the goal was to have at least one proposition and argument pair garner at least 90% agreement on validity regarding the categorization. This signals that a sufficient number of mathematicians reasonably agree the argument is invalidated by the CVI issue, and thereby the issue is not only genuine but universal from the ITP standpoint. The focus group data presented is to add clarification about the effect of each CVI categorization. Certainly, I hoped strict

consensus would be found while speaking with the mathematicians, but it was not a requirement to move forward with the assessment. The set of data presented here is a survey of 228 mathematicians and the two focus group interviews ($N = 4$ & $N = 7$) which followed the survey.

Assuming the conclusion. Of all the CVI categorizations, surveyed mathematicians were most unified in terms of the affect that AC had on the validity of an argument (see Table 12). In seeing a total of 4 proposition and argument pairs containing the validity issue of AC, of the 184 occurring argument validations, only 8 mathematicians disagreed that AC caused the arguments to be invalid. This meant 95.7% of mathematicians agreed AC had invalidated the four arguments from the survey. Additionally, when asked directly in the survey if AC in the ITP context was enough to invalidate an argument alone, 171 of a total 178[21] mathematicians – roughly 96.06% of respondents – answered affirmatively.

Table 12

Mathematicians' agreement with the CVI categorization of AC

Item	Responses	Final Response – Invalid	Reason		% Agree[22]
			Agree	Disagree[23]	
AC1	47	47	47	0	100%
AC2	46	46	45	1	97.8%
AC3	46	46	43	3	93.5%
AC4	45	41	41	4	91.1%
Total	184	180	176	8	95.7%

[21] Note that, though there is data for 228 mathematicians, as was mentioned in the methods chapter, only 178 completed the entire survey. This question and others like it were at the end of the survey.
[22] This number can be calculated by dividing the number of respondents (column 4) who agreed by the number of responses (column 2).
[23] As a category, disagree represents all respondents who either thought the argument was valid or thought the argument was invalid but disagreed with the reason I supplied.

Mathematicians who did not agree with the reasoning of AC, or thought the arguments were valid, typically gave responses more akin to grading. These responses took the form of comments like, "I would say that this argument is almost correct rather than invalid," or "It could be modified quite quickly for the proof to be correct." In both cases, the mathematicians gave the benefit of the doubt to the author of the argument, much like one might do when grading and giving feedback to a student[24] (e.g., Moore, 2016).

One interesting comment pointed out an important issue in which a mathematician implied the conditional statement, which was to be proven, was equivalent to the converse which was proven. This is interesting as there are cases where a conditional and its converse are indeed equivalent. In the two focus groups, mathematicians had little to say on this matter overall, but when asked about the affect at the ITP level, Tyler from focus group one had this to say:

Tyler: It has to be logically sound.
Moderator[25]: Logically sound?
Tyler: It has to actually answer whatever the question was.

Here Tyler takes the conversation about AC and broadens it to a conversation on overall logical consistency – a common maneuver by both focus groups' mathematicians – but points out that an argument needs to answer the question being asked. This stems from the oft repeated fact that the discussion was surrounding students learning proof in an ITP setting. Jeremy later commented, "Depends on what level it's at, too. Like, for a student I would like to see a conclusion drawn, so they're indicating to me at least they understand

[24] As mentioned in the methods, mathematicians were asked not to grade the proof, but solely determine validity instead
[25] In all cases I was the moderator.

what the conclusion is." The mathematicians are not blind to the fact that conditional statements are on occasion equivalent to their converses. Instead, it was their opinion that validity regarding AC in the ITP setting is defined in relation to the process of learning proof more generally, which to them is the goal of the ITP class.

The combined understanding from the survey and the focus group indicates that AC is a valid categorization for the CVI framework. Mathematicians were nearly unanimous in their responses in the survey, the few exceptions focused more on aspects of grading student work. The focus groups, though they mostly framed their comments with regards to logical consistency, stressed the role that AC plays in invalidating arguments at this level. For them, the generality of assuming the conclusion being invalid came down to the curricular aim of the ITP course, specifically learning to prove.

The process of selecting an item which at least 90% of mathematicians agreed on was straightforward as all items met this requirement. Since AC1 was the only proposition and argument pair which was unanimous in agreement it was selected for the open survey as was AC3. In the case of AC2 and AC3, they represented an instance where the arguments were for nearly identical propositions. The only difference between the two occurred in the statement of the conditional phrase where AC2 used the more traditional *if-then* structure and AC3 used the non-standard phrasing *then-whenever* (see Figure 20). The choice in this case was to leave the standard language to see if students could identify the type of structure needed to undertake a direct proof of the proposition for AC3.

Proposition: Let X be a set with subsets A and B, if $A \subseteq X - B$ then $A \cap B = \emptyset$.

Proposition: Let X be a set with subsets A and B, then $A \cap B = \emptyset$ whenever $A \subseteq X - B$.

Figure 20. The first proposition is for AC2 which includes the standard "if-then" structure. The second proposition is for AC3 which supplanted the standard structure with the non-standard but equivalent "then-whenever."

Misuse of notation. This categorization lead to an interesting result for the CVI framing. Mathematicians saw two arguments with the validity issue of MN and were in solid agreement with the CVI framing in both instances (see Table 13). The few mathematicians who disagreed – all of whom thought the arguments were valid – did so under the auspice of grading. Their comments were of the following forms, "If the error was pointed out to the student, they could easily fix the argument," or "This is certainly bad, but could be forgiven," or finally, "I would accept the proof in this case as I would regard it as a minor easily fixable slip." On the other hand, when mathematicians were directly asked about the general effect MN has on the validity of an argument, only 40 mathematicians – less than a quarter of all respondents – thought such an issue would invalidate an argument. Thus, despite the two items' performances, MN had some issues with ubiquity in its effect on validity.

Table 13

Mathematicians' agreement with the CVI categorization of MN

Item	Responses	Final Response – Invalid	Reason		% Agree
			Agree	Disagree	
MN1	51	48	48	3	94.1%
MN2	54	50	50	4	92.6%
Total	105	98	98	7	93.3%

While asking about how different constructs from the CVI framework affect validity, focus group members were asked how mathematical grammar and notation affects validity. Focus group one had this discussion which jumps back and forth between two contexts:

Jon:	Yeah, notation is important too.
Tyler:	Yeah. I'm a stickler about that too. It's gotta be stated precisely.
Jeremy:	Yeah, I grade it but I've seen refereed papers that don't show much care.
Tyler:	If you look at old journals from the 20's and 30's, comparing it to today's, we're now much inferior today across the board.
James:	Inferior.
Tyler:	As a purist, I would prefer to see this done well.
Moderator:	Okay.
Jeremy:	I was going to say, sometimes I like to distinguish between having a proof and a proof sketch. So, a lot of times, I'll say, this can be a sketch.
James:	I call it book proof.
Jeremy:	Oh, book proof, yeah…
James:	Yeah, I don't like book proofs.
Jeremy:	…like all the i's doted and the t's crossed. You know, for homework I want them to do that, but quizzes and exams, I tell them, you don't have time to worry about that. I just want a coherent argument.

The three mathematicians discussing the issue of MN in focus group one started commenting on their practice as educators, but then jumped to talking about mathematics in journal writing, implying MN as a writing issue persists to some degree in contemporary published mathematical works. When refocusing on the ITP level, Jeremy further divided the ITP context into that of homework for the ITP class and assessments in the ITP class. Each of these sub-contexts appear to have their own distinct set of norms about mathematical grammar and notation. From Jeremy's comment, he states his standard is much higher on homework than for assessment due to the requirements and aims of each context. For the purposes of understanding MN with regards to validity, this is problematic, though consistent, with what mathematicians from the survey implied.

The underlying implication is that MN is important in terms of grading much like English grammar[26] is. Seeing there is no direct link yet between grading and validity, it is hard to say what this implies about validity. It is important to note Jeremy's students are being informed about each context and the norms for those contexts, making it easy to believe they have an understanding of what is required in terms of MN and validity, or at very least grading.

In the survey results from the two items, MN1 and MN2 the amount of agreement was sufficient to invalidate them. In fact, all the mathematicians who said the arguments were invalid did so because notation had been misused. I learned that though the MN classification in the CVI framing is not always recognized as a genuine validity issue, the two instances where mathematicians judged, it certainly was ubiquitous as they broke the mold in terms of acceptability for validity's sake. This made including and selecting an MN issue reasonable as both items had high agreement, but also a practice of picking between two good options. In general, the MN categorization is not universally recognized in the ITP setting, but the two items from the survey appear to be universal in the broad-spectrum ITP context. In the end, MN1 was selected for use in building the assessment though MN2 would have been just as good as the two items performed identically.

Circular reasoning. When directly asked about the effect that CR has on the validity of an argument, 170 of 178 mathematicians (95.5%) said that at the ITP level such validity issues are enough to invalidate an argument. The survey itself had two examples of CR for mathematicians to validate (see Table 14). As a group, 92.2% of the

[26] English grammar as this study was conducted in the United States completely in English.

mathematicians agreed that CR invalidated the two arguments, but even more notably, the only disagreement occurred with mathematicians who thought the arguments were valid. In other words, 100% of the mathematicians who though the arguments were invalid indicated that it was because CR had occurred.

Table 14

Mathematicians' agreement with the CVI categorization of CR

Item	Responses	Final Response – Invalid	Reason		% Agree
			Agree	Disagree	
CR1	36	35	35	1	97.2%
CR2	54	48	48	11	88.9%
Total	90	83	83	12	92.2%

Circularity lead to almost complete unanimity in response from mathematicians in CR1, with the lone dissenter not commenting in the open section of the survey. On the other hand, CR2 was less unanimous in outcome with 11 respondents disagreeing that CR caused the argument to be invalid. In the argument for CR2, a portion of the original proposition was included as part of the argument before that portion had been proven[27] (see Figure 21). A few of the 11 mathematicians brushed aside the issue making statements like, "I noticed this, but I think it is sufficiently trivial," or "It is a meaningless question." Other more poignant comments indicated the mathematicians used a contextual paradigm for judging, for instance:

> Invalid is too strong. There is no doubt a lack of detail. But, the student has already identified A with a bounded subset of the natural numbers and hence every subset of A with a bounded subset of the natural numbers. At this point, it seems the assumption is about the structure of the natural numbers (which depending on the course, could or could not invalidate the argument).

[27] Using the proposition as part of the argument for said proposition was defined as one type of CR in CVI framework in Table 2.

This mathematician's comment, as well as others, points to a question concerning *what is already known to be true*, a contextual question which other studies have shown to be an important issue in determining validity (e.g., Weber, 2008). For this survey, the task given to the mathematicians was clarified as being the ITP[28] setting and all propositions should be proven with regards to what is appropriate at that level. This means that the context was in fact fixed, and for a strong majority of mathematicians, the inclusion of circularity was sufficient for the argument to be considered invalid.

Proposition: Let A and B be sets. If A is finite and $B \subseteq A$, then B is also finite, and $|B| \leq |A|$. Furthermore, if $B \neq A$ then $|B| < |A|$.

Argument: Let A and B be sets where A is finite and $B \subseteq A$. Since A is finite, there exists some $j \in \mathbb{N}$ such that $\mathbb{N}_j \approx A$. Since A is finite and $B \subseteq A$, it follows that B is also finite. Since $B \subseteq A$, then either $B = A$ or $B \neq A$. If $B = A$, then $\mathbb{N}_j \approx B$ which implies that $|B| = |A|$. If $B \neq A$, then since B is finite, there exists some $k \in \mathbb{N}$ such that $\mathbb{N}_k \approx B$. We have that $B \subseteq A$ and $B \neq A$, thus $B \subset A$ and $k < j$ which implies that $|B| < |A|$.

Figure 21. The argument for CR2 uses the statement in red, which is part of what is supposed to be proven, in order to prove the proposition.

The few comments made by the two focus groups concerning circularity were often spoken of in terms of extraneous information. For instance, in response to a question both focus groups were asked, the following conversation occurred with focus group two:

Moderator: Are there mistakes that are permissible in a valid proof?
Justin: I can think of one situation. A student writes an extra step that's not even needed for the proof for it to be completely valid, and that step contains a mistake.
Moderator: So just some extraneous thing thrown on there that really didn't affect the larger argument? What about in the case of circularity?
Justin: Well, the standard I'm using is, does that proof establish the validity of the fact being proven? So, the student still made a mistake and I'm still gonna point out that mistake, but to me that doesn't threaten the validity of the proof. It perhaps weakens the quality of the proof.

[28] The term/phrase ITP was defined to mathematicians as, "the university mathematics course where students are *first* formally introduced to advanced mathematical thinking, logic, and most importantly, proof."

Trevor: We can assess the quality of the writing, but if it's inconsequential; it does not render proof *invalid* per se.

In this selection of discourse, Justin and Trevor state that extraneous statements in an argument are generally inconsequential in terms of validity, even in cases where there might be mistakes. The focus groups cared about the effect circularity had on the argument's clarity and flow – as in Trevor's closing comment above – much as was suggested by Dawkins and Weber (2016) or Selden and Selden (2003).

On the other hand, one conversation with focus group one did present a distinctly different understanding on the effect of circularity. This conversation was a result of yet another open-ended question asked of both groups meant to elicit input in areas which might not be covered by the CVI framing. In the first group, it prompted a conversation focusing on CR,

Moderator: Are there other major validity issues you see often?
Jeremy: For me, like the circular arguments. There's a degree of circularity where sometimes there's really bad circularity but then sometimes it goes so far out that it's hard for them to see that that's what they're doing.
Moderator: Okay, so what's the effect on validity then?
Jeremy: So, I don't know. To me, this is like a matter of degree. They're not really proofs when they make these mistakes, I would say. I wouldn't, in an ITP course. I would still have them fix it up.
James: But not count off.
Moderator: *Not* count off?
Jeremy: Not count off, they just ... it's all or nothing, but they get infinitely many tries, so to speak.

Jeremy and James's comments suggest an "all or nothing" approach to validity, something both repeated a few times throughout the hour-long focus group. In use here, the two are suggesting that though circularity does affect clarity, as was previously discussed, for group one it also invalidates the arguments. Jeremy goes so far as to say

that these kinds of arguments "are not proofs" because of their circular natures. This paints a considerably different picture than what was presented by focus group two.

No real consensus exists as to the effect of CR generally in an ITP argument as two groups of independent mathematicians came to different conclusions. Conversely, it is important to note that focus group two's discussion focused more on extraneous data whereas focus group one talked directly about circularity. This latter understanding along with that gained from the survey itself paints a consistent picture that CR is a valid part of the CVI framing, and using the 90% benchmark, this meant that CR1 was included in the student assessment.

Logical gaps. There were a total of four proposition and argument pairs in the survey dealing with LG (see Table 15). The categorization of LG had far less agreement than the prior categorizations, understandably so given that research shows that even in ITP class textbooks details are sometimes glossed over, including even banal matter of stating givens at the beginning of a proof and closing proofs by stating what was proven (e.g., Fagan & Melhuish, 2018). Because of this, I knew going into the process of building the CVI framework and creating items for a student assessment that LG would be a matter of preference, which could widely vary. It was, therefore, my goal from the survey to understand what LGs were more universally atypical in the ITP setting and test the levels of logical gaps until something breaks to better understand the tolerances that exist for ITP arguments. To do this, I intentionally put arguments in front of mathematicians that had what I felt were minor to major logical gaps to give them the ability to set the norm.

Table 15

Mathematicians' agreement with the CVI categorization of LG

Item	Responses	Final Response – Invalid	Reason		% Agree
			Agree	Disagree	
LG1	46	42	42	4	97.7%
LG2	36	26	26	10	72.2%
LG3	35	17	17	18	48.6%
LG4	35	13	13	22	37.1%
Total	152	98	98	54	64.5%

First and foremost, LG1 surpassed the necessary 90% benchmark. On the other hand, item LG3, represented in Figure 22, offered an interesting insight into the split nature of mathematicians on this construct. At a near perfect split, 48.6% of mathematicians reported that the argument for LG3 was invalid. Breaking with the normal structure of the survey for this item – as well as LG2 and LG4 – mathematicians who indicated the argument was invalid, instead of being presented with why I thought the argument was invalid, were presented with an open prompt and asked to explain why they thought the argument was not a valid proof. This probed the tolerance mathematicians had regarding LG on validity. In this open prompt, all 17 mathematicians commented about the lack of detail. For instance, one mathematician said:

> It feels like an assertion. In both cases, the author first substituted into the definition of $A \triangle B$, and then asserted that that expression equaled what we want to show. There's no appeal to the underlying reasoning of why $(A - \emptyset) \cup (\emptyset - A) = A$. In some contexts, I'd give the author the benefit of the doubt, namely if the reader has been making those exact leaps earlier in the text. Absent that, it feels like the heart of the proof is missing.

This comment is indicative of the other open form comments, and points out two important ideas, similar to those about CR; (1) context is an important factor for mathematicians with regard to validity (e.g., Weber, 2008), and (2) the lack of detail

invalidates this argument at the ITP level as students would not prove using "those exact leaps earlier."

Definition: The symmetric difference of A and B is defined as $A \triangle B = (A - B) \cup (B - A)$.

Proposition: For any sets A, $A \triangle A = \varnothing$ and $A \triangle \varnothing = A$.

Argument: Let A be a set, then by the definition of symmetric difference $A \triangle A = (A - A) \cup (A - A)$. But $(A - A) \cup (A - A) = \varnothing$ which implies that $A \triangle A = \varnothing$, as required. Furthermore, also by the definition of symmetric difference $A \triangle \varnothing = (A - \varnothing) \cup (\varnothing - A)$. But here $(A - \varnothing) \cup (\varnothing - A) = A$ which implies that $A \triangle \varnothing = A$. Thus we have shown that for any set A, $A \triangle A = \varnothing$ and $A \triangle \varnothing = A$.

Figure 22. The argument was intentionally trivialized at multiple instances by choosing to not include arguments as to why $(A - A) \cup (A - A) = \emptyset$ *and* $(A - \emptyset) \cup (\emptyset - A) = A$.

After an in-depth review of all the comments from the open responses to LG2, LG3, and LG4, all mathematicians who indicated these items were invalid claimed the arguments lacked detail. One mathematician said of LG3, "It does not contain enough detail for me to feel confident that a student in an ITP[29] course actually understands why the [premise is true]." Similarly, for LG4 one mathematician summarized that, "Well it's valid, but not from a beginning student" further implying the argument had leaps which for an ITP student were not permissible.

When talking with mathematicians about LG during the focus groups, the same salient theme emerged during the general discussion as did from the survey: *details*. The focus groups were relatively unified in their feelings about what amount of detail might or might not be required in an ITP setting. The following conversation occurred early in focus group one,

Moderator:	What aspects does an argument have to have to qualify as valid?
James:	I like details.
Moderator:	Details are necessary?
James:	All the details, yeah. Show me the details.
Jeremy:	[What level student are we talking about?]
James:	Well, is this about the beginning, are they learning to prove or is it about-

[29] This mathematician's use of ITP was genuine and not shorthand transcription.

Jeremy: Yeah, which level is it?
James: Is it graduate school?
Moderator: [The ITP level]. That's the class that we're talking about. That's those very beginning students, the first time they're seeing proof.
James: There we're getting details.
Jeremy: Right, there we need details.
James: That's our main duty.

For these mathematicians, they felt it was important to include details, with James stating the main emphasis of teaching proof in ITP courses is getting students to fill in the details. Additionally, these mathematicians echoed the surveyed mathematians who implied context is important in determining validity requirements. James was rather forceful in his assertion that details matter and multiple times throughout the focus group one discussion started conversations like the following:

James: Details, man, details, we want details.
Moderator: You want details? Okay.
Tyler: Yeah, at that stage, more details.
Jeremy: …the devil's in the details.

The takeaway from the focus groups was that details, or a lack of logical gaps, are a highly desired piece of valid proof writing in the ITP setting. This understanding combined with the results of the survey certainly place a contextual constraint on the categorization of LG, though this categorization is absolutely a key piece of the CVI framing and on the assessment.

There was not complete unanimity when talking about detail and – by extension LG. In focus group one's conversation, Jon commented about a presented example, "What's obvious is relative…It's easy to fill in the little gaps if you think there is one." This comment is noteworthy for a couple of reasons: first, we were exploring argument LG4 from the survey, the one most mathematicians thought was valid; and second, because Jon said little concerning his feelings about the effect of detail on validity.

Looking back, it is hard to determine if he generally felt that details – as a replacement construct for LG – was not an issue, or if he felt that the logical gap in LG4 was insignificant, as did the rest of the mathematicians. I did not collect personal information in the survey itself, so I cannot say if Jon saw this item when taking the survey, and, therefore, I cannot say anything conclusively on his feelings generally. In the worst case, where Jon does generally think that LG is not a validity issues, he is something of an outlier. At best he agrees that LG4 is valid and had no other thoughts to share on the matter. In either case, the replacement construct of detail is considered by most mathematicians to be a significant validity issue. Thus, it is consistent that LG is a valid part of the CVI framing. Once again, using the 90% standard, this means that item LG1 is the only item which carried forward to the assessment phase.

Warranting. Mathematicians were asked about three different ways which warranting could affect the validity of an argument (see Table 16). In each instance, the mathematicians were asked if the warranting issues would necessarily invalidate an argument at the ITP level. The survey indicated mathematicians most consistently felt – 163 of the total 178 surveyed, or 91.6% of respondents – that an argument is invalidated by a warrant incorrectly used to justify a claim[30]. However, stating an incorrect justification[31] or including an arithmetic error[32] invalidating an argument in the ITP setting garnered 71.9% and 69.1% agreement respectively among surveyed mathematicians. This implies not all warranting issues affect validity the same at the ITP level, with more emphasis placed on using a justification consistent with its intended

[30] This group was coded as $\omega 1$.
[31] This group was coded as $\omega 2$.
[32] This group was coded as $\omega 3$.

meaning (*ω1*). It is interesting to note that an arithmetic error could be thought of as using a warrant incorrectly – meaning an *ω1* warranting issue – in which case the divergence of opinion leads to more questions than answers.

Table 16

Mathematician's views about the effect of three types of warranting issues

Warranting Issue	Code	Example	# Agreed	% Agreed
Using a warrant incorrectly	*ω1*	Enacting the converse of a theorem as a means of justification.	163	91.6%
Using the wrong warrant	*ω2*	Stating associativity when commutativity was the proper warrant.	128	71.9%
Arithmetic errors	*ω3*	Incorrectly squaring a binomial.	123	69.1%

From the outset, warranting was an area of the CVI framework I strongly desired mathematicians' input to clarify what I felt were inconsistency regarding the practice generally. For this purpose, I included a total of eight (8) warranting items on the survey (see Table 17). Of these eight items, there were one *ω1*, four *ω2*, and three *ω3* warranting issues. Items W2 and W3 represented warranting items that superseded the 90% threshold for agreement. None of the *ω*2 items had 90% agreement, though 96.6% of the mathematicians – 56 out of 58 respondents – agreed that W1 was invalid. The surveyed mathematicians were generally united on the effect *ω1* issues have on the validity of arguments, which aligns with the CVI framework. Because of this, not only does it seem justified that *ω1* issues are a valid part of the CVI framework, but item W2 also included in the student assessment. What is left is to better understand *ω2* and *ω3* warranting issues.

Table 17

Mathematicians' agreement with the CVI categorization of W

Item	Code	Responses	Final Response – Invalid	Reason		% Agree
				Agree	Disagree	
W1	*ω2*	58	56	27	31	46.6%
W2	*ω1*	61	56	56	5	91.8%
W3	*ω3*	35	32	32	3	91.4%
W4	*ω3*	59	51	48	11	81.4%
W5	*ω2*	59	47	47	12	79.7%
W6	*ω2*	62	46	46	16	74.2%
W7	*ω3*	50	34	33	17	66.0%
W8	*ω2*	60	36	36	24	60.0%
Total		444	358	325	119	73.2%

At first glance $\omega 2$ items, or items where the argument used the wrong warrant or lacked a correct warrant, looked like they performed poorly. There was an overall agreement of 65.3% that the arguments were invalidated due to *ω2* warranting issues (see Table 18). Some of this is attributed to item W1, which quizzically accumulated only 46.6% agreement but was considered invalid by 56 of the 58 mathematicians who saw it. No other $\omega 2$ item – or warranting item for that matter – did as poorly as W1. Because of this I went back and reexamined the item.

Table 18

Mathematicians' agreement with the CVI categorization of ω2

Item	Code	Responses	Final Response – Invalid	Reason		% Agree
				Agree	Disagree	
W1	*ω2*	58	56	27	31	46.6%
W5	*ω2*	59	47	47	12	79.7%
W6	*ω2*	62	46	46	16	74.2%
W8	*ω2*	60	36	36	24	60.0%
Total		239	185	156	83	65.3%

It turns out that item W1 performed poorly because, by my own oversight, this item had two validity issues. There were supposed to be two versions of this item initially, one which had a warranting issue and one that assumed the conclusion. The version that mathematicians saw ended up having both issues. As a result, though 27 mathematicians agreed that there was a warranting issue, of the 58 respondents for this item, 44 (75.9%) said that the argument was invalid mainly because the issue of AC. A few of the respondents who disagreed concerning the warranting issue made statements like, "I saw that it was an attempt to prove the converse and stopped reading at that point." As the warranting issue for the argument occurred a few lines into the argument, it was impossible for mathematicians who enacted this sort of behavior to agree that a warranting issue existed. With W1 removed from the set of warranting issues, items W2-W8 garnered 75.9%[33] overall agreement concerning the effect of warranting issues on the argument's validity. Furthermore, $\omega 2$ are bolstered by the removal of W1, going up to 71.3%[34] agreement. As the data for the $\omega 2$ items were discouraging, none carried forward to the assessment stage. However, questions were brought up during the focus group to better understand this phenomenon.

While $\omega 2$ warranting issues did not have a single item that performed well enough to move on to the assessment, $\omega 3$ issues did. Item W3 garnered 91.4% agreement from mathematicians (see Table 19). Overall arithmetic errors saw 78.5% agreement, with item W7 performing the worst of the three items. For item W3, only 3 mathematicians disagreed in the end, and the following mathematician aptly sums up the thoughts these mathematicians had:

[33] Not significantly different from the original proportion.
[34] Not significantly different from the original proportion.

> I don't know exactly. Technically it's wrong, but if it were another mathematician who had written this proof, I'd never doubt that they were actually confused on that point. If I were grading this as a student proof, that would come into play, but the instructions said explicitly not to. As to its validity, the erroneous distribution of (2k+1)^2 doesn't derail the essence of the proof, and it successfully gets another mathematician through to the correct proof, and so I put it in the category of giving the author the benefit of the doubt. At the same time: if I were refereeing a paper with this error, of course I would note and correct it for the author, because it's actually not correct. But I'd classify it as a typo and not a conceptual proof error, which I suppose is the level I care most about.

This mathematician was clearly conflicted and recognizes the mistake made. It is also clear the mathematician attempted to validate and not grade the proofs, but doing so at a level beyond what was asked, as they reference seeing other mathematicians perform such an error. Their comment does indicate that the potential validity issue is genuine for them, though they do not feel it sufficient enough to invalidate the argument. Despite the three who disagree, this item was taken to the next round and included in the open-pilot.

Table 19

Mathematicians' agreement with the CVI categorization of $\omega 3$

Item	Code	Responses	Final Response – Invalid	Reason		% Agree
				Agree	Disagree	
W3	*ω3*	35	32	32	3	91.4%
W4	*ω3*	59	51	48	11	81.4%
W7	*ω3*	50	34	33	17	66.0%
Total		144	117	113	31	78.5%

Focus groups. As was mentioned from the outset, much of the conversation surrounding LG and W were re-framed as *details* by the mathematicians during the focus groups. This meant it is hard to determine which specific construct they might have been referring to during the analysis process. This possible dual meaning is understandable as the two constructs, LG and W, have intentional overlap, as both are defined with the understanding that warranting is part of the argumentative structure (Toulmin, 1964) for

which an LG is implicitly missing a series of warranting. For mathematicians who not accustomed to the vernacular I adopted, the two ideas may not be distinct in their mind. There are clear moments where the conversation was about one or the other construct, and in those situations, I was sure to include the contextual clues to which specific construct they were discussing.

In the focus groups, the conversations about *ω1* and *ω2* warranting issues in the ITP setting revolved around the idea of novelty with respect to specific justifications. This first segment begins with focus group two refocusing on warranting requirements:

> Trevor: Let me go back to your last question, which is, "in papers that are published, is that the same level of rigor as the classroom or classroom expectations?"
> Moderator: Yeah, that was the question.
> Trevor: I don't think so, but I think that's a different standard. In a math paper the author may say, "well, it's easy to see that <boom>." We don't necessarily want that from our students.

This excerpt begins the conversation on the differing requirements concerning warranting in relation to context. Trevor is discussing the differences between journal writing, writing proofs in the classroom, and the expectation of students' own written proof in the ITP setting. His comments set up the idea that, to some degree, the level of rigor is higher in the classroom than in journal writing with respect to warranting.

From this point, Trevor discussed further his ideas about the juxtaposed nature of classroom expectations and journal expectations. Note again, that while this conversation began with a discussion on warranting, it could very well support the role which LG plays in validity at the ITP level as well:

> Trevor: But I think that's a matter of, what's the audience, and also, how far out is your resource? In classes we're seeing very old results, in some cases, not all that difficult, so we can afford to ask for all the details. But if you're publishing a paper with some brand-new theorem, then you may need to get heavy on the details of that

new part, but not in the part you need to go through quickly to get to what is your actual contribution. I think it's a hard question to answer because it depends on what part of the proof you're referring to and what's the audience for that proof. So, for the student, yeah this stuff we are seeing is old, but to them it is new and we need to see that they know, just like in a journal, you know, the details need to be there.

Moderator: So, when justifying for students and in the larger world of mathematics, the proximity has a real effect on how explicit you need to be and when? Would you guys – I see a lot of nodding heads – do we agree with that? That justification itself is somewhat dependent upon what thing is novel within a proof?

Justin: Yeah.

Here Trevor introduces the idea that novel ideas, whether in the classroom setting or in the larger world of published mathematics, are the ideas for which details – details like justifications and stronger logical connections – are required for the argument. In fact, this idea was an echo of a similar idea presented by focus group one where they discussed the effect of a *ω2* warranting issue:

Moderator: There is going to be a point where your ITP student might later be sitting in an algebra class and they're doing a proof that requires them to call upon the first isomorphism theorem. Is it enough that they use it correctly or do you want to see that they know that they're pulling from the first isomorphism theorem?

Jon: If I just taught it, then I want them to state it on the problem that follows that. But later, no.

Moderator: Later, no?

Jon: And it's actually quickly later. Not anymore.

Moderator: Okay

Jeremy: I would circle the justification and note that everything's good except for that, fix it. But I'd let them turn it in again and try to fix the thing, so it'd become a proof. On the other hand, I would say technically that what we are looking at here is not a proof because if there was a compiler or something like that, it would definitely reject it. If you have a theorem prover it would say, 'invalid.'

Here Jon is indicating, much as was inferred by focus group two, that the effect of warranting on the validity of an argument is dependent upon the novelty of said justification. Thus, from these two conversations, the implication is that *ω2* issues only

have any large scale bearing for warrants which are novel in the ITP setting. This means that new definitions and theorems are important as explicit justifications so long as they are recently learned.

This is an interesting point, as in the ITP setting, some of the definitions and theorems may not truly be novel for the students, and in some cases are not strictly defined as in the case of field properties (Fagan & Melhuish, 2018). Focus group one discussed the need for specific justifications which were included in an argument they read (see Figure 23) during the focus group:

Jeremy:	You don't need those *words* at all do you?
Jon:	Exactly. You don't need them at all. So just delete it.
Moderator:	So, when you're teaching an ITP class, or other classes like it, do you try to shy away from specific justifications?
James:	No, I like to have exactly accurate justifications.
Jeremy:	You have to have a justification. And I do mine like all or nothing at that level.
Tyler:	It's a good point this one. Because in that class, very often I'll tell the students that I think I'd be happy with a period after such and such a claim, that does not bother me. But I tell the students, 'but you have to understand that when I read the proof, I need to be convinced that you know what you're talking about. So maybe you should put more information to let me absolutely know for sure that you do know.' Because if you put a period there and I'm thinking, well how did they know that. In this case that is fine, but there are some gray areas where I tell them, give me more information, then I'll know for sure that you knew where it came from.
Jon:	Fully agree.

This conversation, as well as the overall discourse across the two focus groups, gives the impression of a moving target where at points more explicit warranting is required and in some instances less is better. This understanding supports the previous notions that (1) warranting is explicitly dependent upon the novelty of the warrant, and (2) at the ITP level, the set of justifications from which one can warrant are somewhat inconsequential in nature meaning that at times less explicit warranting is actually preferred. From a

validity standpoint, this means that *ω2* as a categorization is too problematic and inconsistent. All this combined with the results from the survey indicate that such notions should not be carried forward to the assessment.

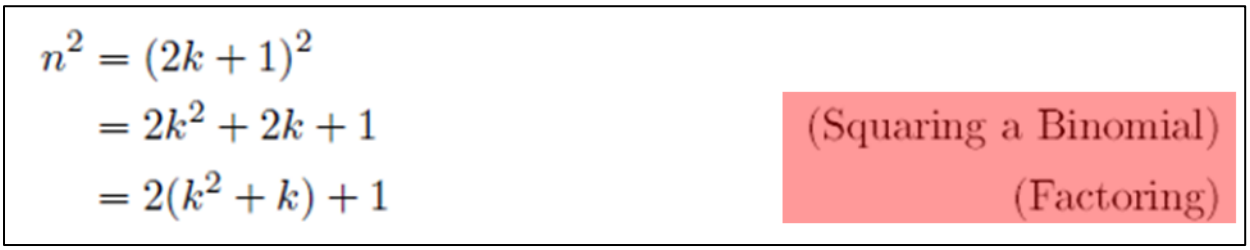


Figure 23. Portion of prompt which included specific justification (red) or "words" as Jeremy called them. The argument included an incorrect squaring of the binomial.

Unfortunately, dealing with arithmetic errors – meaning *ω3* warranting issues – the conversations in both focus groups were short. The following came from focus group one near the beginning of the general discussion of validity,

Moderator: What about correct arithmetic manipulation?
Jeremy: Yes.
Moderator: Yeah?
James: Yeah, that better be there.
Tyler: If it's a part of the argument.

Here, the mathematicians, and especially Tyler, imply that arithmetic is important, so long as the arithmetic is an important part of the argument. This idea is similar to the idea of novelty previously discussed, where new ideas about arithmetic manipulations should be done correctly, but things which are not contextually new do not hold as much sway. The conversation continued as they looked at item W3:

Tyler: It's a trivial mistake, this one.
Moderator: This one you think is a trivial mistake?
Tyler: Yeah, I mean it is a mistake…
James: That's an arithmetic mistake
Tyler: …but it's a silly mistake.
James: It doesn't count off that much.
Tyler: Yeah, it doesn't count that much.
Moderator: But is wrong?
Tyler: It's a terrible mistake though.

Moderator: It's completely wrong though, right?
Jeremy: It is completely wrong. That's right.
Moderator: There isn't an ITP world where it's right?
Tyler: Yeah, yeah.
James: Right, right.
Jeremy: So again, if it was on an exam, I could clear that but if it was homework, I would send it back.
Tyler: That's the kind of mistake though that would infuriate me, really.
Jeremy: Once they're in an ITP course they should be able to square a binomial.

The conversation continued as Tyler presented an anecdote about students who are unable to perform basic computations. It became very clear that these mathematicians had an amount of vehemence about this subject. Their thoughts, once again, indicated that proximity is important, but this did not concur with the outcome of the survey as W3 was considered almost unanimously to be invalid because of the arithmetic error. Additionally, Jeremy brought up the point about the differing context of homework and exams and the different requirements of each context. As with before, the homework context has a higher level of rigor than the exam context.

In the end, as W3 did surpass the benchmark of 90%, this leaves some questions as to the effect of the *ω3* warranting issue. The focus groups indicated they would apply negative grades and feedback in response to homework with these sorts of mistakes, and as has been pointed out, grading to some degree is a surrogate construct for validity. It is possible then that students might recognize arithmetic as a genuine validity issue. As such, and because W3 did preform as required, it was included as part of the student assessment.

Weakening the theorem. The final construct from the CVI framework is WT. The items from WT were split into two groups: those that performed very well, WT1 and WT2, and those which did not (see Table 20). The items which did not perform well were

to some degree designed to test the bounds of mathematics, especially from a linguistic standpoint.

Table 20

Mathematicians' agreement with the CVI categorization of WT

Item	Responses	Final Response – Invalid	Reason		% Agree
			Agree	Disagree	
WT1	47	46	46	1	97.9%
WT2	53	52	52	1	96.2%
WT3	43	33	11	22	27.9%
WT4	36	25	23	13	63.9%
WT5	49	18	15	34	30.6%
Total	228	174	147	71	64.5%

Item WT5 (see Figure 24) represents what was perhaps the most interesting of the WT items. In designing this item, I intentionally left the nature of the proposition ambiguous with regards to the term odd – does this term automatically imply odd natural number or odd integers? Additionally, in the argument itself the phrase, "The $x = 2a + 1$ for some $a \in \mathbb{N}$" was included again to probe the often-inconsistent definition of the set of natural numbers with regards to the number 0. Clearly, most mathematicians thought this meaningless in terms of validity, most commenting that the structure of the argument was the most important thing, not the individual pieces. One mathematician commented:

> The heart of the argument is understanding that odd numbers are 1 mod 2 and that an odd number squared is 1 mod 2, which remains valid. The error is minor because of its consequence. If this was a proof involving absolute values and the negative numbers [were] not properly dealt with that would be much more damning.

This is not the consensus of all the comments, as some mathematicians were perturbed by this prompt, and a large amount of open comments were made. The implication is that,

yes there seems to be some sort of disagreement in terms of implicit language, but overall the arguments structure supports the proposition.

Proposition: If x is odd, then x^2 is odd.

Argument: Suppose x is odd. Then $x = 2a + 1$ for some $a \in \mathbb{N}$. Thus we have

$$x^2 = (2a + 1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1.$$

Since $2a^2 + 2a \in \mathbb{N}$, then $2(2a^2 + 2a) + 1$ is odd, and therefore so is x^2.

Figure 24. Item WT5 was designed specifically to test the bounds of WT in regard to implicit language.

Due to time constraints on the mathematicians, neither focus group had much of a conversation concerning WT generally. The conversation associated with WT, though interesting, falls outside the consideration of this study and will therefore be reported on at some other time. That said, though there were items which performed poorly on the survey, it was by design. Therefore, it is understandable that certain of the items did poorly. On the other hand, items WT1 and WT2 garnered enough agreement to easily be included in the open-pilot. It is easy to exclude certain items which performed poorly and accept WT as a legitimate CVI categorization, with the caveat that more focus group data would be helpful in further legitimizing this categorization.

Valid items. As defined in the framing for this study, valid items were those items which did not include a CVI issue. No one of these items were universally judged to be valid, but two of the five, items V1 and V2, surpassed the 90% threshold (see Table 21). Item V3 only missed the threshold by the thinnest of margins – one response. Mathematicians who thought item V4 to be invalid grappled with the definition of odd, an important fact in the argument, which is interesting considering the difficulty that arose from item WT5 in the previous section. This only strengthens the point that implicit

mathematical language can be problematic in terms of agreement among mathematicians. Finally, mathematicians who thought item V5 was invalid did so under the auspice that the conclusion was not expressly stated and was therefore incomplete, as stated by several mathematician, “The argument is incomplete,” or “Did not finish the argument,” and finally, “Does not finish the proof.”

Table 21

Mathematicians’ agreement with items which do not have a CVI issue included

Item	Responses	Final Response – Invalid	% Agree[35]
V1	37	2	94.6%
V2	37	3	91.9%
V3	37	4	89.1%
V4	37	6	83.8%
V5	41	9	78.0%
Total	189	24	87.3%

Though only two items garnered the requisite 90% agreement, the valid items performed admirably. Importantly, as will be discussed more in-depth in the next section, no new categorizations arose from the presentation of valid items. Part of presenting mathematicians with valid prompts was to account for the possibility that there are other categorizations that needed to be accounted for in the CVI framing. and it would be possible that other categories might arise. As 87.3% of the mathematician who took the survey deemed this group of five items valid, and comments made by mathematicians who thought the arguments were invalid aligned with a framework categorization or could be accounted for in other ideas like grammar or clarity (e.g., Moore, 2016), it is safe to assume this process yielded no new categorizations. For the purpose of building

[35] Here *% Agree* represents the percentage of mathematicians who agreed the arguments were in fact valid.

an assessment, items V1 and V2 were carried forward for inclusion in the final assessment as both garnered the required agreement.

Other possible categorizations. During both the survey and in both focus groups, mathematicians had the opportunity to point out some other validity issue or issues that I did not take into account when putting together the CVI framework. In the survey, I explicitly asked mathematicians if there were other types of validity issues they did not see in the survey which they thought were common at the ITP level. In total, 25 mathematicians responded in the affirmative to this open question, asserting various ideas as common validity issues. Of these 25 responses, 15 included comments inferring existing categorizations in the CVI framing (see

Table 22). For instance, the following comment was coded as a *logical gap*, "Incomplete arguments that leave major steps out and unexplained." All other codes which were not specifically about a CVI issue dealt with comments that mathematicians made either about the survey generally, or about specific ideas concerning validity. Responses coded as a "comment" were like the following response which was coded as *comment on severity of error*; "What would invalidate a proof is an error that cannot be easily recovered from by a small correction." The mathematician in this comment is not submitting a novel categorization, but rather commenting on the general effect of the severity of a validity issue. From coding and analyzing these responses, no new categorizations for the CVI framing were put forth by mathematicians in the survey.

Table 22

Coding of other possible categorizations for CVI from mathematician survey

Code	Definition	Count
Logical gap	As defined by the CVI framing.	5
Comment on survey	The mathematician used the open prompt as an opportunity to comment on the survey generally.	4
Warranting – Justifying generally	As defined by the CVI framing.	4
Comment on the severity of error	The mathematician indicates that there is a difference the effect of any issue based on severity of the error.	4
Comment on arithmetic or grammar	The mathematician express ideas about the effect of arithmetic errors or flaws in grammar as a direct response to what they saw in the survey.	3
Weakening the theorem	As defined by the CVI framing.	3
Comment on warranting	The mathematician express ideas about the effect of warranting as a direct response to what they saw in the survey.	2
Misuse of notation	As defined by the CVI framing.	2
Warranting - arithmetic error	As defined by the CVI framing.	1

During both focus groups, mathematicians were also asked if there were validity issues they had not seen in the survey or discussed during the focus group that should also be considered. The idea most discussed was the idea of details. As has already been stated, mathematicians used the term detail extensively and, in all cases, seemed to infer the CVI issues of LG or W. Another idea which was voiced in both groups was grammar and clarity as was discussed by the mathematicians in Moore's (2016) study on grading practices. While individuals in both groups expressed frustration on this point neither

concluded that it was a worthwhile candidate for invalidating an argument. Instead, the discussions turned to grading, echoing the conclusions drawn from Moore's (2016) study. No other ideas were presented by either focus group, and thus when combined with the results from the survey, the conclusion is drawn that the six categorizations initially presented in the CVI framing form a sufficient basis for the subject of validity at the ITP level.

Conclusion. Though no categorization for the CVI framing had universal acceptance from all mathematicians – which was never the goal to begin with – there certainly were areas which met the requirements for this study. Not only did the CVI framing meet the requirements set for it, each categorization had at least one item with 90% agreement, with most of the categories generally agreed upon by mathematicians as areas which can and do cause invalid arguments at not only the ITP level, but in other more general contexts too. From the survey and the two focus groups, it is clear that the relationship between a mathematician, context, and expertise play a role in validity judgements as has been shown in previous studies (e.g., Inglis & Alcock, 2012; Inglis, et al., 2013; Weber, 2008). Because the CVI framing is the basis for the assessment, and it performed as expected, there is a strong case for the assessment in terms of validity, as each item for the assessment comes from this process.

Assessment Development and Piloting

The open-ended pilot consisted of twelve (12) arguments from the CVI framing survey given to mathematicians (see Table 23). A total of 68 students took part in the open-ended pilot whose results were analyzed to create a set of distractors for the closed multiple-choice assessment that would follow. It is too cumbersome to detail the process

for each item, so instead I present a set of results from one item, pilot item P1, and the process undertaken to create distractors for said item.

Table 23

Items for the close-ended pilot from the CVI framing

Pilot Item	CVI Name	CVI Validity	ITP Topic	% Mathematician Agreement
P1	AC1	Invalid	Basic Number Theory	100.0%
P2	WT1	Invalid	Basic Number Theory	97.9%
P3	LG1	Invalid	Set Theory	97.7%
P4	CR1	Invalid	Set Theory	97.2%
P5	V5	Valid	Relations	94.6%
P6	WT2	Invalid	Set Theory	96.2%
P7	MN1	Invalid	Relations	94.1%
P8	AC3	Invalid	Set Theory	93.5%
P9	MN2	Invalid	Basic Number Theory	92.6%
P10	V4	Valid	Set Theory	91.9%
P11	W2	Invalid	Set Theory	91.4%
P12	W3	Invalid	Basic Number Theory	90.3%

Example process. Item P1, AC1 from the CVI framing, presented students with an argument for the converse of the proposition being proven (see Figure 25). As with all items for the open-ended assessment, students were first asked to validate each argument and then to describe in detail why they thought the argument was or was not a valid proof. It was from this second, open-ended question that the distractors were created. Each of the 68 responses were coded multiple times throughout the analysis phase to consider new codes that arose in the coding process (see Table 24). Note that the codes in Table 24 do not represent all codes from the open-ended assessment, just the set of codes which pertained to item P1.

Table 24

List of codes and responses for P1

Code	Definition	Responses
Conclusion	The comment alludes to the proof assuming the conclusion.	31
Def	The comment references definitions or concerns the definition of a concept used or not used in an argument.	19
Produced	The comment consisted of a student produced proof, outline, examples, diagram, counterexample, non-proof.	14
Logic	The comment deals with the logic/logical flow of the argument.	13
Sense	The comment includes a phrase similar to "makes sense," or the validity is based upon understanding the proof. Often the comment is an explanation of what was done in the argument but does not restate the entire argument.	11
Comprehension	The student's claim is based upon their ability to restate the argument and understand it. In most cases the comment is a copy of the argument, perhaps with some explanation but does not necessarily include it.	7
Direct	The comment references the argument was a direct proof or that it should have been a direct proof.	4
Algebra	The comment makes remarks about the use of algebra - computations and manipulations	3
Explanation	The comment references the inclusion of or lack thereof of a sufficient explanation.	3
Null	No comment was made.	2
Detail	The comment references the inclusion of or lack thereof of sufficient detail. Usually including a statement similar to “needs more detail.”	2
Clarity	The comment references the clarity or that it is easy to follow.	2
\in	The comment references elements belonging to sets.	2
Case	The comment references that all cases in a proof by case were accounted for or that there were cases that were missed.	1
Looks good	The comment is a non-mathematical statement about how the proof looks, often saying things like, "looks good" or "appears to be good."	1
End of proof	The comment indicates the desire for a “therefore” statements at the end of proof or some sort of clear closing remark.	1
Prop true	The comment references a conviction in the truth of the proposition rather than the argument.	1
Warrant	The comment references the justification and warranting in some direct way. "They didn't justify the statement or claim"	1
Axioms	The comment references basic axioms, and specifically uses language including the word axiom.	1
Previous	The comment refers to something from an earlier argument either within the assessment or from prior knowledge, or is a comment concerning some other argument in the assessment.	1

Proposition: If x is odd, then the sum $x + 4$ is also odd.

Argument: Assume that $x + 4$ is odd, then there exists an integer n such that $x + 4 = 2n + 1$. Thus we have that $x = 2n - 4 + 1 = 2(n - 2) + 1$. Since $n - 2 \in \mathbb{Z}$, then x is odd.

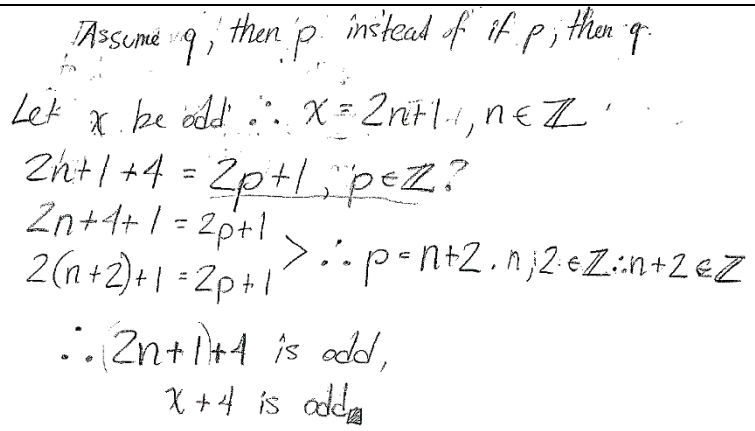
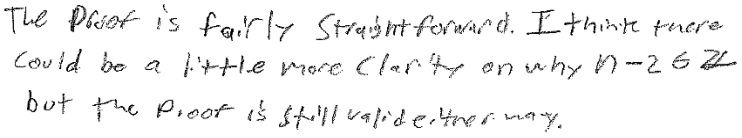
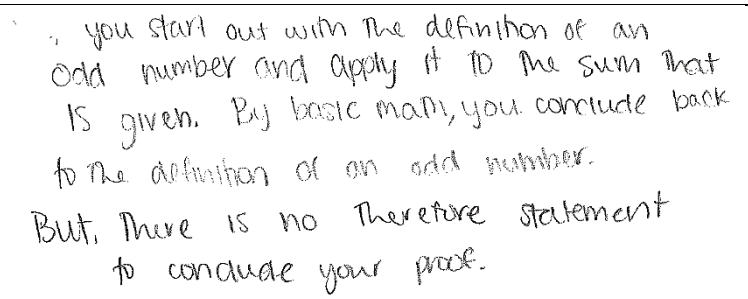
1.) Do you think the above argument is a valid proof for the included proposition (select only one)?
☐ **Yes**, it is a valid proof.
☐ **No**, it is not a valid proof.
2.) In as much detail as possible, explain why the argument is or is not a valid proof.

Figure 25. Item P1 from the open-ended assessment was AC1 from the CVI framing process.

Of the 68 comments from the open survey, 31 grappled with the fact that the argument assumed the conclusion. The next largest grouping of comments dealt with definitions (19) or contained a student produced argument (14). Open responses were often coded using multiple codes as it was entirely possible for students to include multiple ideas in their responses. For instance, Table 25 includes a set of three different responses and their codes for P1, where each response was coded using two or three different codes.

Table 25

Three example open responses for P1 from open-ended pilot with response coding

Open response	Response coding
	• Conclusion • Produced
	• Looks good • \in • Clarity
	• Comprehension • End of proof • Def

Once I completed coding the entire open pilot, I compressed the codes[36] to form categories where each category included multiple codes which focused on similar themes throughout the open pilot. For item P1, the counts for each category are found in Table 26. Each category listed in Table 26 includes many ideas not present in P1, but which arose for other items in the pilot. For instance, no code in Category 8 appeared in item P1 as students did not make comments for this item on ideas concerning symbols, parameters, notation or variables though they did for other items.

[36] The code *Null* was excluded from categorization compression.

Table 26

Condensed category responses for open pilot item P1

Category Number	Category Name	Codes in Category	Counts	% responses ($n = 68$)
1	Logical / Structural	Logic, Conclusion, Case, iff, Structure, General, Gap, Initial Conditions	45	66.2%
2	Understanding	Sense, U/G, Comprehension, Prop True	19	27.9%
3	Manipulation / Algebra / Set Theory	Algebra, Element, Set, Substitution, \in, \notin, |	5	7.4%
4	Conceptual	Def, Concept, Axioms	20	29.4%
5	Presentation / Clarity / Conventions in Proof Writing	Clarity, Looks Good, End of Proof, Wording	4	5.9%
6	Created / Copied / Reiterated	Produced	14	20.6%
7	Detail / Explanation	Detail, Explanation	5	7.4%
8	Variables / Parameter/ Symbols / Notation	Symbols, Parameter, Notation, Variables	0	0.0%
9	Proof Type	Direct, Induction, Contrapositive	4	5.9%
10	Justification / Warranting	Warranting	1	1.5%
11	Holdover Knowledge	Previously, Taught	1	1.5%

It was precisely these categories in conjunction with the CVI framing that lead to the creation of the set of distractors for the SCAP. For item P1, most of the comments students made concerned Category 1 dealing with the logical and structural elements of the argument, Category 2 a basic understanding of the argument, Category 4 dealing with mathematical concepts which were important for the argument, and Category 6 which entailed students creating an entirely new argument or reiterating ideas from the argument itself. As this assessment is concerned with validity and not comprehension, Category 2 was always left out of the distractor creation process. Similarly, Category 6

was left out as it aligns with a portion of the framing of the proof comprehension test (Mejía-Ramos, et al., 2018). Additionally, as many of the comments for Category 1 were correct as the argument does have the CVI issue of AC, this category was partially removed as a means for creating distractors.

The treatment of the key for P1 was different than most because many of the comments in Category 1 separately referenced that the argument first began by assuming the conclusion, and then ended by showing the antecedent. Additionally, from the efforts to verify the CVI framing, mathematicians often stated that once they realized that the argument began with the conclusion, they stopped reading the argument as they knew then that it was invalid. From these two pieces of information, it felt consistent to include the key idea of AC as two separate responses, specifically, "the argument begins by assuming $x + 4$ is odd," and "the argument concludes by showing x is odd." As this was not the common structure for the SCAP, it presented an opportunity to insert a novel structure into the test helping to safeguard against students learning to take the test.

As it was the desire for students to see only four options for *why* questions, this left the possibility of two distractors. Since the three largest categories from the open pilot were eliminated for P1, this meant the lesser categories would have to be used, and importantly those categories needed to align with the CVI framing. As such, distractors were selected dealing with Category 4 in conjunction with Category 10, and Category 1 (see Table 27).

Table 27

Distractors, open survey categorization, and CVI framing for P1

Distractor	Open Survey Categorization	CVI Framing
The definition of odd was used incorrectly.	• Category 4 • Category 10	W
The argument both assumes and shows that $x + 4$ is odd.	• Category 1	CR

This process was undertaken for the entire set of twelve items from the open-ended survey except for P8 and P9. In the end, the decision was made to cut these two items for several reasons. First, many of the open pilot students commented on the length of the time it took to validate and comment on each argument. From this feedback it became clear that a few items needed to be removed. To make the truncation easier, I designated that the 6 highest performing items from the CVI framing could not be cut, thus items P1-P6 were excluded from consideration. This left the bottom half, P7-P12 to cut. Items P11 and P12 were both warranting issues, but they were *ω1* and *ω3* issues which made it hard to cut either. I wanted two valid items, thus P10 could not be cut which left P7-P9. Since P1 is an AC item I felt comfortable cutting P8, and in similar fashion, because P7 had slightly better numbers from the CVI framing, and because item P9 dealt with modular arithmetic, a topic which some students said they were unfamiliar with in the pilot, I chose to cut this item as well.

Additionally, item P7 represented an interesting opportunity to introduce variety into the SCAP. This item was an argument for the proposition in Figure 26 about an equivalence relation; thus reflexivity, symmetry, and transitivity were shown in the argument. This meant the argument was lengthy and therefore time consuming for students to read, let alone validate, a problematic issue for a low-stakes assessment. To

offer a solution to this, I truncated the proposition as shown in Figure 26 so the argument focused only on reflexivity and transitivity which included the portion the original CVI issue.

Proposition: Let R be a relation defined on $\mathbb{Z}$ by $a\ R\ b$ if and only if $a + 4b$ is divisible by 5, then R is an equivalence relation on $\mathbb{Z}$

Proposition: Let R be a relation defined on $\mathbb{Z}$ by $a\ R\ b$ if and only if $a + 4b$ is divisible by 5, then R is both reflexive and transitive.

Figure 26. Item P7 was an argument for this proposition which was lengthy and thus on the SCAP two version appeared. The first was the full argument, the second for a truncated version of the proposition and argument.

Semi-Closed Assessment Pilot Results – Reliability

It is important to note the assessment pilot was an anchored assessment meaning that some participants saw one version of the assessment while others saw a different version, though all participants saw the same set of five testlets; T1-T3, T6, and T7. Though a grounded assessment, I first present the overall results of the IRT analysis assuming a non-anchored structure. Below in Figure 27 are the characteristic curves from the LTM analysis with the loadings for discrimination and difficult in Table 28 as well as percentage breakdowns of performance on each testlet in Table 29. The items as a complete grouping accounted for a moderate spread of ability, where difficulty ranged from $-0.7764 \leq d \leq 1.6590$. Notably, only testlet T7 had a discrimination lower than the requisite 0.5 at 0.4874 while testlet T5 had an inflated discrimination of 15.5592. All other discriminations were above the required 0.5.

Exploring the information curves presented in Figure 27 with testlet T5 present is fruitless as information in the two-parameter model of IRT is based on the square of the discrimination. This meant that T5 had an extraordinary peak information of 60.52218 at an ability of 0.4590. While an over simplification, Figure 28 removes T5 and presents the information curves giving a sense of overall reliability that is higher for lower abilities

$[-2,0)$ and slightly lower for higher abilities $(0,2]$. Due to T7's low discrimination, its peak information is nearly sustained for the entire range of abilities.

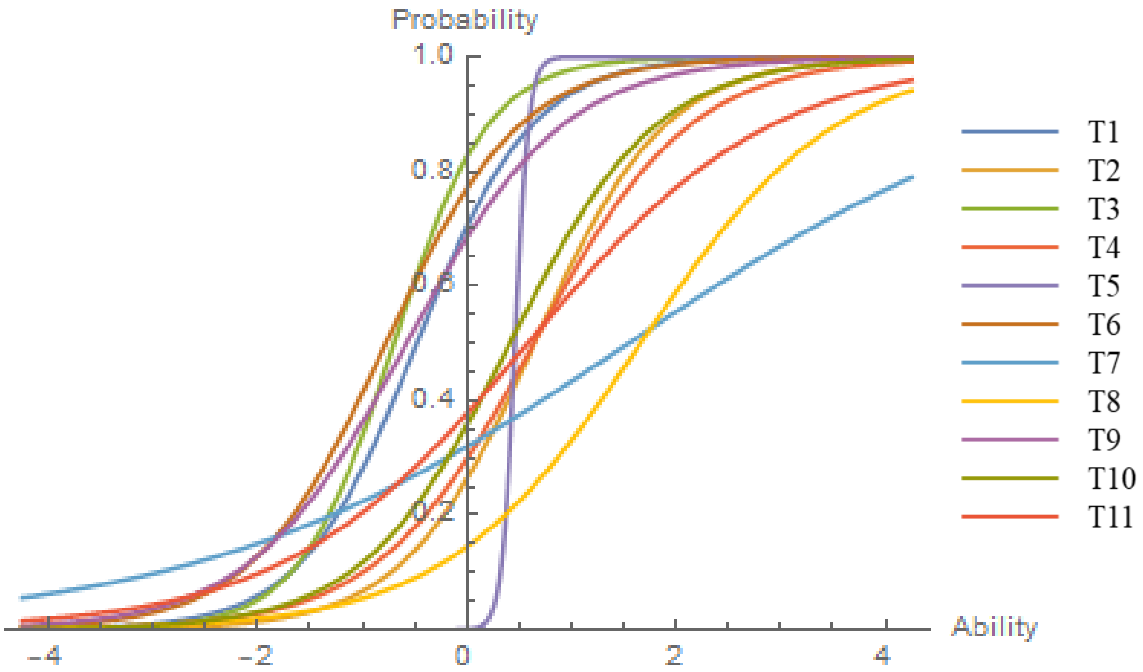


Figure 27. Item characteristic curves assuming complete data – non-anchored analysis.

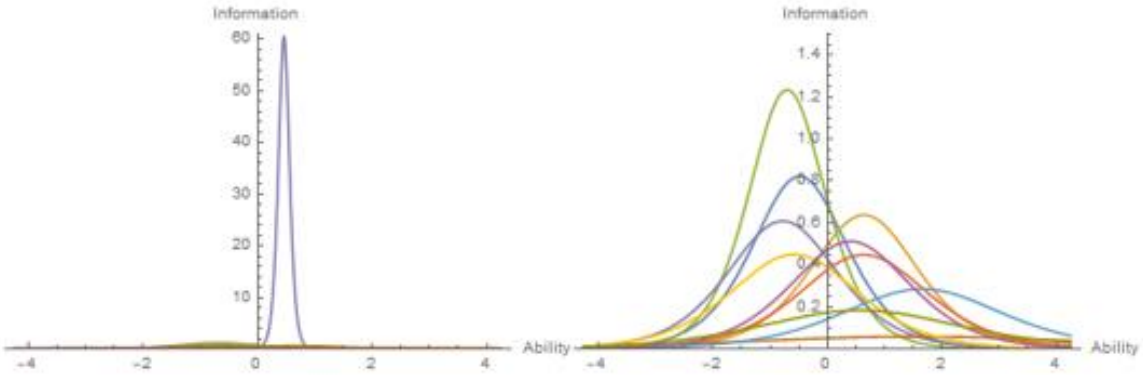


Figure 28. Left – Item information curves assuming complete data – non-anchored analysis. Right – Item information curves with testlet T5 removed.

Table 28

Testlet difficulty and discrimination loadings

Testlet	Difficulty	Discrimination	Peak Information	Pilot Item	Validity
T1	-0.4803	1.8135	0.822196	P4	Invalid
T2	0.6435	1.5979	0.638321	P1	Invalid
T3	-0.6997	2.2218	1.234099	P5	Valid
T4	0.372	1.3402	0.449034	P7	Invalid
T5	*0.4590*	***15.5592***	***60.52218***	*P7*	*Invalid*
T6	-0.7764	1.5616	0.609649	P10	Valid
T7	*1.5452*	***0.4874***	***0.05939***	*P3*	*Invalid*
T8	1.6590	1.0693	0.285851	P6	Invalid
T9	-0.5783	1.3431	0.450979	P2	Invalid
T10	0.4080	1.4315	0.512298	P11	Invalid
T11	0.5799	0.8593	0.184599	P12	Invalid

Table 29

Testlets breakdown of percentages by score

Testlet	CVI Validity	*N*	Scores by %		
			-1	0	1
T1	Invalid	187	24.1%	12.8%	63.1%
T2	Invalid	187	57.2%	10.7%	32.1%
T3	Valid	187	27.8%	1.1%	71.1%
T4	Invalid	94	50.0%	16.0%	34.0%
T5	Invalid	92	44.6%	20.7%	34.8%
T6	Valid	183	28.4%	1.1%	70.5%
T7	Invalid	183	60.1%	7.1%	32.8%
T8	Invalid	90	62.2%	23.3%	14.4%
T9	Invalid	93	26.9%	5.4%	67.7%
T10	Invalid	88	46.6%	15.9%	37.5%
T11	Invalid	95	37.9%	21.1%	41.1%

With a discrimination of $d = 15.5592$, it first seemed as though something had gone wrong with testlet T5 as such a high discrimination is not expected on human testing. This prompted further investigation, including an analysis of T5 scores against

overall scores, as reported in Table 30. The result of this analysis suggests that testlet T5 is performing very well, and that the calculated discrimination is a fair understanding of this testlet's performance in relation to the other testlets from the assessment. Students who received a -1 (i.e., 100% incorrect) for this testlet overall did poorly on the assessment whereas students who received a 1 (i.e., 100% correct) did well on the assessment, and those students who received a 0 (i.e., corrected) on the testlet were somewhere between. This implies that testlet T5 is an item that should be carried forward into future data collection for this instrument.

Table 30

Testlet T5 performance against overall performance[37]

	Total Score																
T5 Score	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
-1	4	3		1	5	4	8	4	5	5	2						
0			1	2		1	5	1	3	1		1	1	2	1		
1											4	2	7	1	14	1	3

The baseline reliability for the SCAP is not a straightforward calculation from a CTT standpoint as there is a large amount of missing data when not accounting for the anchored nature of the assessment. In order to get a general understanding, scores for testlets which were paired and presented at random to students were merged into a single variable in SPSS so that the calculation for Cronbach's alpha could be performed (see Table 31). This process meant taking testlets T4 and T5 – testlets which students randomly took exactly one of during the assessment – and merging their scores into a single variable R1, and similarly for the testlet pairs T8 and T9 into R2, as well as T10 and T11 into R3 as described in Table 31. Having done this merging, the score for reliability for the assessment was $\alpha = .723$ as presented in Table 32. This statistic is only

[37] To clarify the pattern of responses, cells that are left blank have a value of zero (0).

bolstered by the removal of testlet T7 to $\alpha = .744$ (see Table 33), though in either case these reliability scores are both in the desired range. From this rudimentary analysis it seems possible that the assessment is reliable in its measurement.

Table 31

Testlet randomized pairs

Merged Variable	Testlets	Split Responses
R1	T4	94
	T5	92
	Total	**186**
R2	T8	90
	T9	93
	Total	**183**
R3	T10	88
	T11	95
	Total	**183**

Table 32

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0.723	0.725	8

Table 33

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
T1	0.16	14.050	0.473	0.268	0.685
T2	0.80	13.884	0.445	0.217	0.689
T3	0.14	13.668	0.495	0.299	0.679
T6	0.15	14.236	0.397	0.194	0.699
T7	*0.84*	*15.629*	*0.174*	*0.042*	***0.744***
R1	0.69	13.367	0.548	0.308	0.668
R2	0.60	13.792	0.451	0.223	0.688
R3	0.60	14.473	0.359	0.156	0.707

Predictive validity. From a predictive standpoint, there are several factors that point to the assessment working in a desirable way. First, the number of proof-based courses both taken previously and enrolled in at the time of taking the assessment are positively correlated with better scores on the assessment (see Table 34). Thus, the more proof-based classes a student took or was taking the better they performed on the assessment. This outcome is reassuring as it seems possible that the more a student encounters proof and proving generally, the better they are at identifying arguments which meet the standards they have incorporated in their proof schemas. Also, this group of students most likely have better defined proof schemas, giving them a better understanding of validity in general. Similarly, student's GPA, though self-reported, were positively correlated with scores on the assessment.

Table 34

Item Pearson correlation with assessment scores

Item	Pearson Correlation	Effect Size[38]
Number of Proof Courses Taken	.235**	0.0585
Number of Proof Courses Currently Enrolled	.215**	0.0485
University GPA	.200**	0.0416

** Correlation is significant at the 0.01 level (2-tailed).

When grouping students by their majors, two groups had significantly different results from their counterparts as show in Table 35. Students who indicated they were pure mathematics majors ($M = .88$, $SD = 3.415$) did significantly better than those who did not indicate they were pure mathematics majors ($M = -1.02$, $SD = 3.879$), $p = .001$. Conversely, students who identified as mathematics education majors ($M = -2.31$, $SD = 2.706$) did scientifically worse than those who did not identify as mathematics education majors ($M = -.13$, $SD = 3.784$), $p = 0.46$. Finally, students who identified as applied mathematics majors had no significant different from their compliment group.

Table 35

T-test – Mathematics majors mean assessment scores[39]

Major – Mathematics	Specification	N	Mean	SD	E.S.[40]	Sig. (2-tailed)	95% Confidence Interval	
							Lower	Upper
Pure	Major	73	.88	3.415	.519	.001	.798	2.990
	Not Major	114	-1.02	3.879				
Applied	Major	65	-.80	3.576	.213	.172	-1.951	.351
	Not Major	122	.00	3.914				
Education	Major	13	-2.31	3.706	-.582	.046	-4.325	-.038
	Not Major	174	-.13	3.784				

[38] Cohen's f^2 effect size – ≥ 0.35 Large; ≥ 0.15 Moderate; ≥ 0.015 Small.
[39] Keep in mind that mean scores reflect an overall possible range from -8 to 8. Mean values close to 0 are both possible and likely.
[40] Cohen's *d* effect size – ≥ 0.8 Large; ≥ 0.5 Moderate; ≥ 0.2 small

The differences in sample clustering based upon classes is reported in Table 36. Students enrolled in abstract algebra ($M = 0.5$, $SD = 3.907$) performed better than students in the compliment group ($M = -0.77$, $SD = 3.681$), $p = .027$. Similarly, results indicate a significant outperformance by students enrolled in abstract algebra ($M = 0.98$, $SD = 3.683$) over those not enrolled in abstract algebra ($M = -0.84$, $SD = 3.743$), $p = .002$. Students enrolled in topology ($M = 1.71$, $SD = 3.869$) outperformed those who were not enrolled in a topology course ($M = -0.48$, $SD = 3.758$), $p = .024$, and interestingly, having previously taken topology ($M = 0.5$, $SD = 3.907$) had a moderate effect size ($ES = 652$), but the group did not outperform their counterparts ($M = 0.5$, $SD = 3.907$) in a significant way, $p = .054$. This last outcome is most likely the results of a small sample size ($N = 9$) of students who previously took topology. On the other hand, students who were enrolled in an ITP course ($M = -0.95$, $SD = 4.136$) did significantly worse than students who were not ($M = 0.89$, $SD = 4.121$), $p = 014$. This implies that students who were newest to proof scored lower on the assessment than established students. No other class had a significant difference in mean scores from their counterparts including former or current number theory and analysis/real analysis students. This analysis implies that students who had taken part in more – as well as more rigorous and/or difficult – proof classes performed better on the assessment while students who were truly novice with regards to proofs performed worse. These outcomes are consistent with what would be expected and only add to the case that the assessment is a reliable measure.

Table 36

T-test - Classes mean assessment scores

Class	Specification	N	Mean	SD	E.S.	Sig. (2-tailed)	95% Confidence Interval	
							Lower	Upper
Algebra	Taken	72	.50	3.907	.335	.027	.148	2.383
	Not Taken	115	-.77	3.681				
Algebra	Enrolled	58	.98	3.683	.490	.002	.666	2.989
	Not Enrolled	129	-.84	3.743				
Analysis	Taken	72	.32	3.801	.256	.090	-.152	2.095
	Not Taken	115	-.65	3.784				
Analysis	Enrolled	117	.05	3.859	.233	.127	-.252	2.011
	Not Enrolled	70	-.83	3.687				
Num. Theory	Taken	25	.52	4.073	.234	.262	-.693	.2535
	Not Taken	162	-.40	3.766				
Num. Theory	Enrolled	6	-.83	4.579	.136	.718	-3.700	2.552
	Not Enrolled	181	-.26	3.795				
Topology	Taken	9	2.11	3.919	.652	.054	-.039	5.059
	Not Taken	178	-.40	3.775				
Topology	Enrolled	17	1.71	3.869	.574	.024	.292	4.073
	Not Enrolled	170	-.48	3.758				
ITP	Taken	117	.45	4.201	.033	.834	-1.382	1.117
	Not Taken	70	.59	4.175				
ITP	Enrolled	39	-.95	4.136	.446	.014	-3.298	-.369
	Not Enrolled	148	.89	4.121				

Conclusion. While this analysis represents an incomplete analysis, it does indicate that overall the testlets performed well together to assess a variety of ability levels and in a reliable manner, save for testlet T7. To complete the statistical analysis, it is important to look at the various forms that arise from the anchored structure of the

assessment. As was mentioned in the framing, since there are three pairs of random testlets this means there are a total of 8 different assessment forms (see Table 9).

Before presenting the analysis of each form, it should be noted that because there are 8 forms, this means each form on average has a sample size of around 23, with one as low as 17. This coupled with the fact that there are only what amounts to 8 items in each form means each form is well below what is required for accurate parameter estimation in IRT-based test development (Şahin & Anil, 2016). Thus, it is important to take the following analysis in context as more data is required to better understand the behavior of the assessment in its anchored arrangement.

Anchored Analysis in Brief

Due to the fact that the sample size and test length requirements have not been met for any of the 8 forms, I will not present the individual loadings for all testlets in all forms but will instead present the characteristic curves for each form and point out any abnormalities from the analysis (see Figure 29). Employing the eye test on the characteristic curves, there is no one form that looks perfect, though there are forms which look better than others as well as those that look much worse. Forms 3 and 5 are perhaps two of the worst looking as testlets T7 and T10 in Form 3 and T7 and T8 in Form 5 had negative discriminations. Once again this raises a question about the effectiveness of T7. This anchored testlet appeared in all 8 forms but in only 4 forms had a discrimination above the required 0.5, twice by the slightest of margins – Form 4 with a discrimination of 0.55939514 and Form 8 with a discrimination of 0.51207391. Additionally, though there are no testlets in Form 8 which have negative discrimination, three testlets, T2, T3 and Q6 all had lower than desired discriminations. As half of the

items in Form 8 were performing at or near suboptimal levels, it is no surprise then that this form also had the lowest alpha score at $\alpha = 0.522$ (see Table 37), which is outside the desired range of $0.6 \leq \alpha \leq 0.9$.

On the other hand, there were at least two promising forms as Form 4 and Form 6 performed well with all 8 testlets in each item having positive discriminations and reasonable coverage of ability. Form 4's difficulty range was $-0.47796278 \leq d \leq 1.19052788$ and Form 6's difficulty range was $-1.52922547 \leq d \leq 1.17473058$. These two forms also had the highest alpha scores, with $\alpha_4 = 0.847$ and $\alpha_6 = .804$ respectivly.

Table 37

Form reliability

Form		Cronbach's Alpha	N of Items
Name	$\{T4, T5, T8, T9, T10, T11\}$		
Form 1	{1,0,1,0,1,0}	0.672	8
Form 2	{1,0,1,0,0,1}	0.707	8
Form 3	{1,0,0,1,1,0}	0.624	8
Form 4	{1,0,0,1,0,1}	0.847	8
Form 5	{0,1,1,0,1,0}	0.666	8
Form 6	{0,1,1,0,0,1}	0.804	8
Form 7	{0,1,0,1,1,0}	0.668	8
Form 8	{0,1,0,1,0,1}	0.522	8

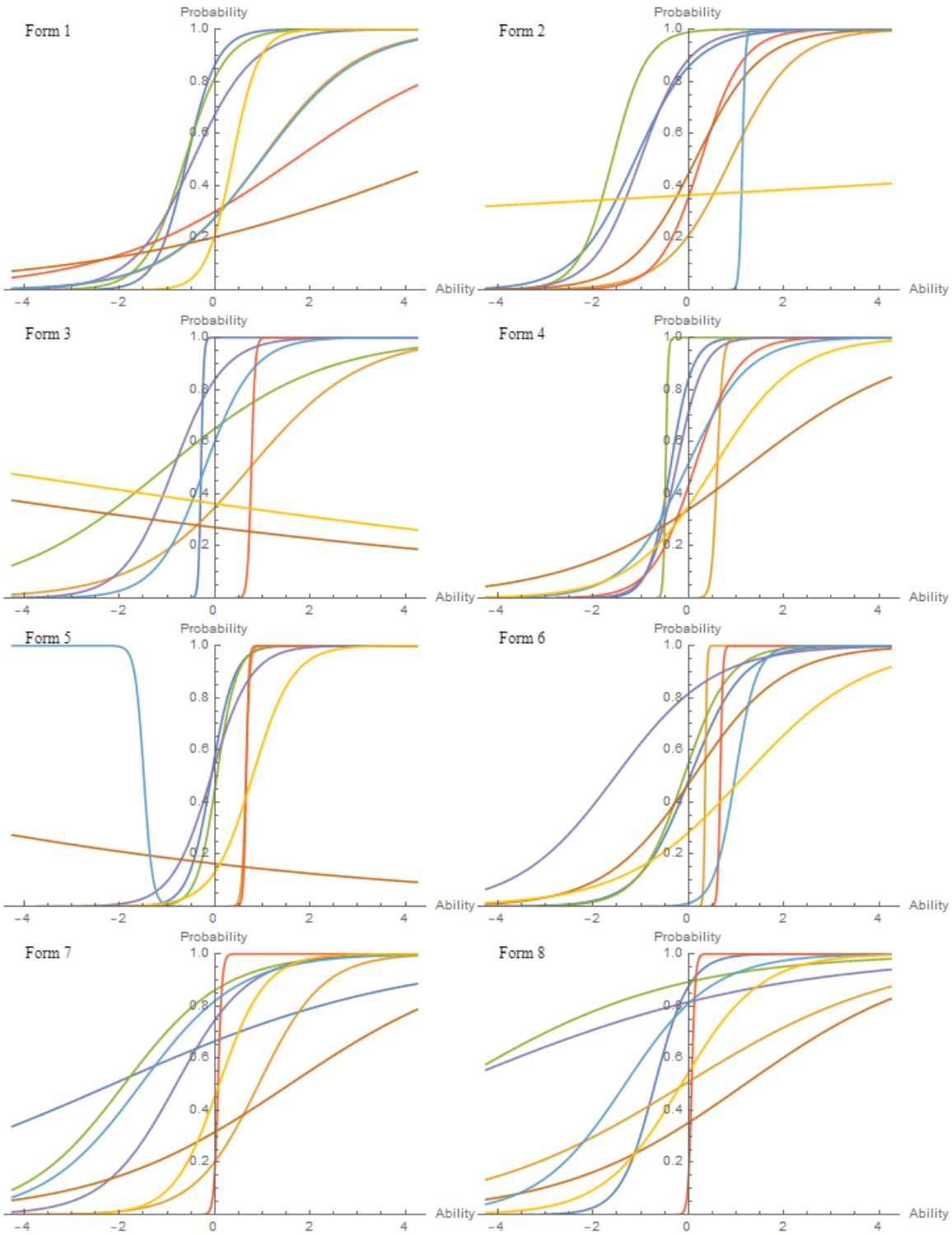


Figure 29. The characteristic curves for all 8 forms of the anchored assessment.

Conclusion. While more data would help identify a specific form, between the overall analysis, the predictive analysis, and some of the form analysis, the case begins to

build that in pilot form, the assessment is a reliable measure. It should be noted that considering the complicated structure of the assessment, having an overall reliability of $\alpha = .723$ is a major feat and strongly hints at an assessment which is consistent in its measurement. Adding to the reliability argument, when compared to the validating process for the Proof Comprehension Test by Mejía-Ramos, et. al (2017), they were able to generate reliability scores for their three versions of $\alpha = .71$, $\alpha = .74$, and $\alpha = .72$, all of which are close to the scores obtained in this study. Considering that their sample size was similar, but their assessment had significantly more items, this only bolsters the results concerning reliability for this instrument even in its incomplete form. What is left next is to determine if the assessment is measuring validity as is the original claim.

Measurement Validity – Student Interviews

The student interviews which proceeded the assessment piloting helped identify the process by which students went about validating from moment to moment as they proceeded through the assessment (see Table 11). Again, the process is the moment-by-moment approach that the students took in verbalizing how they viewed the argument. The process could take on many different colors, from performing line-by-line checks of the argument to simply reading or restating the argument. The latter process might indicate the student was doing something more akin to the action of proof comprehension as restating is comparable to what Mejía-Ramos et al. (2012) called summarizing via high-level ideas. However, line-by-line checking aligns more with the action of validating (Inglis & Alcock 2012; Selden & Selden, 2003). By identifying what processes students took, it should be clear whether they were in fact genuinely validating the arguments, and thus the data would corroborate that the instrument itself is valid. It is

possible students were doing some other activity like, for instance, comprehending the arguments[41], though this is not necessarily problematic so long as they did in fact genuinely validate the arguments throughout the course of the interview. Additionally, the data presented helps clarify why students might have received the scores they got from the assessment.

In order to simplify the process of understanding the results of the interview, I present the findings with regard to testlets T1 and T7 (see Figure 30 and Figure 31). To be clear, this does not mean the analysis is wholly about these items, but rather that each example will be taken from the interview where it concerned the arguments for these testlets. I do this because, (1) these both were anchored testlets meaning that all students responded to these testlets while taking the assessment, and most importantly (2) the diverse yet characteristic responses that occurred to these prompts paint a clear and representative picture about the ways students went about taking the assessment. The results generally follow the patterns set out by these dual examples, so it does not feel artificial to present the results in this fashion.

Proposition: $\{x|x \in \mathbb{Z} \text{ and } 18|x\} \subseteq \{x|x \in \mathbb{Z} \text{ and } 6|x\}$.

Argument: Suppose that $a \in \{x|x \in \mathbb{Z} \text{ and } 18|x\}$, from this we have that $a \in \mathbb{Z}$ and $18|a$. As $18 = 6(3)$ and since a is an integer and is divisible by 18, this implies that $a \in \{x|x \in \mathbb{Z} \text{ and } 18|x\}$. Therefore, we have shown that $\{x|x \in \mathbb{Z} \text{ and } 18|x\} \subseteq \{x|x \in \mathbb{Z} \text{ and } 6|x\}$.

Figure 30. The proposition and argument for testlet T1.

[41] Not only is it possible, but it is also highly likely that students were involved in comprehending activities throughout the assessment process. It only makes sense that students might first try to understand the argument before validating them. Thus, the larger question should therefore be focusing on whether students *eventually* engaged in genuine validating activities or not.

Proposition: For arbitrary set A, B, and C;

$$(A - C) - (B - C) \subseteq A - B.$$

Argument: Let $x \in (A - C) - (B - C)$. That means that $x \in A$ and $x \notin (B - C)$ which implies that $x \notin B$. Therefore, we can conclude that $(A - C) - (B - C) \subseteq A - B$.

Figure 31. The proposition and argument for testlet T7.

As I read through the interviews and the coding of the interviews, there are distinct groups of interviewees regarding what they were reporting on during their validating process. In sections to follow, I present these groups, their actions, processes, and my analysis of each. For brevity I will not present all six-validating processes for T1, and T7 – twelve total validating processes – but instead will pick those which are representative of the individuals and their groups.

Group 1. First there were students like Al, Gerald and John (see Table 38). These three individuals all had an intuitive understanding of the how a proposition affects the consistency of an argument and were able to rectify that understanding with the argument they were seeing. Al, Gerald and John's process most typically stemmed from their intuition which lead them to check things like the proof framework or do a line-by-line check of the argument, though on a few rare occasions, verbalized their understanding was a part of their process as well. In those instances where their processes were that of comprehending, they typically would quickly resolve their comprehension issues and return to a process of intuition-led checking of the argument. This meant most often these students were in the action of authentically validating the arguments. For instance, Al had the following to say when validating the argument for T1,

> Okay. Let's see. Yeah, so the first thing I see is that we're trying to prove a set is a subset of another set. In my head, I'm looking for things like, are they taking an arbitrary element from the first set and showing that that arbitrary element's going to be in the second set? Other things I might be looking for is whether or not

> they're applying the properties of being in that first step properly. So when I was reading this argument, they started out the way I would want and I'm thinking, okay, this is looking good so far, they're taking that arbitrary element from the first set, but I think then they kind of veer off course a little by showing that it's an element of the same set that they're taking the element from. They're not showing it's in the second set.

In this excerpt, Al started by stating what he was looking for based upon what he knew about arguments that dealt with set containment, a key feature of the argument in this T1. This understanding was instrumental in his analysis of the argument and lead him to conclude that the issue with the argument had been because of its circular nature, the correct option for this testlet. This was a common practice for Al, Gerald and John, where they would make a statement of what they were generally looking for and then present an analysis of the argument based upon this understanding.

Table 38

Group 1 assessment interview results

Student	Form	Process	Action	Score
Al	4	Intuitive proof framework and line-by-line check	Validating	8
Gerald	4	Intuitive proof framework and line-by-line check	Validating	6
John	4	Intuitive proof framework and line-by-line check	Validating	6

These students had specific things they expected to see within the argument, based upon what they thought the proposition implied about any argument for said proposition. Then during validation, they went about checking for these things. The quality of these checks might be thought of in terms of checking the logical structure and flow of the argument, or as Selden and Selden (2003) referred to as checking the various proof frameworks. If they found what they were seeing in the argument aligned with what they felt the proposition required, they would then move on to check other aspects

via a line-by-line check, or they might have attempted to do this all simultaneously. For instance, in checking the proof frameworks, they might also check the consistency of the algebraic manipulations or the quality of the justifications for claims, or, as in the previous example with Al, check that the overall logical structure supported showing one set is a subset of another, or called "element chasing" by the students. Al checked that the element which was selected for the element chasing process was an arbitrary element, thus part of his procedure for checking was somewhat simultaneous, or at the very least, could not be broken apart from the interview. This line-by-line checking was described by Inglis and Alcock (2012) as a common approach by both the mathematicians and students in their study, but it was the mathematicians who seemed to have more between line checks than their novice counterparts.

From the analysis of their verbalized processes coupled with their scores, Al, Gerald, and John were mostly going about the process of validating the arguments as opposed to some other process. Without even being introduced overtly to the CVI framework, they were authentically checking ideas of comparable nature to the categories from this framing. Within this group of three, there were examples of deviation, as all three individuals had instances where they were not fully engaged in validating. For instance, Al and John on T7 and Gerald on T10 demonstrated a pattern of validating mingled with that of argument comprehension. This can be clearly seen as Al describes his process for T7, saying,

> They start off all right by assuming ... By taking the arbitrary element from the first. Then they get a bit confused. So they say x is not an element of $B - C$. That's true. That's what... That's true. But then they say that implies x is not ... What now? ... x is not an element of B? ... What went wrong here? There's a lot of double negations. What went wrong here? So, x is an element of A ... so x is an

> element of A subtracted by C, and x is not an element of B ... C? Well, if x is not an element of this, that means x is ... Yeah. Give me one second here, I'm just…

In this first portion, Al grappled with the argument for some time as he both attempted to validate and to understand the argument. He certainly makes the statement, "What went wrong here?" indicating he is working toward validating the argument, but what follows is more along the lines of Mejía-Ramos et al. (2012) called "Meaning of terms and statements" where Al is restating portions of the argument in a "different but equivalent manner" (p. 8). During the interview he struggled at first to narrow in on any one reason he felt caused the argument to be invalid[42] without first taking some time to further comprehend the argument. He continued his validating attempt after a lengthy pause,

> Oh okay. All right. Yeah, so okay, so that means x ... x is not an element of B or x is an element of C ... Yeah, this is where it went wrong. Okay. So, when we say ... So if I haven't gone off, or if I remember correctly, x being an element of $(A - C) - (B - C)$ means that x is an element of $A - C$ and x is not an element of $B - C$. So they're okay there. But when we say x is not an element of $B - C$, that means x is not an element of B or x is an element of C. So that's where they're not ... That's where I said it's invalid. They're not considering both cases.

In the end, Al's analysis was spot on, but it took him some time to better understand the argument. His cycle of argument comprehension seemed to lead him to make a correct validation judgment. This process was similar for Gerald and John. During the interview as they were able to identify correctly what validity issues had occurred in T7 and T10 respectively, even though in some instances, they were not able to do so on the assessment itself[43].

---

[42] During a portion of the interview, students were asked what they were attending to and thinking about while validating. They were only seeing the argument and not the possible reasons why the argument might be invalid.

[43] As was noted in the methods, all interviewees were given access to their completed assessments prior to the interviews and asked to review them. This means it is possible students could have found new meanings while reviewing and those new meanings became the "facts" they presented in the interviews. Though this may be the case, Al, Gerald, and John seemed to be the group that most gained from this as the other groups still had difficulties identifying validity issues during the interviews.

Al, Gerald, and John not only went through a phase of argument comprehension a few times, but, in most cases, they did not verbalize this process; this is true of most of the interviewees. Typically, these three participants thought to look for and tackle the validation process by orally referencing that they were checking the proof frameworks, logical flow, justifications, algebraic manipulations, use of parameters and notation, and other issues typically tied to validity. They characteristically did so without making statements about their manifest understanding of each step of the argument or the argument as a whole. This is in sharp contrast with some of the other students interviewed for this study, especially Shannon. Only in a few rare cases did these three deviate by verbalizing their process of understanding the arguments. As it turns out, for Gerald and John, two out of the three arguments they lost points on during the assessment were also arguments they made explicit verbalized attempts to comprehend during the interviews. Comprehension is not a bad process to undertake during validating, in fact, it seems almost impossible to validate something one does not understand. What this might suggest is that when students struggle to comprehend, they might be using comprehension – or even lack of comprehension – as a que for their validation judgments. While comprehension is not the worst replacement construct for validating, it certainly is not an exact replacement and can lead to incorrect judgements.

Group 2. Brent (see Table 39) was similar to the first group in that he had an idea of what the proposition implied about an argument, but, unlike the first group, he often had difficulty reconciling his own idealized approach with the approach taken in any given argument. His ability – or indeed lack of ability – to transfer his intuition to the process of validating hampered

his ability to correctly validate the arguments in this assessment. Many of his validating attempts included statements about what he would have done but those statements rarely connected to the task of actually validating the argument. His process was at times a little messy and can best be described as non-resolved intuition often followed by line-by-line checking of the argument. Most often the action Brent undertook was that of validating the argument, though he made overt attempts at proof comprehension as well. For instance, when validating the argument for T1, he went through the following process:

Interviewer: So yeah, what were you looking at? What were you thinking about?

Brent: Well, I mean, basically, whatever we call that quotient, three times that quotient is also an integer, and that's what you would get from doing it with six. So ...

Interviewer: Okay.

Brent: If you were doing it like, constructively, for each a that's how you would do it. But I think there's something where it was true, but one of the steps is wrong?

Interviewer: Okay.

Brent: a is an integer [inaudible] divisible by ... solve for, okay... [long pause] No, no, this ... should it go the other way? Okay, I could have misread it such that it was 18 going into x rather than vice versa. But ... No. 'Cause that would be valid. So, yeah, I think it's just what I said earlier.

Here Brent starts with an explanation of his understanding of what he would do, though he does state it would be a laborious construction, implying it was not fully constructed in his mind, and he was unwilling to share more than he did in the exchange presented above. Interestingly, it seems as though he did not actually look at the argument until his final line of dialog in this exchange where he grapples with what $18|x$ means, but then hastily concludes something he previously reported is why the argument is invalid, though it is unclear what that thing is.

Table 39

Group 2 assessment interview results

Student	Form	Process	Action	Score
Brent	4	Non-resolved intuition often followed by line-by-line checking.	Validating – w/ minimal amounts of comprehension	1

In the end, Brent was unable to tell why the argument was invalid without being allowed to see the set of options, at which point he said, "Oh, right, that's right, so our desideratum would be x is congruent to six, so that's the wrong line, because ... yeah, we just did a circular thing." Whether his original rational about the argument's validity was an intuition or a guess is unclear, but during the assessment, he was able to select the correct answer for why the argument was invalid. His statement here was partially nonsensical as the argument had nothing to do with being congruent to six – this could stem from his difficulty with the expression $18|x$ – but certainly there was circularity in the argument. Ultimately, Brent had a difficult time conveying his rationale and even had difficulties grasping meaning in a few key situations. However, he had no shortage of intuition about how an argument should be accomplished.

A similar idealized exchanged happened when he commented on T7. Comparable to his process for T1, Brent makes a statement about the proposition – it being true – and stating how he would have gone about the proof:

> I'm pretty sure this is, in fact, true, but it might not be … how I would have done it. But yeah, basically the thought process is, the points in C are irrelevant because it's excluded from both, so those are just out of our consideration. To prove that you're a subset of $A - B$, you just have to prove that, you know, your arbitrary element is in A, 'cause it's not in B, so that should follow. I don't know ...

Once again, Brent's first process was that of an idealized and ultimately unfulfilled argument. In some ways his intuition about the process was correct as the set C becomes

the lynch pin for the argument in T7, but it does not seem like his comments are even part of exploring the argument so much as exploring his own mental image. In short, Brent seemed to have a disjoint union between his idealized argument and the argument he was validating. He often had ideas or intuition about how he would go about proving a proposition, some of which were very different – though insightful – from what was being presented, but in the interview, he was never able to reconcile his image with the argument he was validating. Ultimately, this disjoint union of idealized and actual argument was the cause for Brent's score of 1 on the assessment.

Group 3. The final students, Christopher and Shannon (see Table 40) differed from this first two groups in that they never verbalized any sort of intuition, insight, or expectation that they might have had about the argument they were about to validate. Instead, they simply dove into the procedure of validating each argument. Their actions for each argument was that of validating, but Shannon's process consisted of spending a considerable amount of time during the interview in verbalized comprehension as she restated the entire argument – emblematic of summarizing via high-level ideas (Mejía-Ramos et al. 2012) – before making any sort of validity judgement. Christopher's process for validating consisted of strict line-by-line checking where he started at the beginning of the argument and checked that it "made sense" and then continued through each portion one line and validity judgement at a time. For instance, Christopher's verbalized process for T1 went like this:

> When I was looking at this, I saw that it said, "Suppose a is in set of x such that x is in the integers and that 18 divides x" and I was like, that's good. And from this, we have the "a is an integer, and 18 divides a;" I was like, that's also good. And then when we got to this part, when it says that, "this implies that a is in the set of x such that x is in the integers and that 18 divides x," that wasn't the conclusion that we wanted to get to. And so, that's kind of a problem.

As was indicated before, Christopher's and Shannon's processes were a methodical line-by-line checking of the argument. He began with the first statement and identified it was a reasonable start, and then moved on to the first implication stating it was reasonable, and, in characteristic fashion, walked through each instance until he reached something he did not agree with. In this case it was precisely the circularity which existed within the argument.

Table 40

Group 3 assessment interview results

Student	Form	Process	Action	Score
Christopher	4	Line-by-line checking	Validating	2
Shannon	1	Line-by-line checking and probing for local and holistic understanding	Validating – w/ a considerable amount of proof comprehension	-2

The only real difference between Christopher and Shannon on T1 was that Shannon, like Brent before her, had a difficult time recalling the exact meaning of the expression $18|x$. In her process of line-by-line checking, Shannon stated that, "I thought this meant 18 is divisible by x, and there's no remainder left over." This misconception led her off course for a portion of the validating, and then after a brief discussion, she was able to get back on track and complete the validation.

Both Shannon and Christopher had issues validating T7 where they both took time to make diagrams to help them understand the argument better[44]. For example, here is Shannon's verbalized process,

> Shannon: Okay. Let x be contained in $(A - C) - (B - C)$. That means that x is contained in A, so definitely agree with that, and then x is not contained in $B - C$. Let's do … why do I agree with that? I agree

[44] Unfortunately, because the interviews were not done face-to-face, but over an internet communication system, I was unable to get copies of either diagram even though I repeatedly ask both participants to send me a picture of their drawings.

with that because, well, definitely we factor out $B - C$, and there's no, I guess risk that it would be contained in there at all, because it's factored out entirely, which implies that x is not contained in B. So, why does it imply that? Because ... [long pause] x not contained in B? I wonder if I drew a picture for this, or if I did that again?

Interviewer: Yeah, go ahead and draw a picture, please.

Shannon: Okay.

Interviewer: If you do draw a picture, I'm going to ask you to take a picture of it and send it to me, just so I have reference of it.

Shannon: Sure, let me just grab a pencil here. All right. So, I have sets A, C and B. [inaudible] A, C. So, x is in A excluding C, excluding ... So, it's definitely in this area. Also excluding B. Then if this was the ... I'm just talking to myself at this point.

Interviewer: You're fine.

Shannon: Okay, cool. Yeah, here's some pictures. Now I disagree that it implies that x is not contained in B, because I believe that I drew a picture where there's some non-empty intersection between A and B, but B doesn't share anything with A and B. Is it okay to be thinking about intersections in this case?

Interviewer: For sure.

Shannon: So, if it was the case that A and B shared a non-empty intersection, then I would definitely; x is in A, x is not in $B - C$ because that set is empty, but then it's x is in $A - C$ minus … Then that would be the empty set minus the empty set, so that would also not let x be in A? … Okay, actually, maybe it does imply that x is not in B. I'm struggling with this one. But if that was correct, it implies that x is not in B… x is in $(A - C) - (B - C)$. I would think it implies that that's included in $A - B$ because it's in A, and it's not in B.

Shannon's process commences in an emblematic fashion for this third group where she begins a process of line-by-line checking. This process deviates when Shannon begins to struggle with the direct implication of $x \notin (B - C)$ – this was the same case for Christopher. Shannon grappled with this until she concluded that she is fine with the implication that $x \notin B$ and that overall the argument is fine.

Shannon's typical approach to validating – her process – involved restating the argument, what Mejía-Ramos et. al (2012) called summarizing via high-level ideas, followed by some amount of validity judgments about parts of the argument. In the

argument for T7, her process more closely approached line-by-line checking which she also did, but again, she went through a process of comprehension which could be construed as "identifying examples that illustrate a given statement" (Mejía-Ramos et al., 2012 p. 8).

In some ways Christopher is an anomaly as it is difficult to reconcile his performance on the assessment with his interview. He struggled at multiple points on the assessment, ending with an overall score of just 2, but in every case during the interview he was able to correctly validate and justify his validations. One possible reason his interview was better than his performance on the assessment is because he took time to go over his assessment before the interview, as he stated many times during the interview. Despite making them available, most students did not review their assessments before the interview, but Christopher explicitly stated he did. I am not saying Christopher was knowingly or willfully misleading me, but was perhaps giving me something more approaching a Hegelian notion of synthesis (Corbett & Connors, 1999), accounting for his thoughts about each argument after having further reconsidered each argument rather than giving an insight into his process from his time taking the assessment. This is completely understandable, as a few months had passed since the students had taken the assessments. Most students were doing a mixture of giving an accurate account of their assessment validating process, while also giving a synthetic account of their validating processes during the interview.

Conclusion. Ultimately, all three groups were validating as an action, but often did so in very different ways. The first group typically went into validating by first understanding the implications of the proposition and then checking to see if those

implications were met within the argument. The second group also understood the implications of the proposition, but Brent often had difficulty reconciling his understanding with the argument he was validating, forcing him to either take a different approach or to abandon the validating process altogether. Finally, the third group validated by line-by-line checks with periods of construct or argument comprehension included.

As far as the effectiveness of each approach, the first group's approach is the strongest. The first group was able to not only complete the task of validating during the interviews, but they were able do so effectively both on the assessment and in the interviews. Brent in the second group was not yet at the ability of the first and it affected his ability to validate arguments both on the assessment and in the interview, though he did have the intuition which will eventually support the approach the first group took. Finally, the third group never made any clear statement about expectation or their intuition concerning the argument, but went about validating in a very methodical line-by-line process, which in Shannon's case, was accompanied by attempts to comprehend the argument moment-to-moment. This sometimes left them want for what to do next.

VI. Discussion

The main goal of this study was to develop an instrument for measuring ITP students' ability to validate deductive mathematical arguments. This process involved developing a cohesive analytic framework which would act as a basis for constructing said instrument in the form of a multiple-choice, closed-form assessment. The process of developing this framework involved surveys and focus groups of mathematicians to define specific categorizations of common validity issues in the ITP setting. In the sections to follow, I discuss some implications of the findings from this study, any limitations the study or assessment may have, and the implications for future research this study supports.

Implications of Findings, Future Work and Limitations

While the CVI framework is not perfect, as no one categorization had 100% agreement, it is certainly a considerable achievement in understanding what mathematicians count as important factors in determining the validity of an argument. Certainly, the framing is aimed at the ITP level, but so often the mathematicians in the focus groups would take these ideas and apply them in other contexts where the framing seemed to hold up in their estimations. More work would need to be done to make sure the framing is robust enough to be useful in other contexts, but it seems that in its current form, it certainly could be applied to other undergraduate settings, like algebra or analysis. As a research tool, the CVI framing could have practical implications in understanding students' produced proofs in a verity of contexts from interview analysis where producing arguments is the focus, to analysis of work from students in an academic setting.

Universal and contextual features of validity. One of the most striking results of the CVI framing is the modality of each of the categorizations with respect to universal effect versus contextual effect. From the results of both the survey and the focus groups, it seems that the categories of AC and CR have a universal effect on the validity of an argument; these issues always invalidate arguments. The other categorizations, LG, MN, W and WT, have a more social or contextually defined effect, where each instance or circumstance is considered individually with regards to effect on validity. This dichotomization leads to an important takeaway: validity issues are not viewed equally.

The view mathematicians have about validity issues came through very clearly in this study. Most mathematicians agreed that assuming the conclusion and circular reasoning invalidate an argument. Because mathematicians have no tolerance for these types of validity issues, students should be aware of these issues early in their studies and be wary of them in their own writing or in arguments that they read – this should be come through very clearly in instruction to the students throughout an ITP course. During the closed pilot, this appeared to be the case for CR as 63.1% of students got a 1 on T1, and slightly more than three-quarters (75.9%) of all students received a 0 or 1. This means students did relatively well at identifying that an argument with CR in it was invalid, or, at very least, they knew CR had occurred in the argument meaning it was invalid. However, students had a difficult time recognizing the argument for T2 was invalid though it had the issue of AC. Less than half (42.8%) of students scored a 0 or 1 with less than a third (32.1%) scoring a 1. This latter result suggests that Selden and Selden's (2003) estimation that students are no better than chance at validating is an overestimate for something that mathematicians indicate is a universal flaw in mathematical proofs.

This is a troubling result as the students who took this assessment were not just ITP students. Some were students in their final undergraduate mathematics courses. This implies students are either not being taught properly or failing to learn what to mathematicians is a basic tenant of sound proving. Whichever case it is, it seems prudent to right this trend. Future research should look for curriculum and instruction which support students in learning basic and fundamental features of sound argumentation in the mathematical setting, and how to recognize these ideas in their own proof writing and in the arguments that they read.

Because not all categorizations are universal in effect, but instead are contextual or case-by-case examples, the case could be made that these categorizations are less important. On the contrary, I would argue that due to their ambiguous natures LG, MN, W and WT are no less important than their more universal partners – AC and CR. For instance, Alcock and Weber (2005) suggested that inferring and checking warrants as a classroom taught concept is important in validating for real analysis if for no other reason than it enhances students' understanding of proofs, a major goal of undergraduate mathematical instruction. I would further argue that these other three, LG, MN, and WT, as socially defined features of proof, are important in the same regard. It is important for students to have a grasp on what makes a mathematical proof different from arguments in other settings and to delineate what features of arguments are important to the viability of an argument as a proof. This knowledge will also aid students in their quests to understand proof and proving to a greater degree.

Though there is certainly an amount of subjectivity in validating arguments, I feel that this does not take away from the assessment from this study which askes students to

definitively validate arguments. This misalignment was overcome through the consensus process. While not all categorizations are universal in affect, all arguments which were carried forward to the assessment process were agreed to be invalid for the defined reasons by at least 90% of the mathematicians who saw them on the survey. Therefore, even though validity has normative characteristics, the arguments students were being asked to validate were either universally valid or invalid.

Proximity and warranting. Another interesting takeaway is the idea of proximity and warranting in mathematical argumentation. While Alcock and Weber (2005) point out that warranting is an important practice in mathematics and moreover an important part of the validating process, it was a point of consternation as explicit and over warranting was found to be of little use generally to the mathematicians in the focus groups. Their comments imply that explicit justifications were only needed in cases where the argument was dealing with novel ideas. This adjacency to novelty was expressed as a consistent factor of mathematical proof writing whether the context was an ITP class or a peer-reviewed journal article. This idea makes it feel artificial to attempt to assess students' warranting in some ways, especially with regards to $\omega 2$ warrants. But it is precisely these sort of validating checks which Alcock and Weber (2005) say are important. This means because no arguments were included of the $\omega 2$ variety, in future instance of the assessment, it might be an important addendum to include some sort of question that probes at student's ability to identify warrants which justify claims being made.

Dual measure of validity. Brent's response and need for cues in the form of options on the level 2 and 3 questions poses the reality that the assessment in this study is

in fact looking at validity from two distinct angles. The first measure, which appears in the level 1 questions (see Figure 14), is of students' ability to validate in a very pure and simple way, "is the argument valid or not?" It is a traditional validity judgment. But, from an assessment and deeper cognitive point of view, this first question answers little about what criteria students are using to make said judgement and if those criteria are valid in the context of the ITP setting. This lack of criteria for validity meant a second measure was needed to better understand students' true ability. This second measure appears in the level 2 and 3 questions and is a measure of students' ability to identify or recognize validity issues when presented with such options, in effect forcing student to contextualize their initial judgments. In this way, the assessment is always measuring validating ability, but in different ways.

In future iterations of this instrument, the scoring of the assessment needs to reflect this dual nature. One possible way to accomplish this would be to give individual scoring for the two judgment types, and then include a composite score. By breaking apart the scoring in this way, the assessment would give more information about what strengths a student has in validating arguments. If this scoring system was put into place, a means for overcoming the imbalance of a 50-50 decision (i.e., valid or invalid) and 3:1 odds on the latter questions (i.e., which one of these four invalidates this argument?) would need to be derived to increase the true meaning of the scores.

Leading students to correct judgements. Selden and Selden (2003) posited that students could be led to validate arguments through prompting and questioning, and in a like manner, this assessment is accomplishing a similar feat by utilizing the level 2 and level 3 questions. While the analysis undertaken to determine the reliability of this

assessment does not clearly answer, on a larger scale, whether students can be led to make correct validation judgments, it does hint at a deeper understanding. Of note, for 7 of the 9 invalid arguments at least 10% of the students changed from an incorrect validation to a correct one through the course of each testlet, with three testlets having more than 20% making the change (see Table 29). Future efforts are needed to determine what this means, and if there is significance in these findings. At the very least, it seems to indicate that at least some of the students utilized the distractors and keys as a means to further analyze each of the arguments and did so with a degree of success.

Validating through comprehension. During the assessment, there were students like Shannon, trying to understand the arguments and at times using comprehension in place of a more appropriate action for validating. In this study, I argue that although they were not necessarily doing what I, as a researcher would term validating, they were doing what they knew how to do in order to determine if an argument was valid or not. While this is imperfect, it is not unexpected as validity and proof validation are not overt parts of the curriculum. Moreover, this understanding corroborates Selden and Selden's (2003) findings that making sense was sometimes an important criterion for student in making these sorts of judgements.

In the open pilot of this study, students often made comments that suggested they were using their understanding as a means of validating, often suggesting that if they did not understand the argument, it was because it was above them, and was probably valid. On the other hand, during the interviews group 1 and 2, students used a lack of comprehension as a que that something might be wrong with the argument and in the

case of the group 3 students, this often was sufficient to invalidate the argument. This says something considerable as the average score on the assessment was 0.5.

All three groups of students from the interviews saw differing levels of success. Each attempted to validate the arguments in their own way in most every case. There were moments where we see some of the students struggled during the interview with comprehending portions of the arguments – though most often *local comprehension* (see Mejía-Ramos et al., 2012) was the issue – but at no point were any of the explanations wholly focused on understanding the argument as a means of explaining their validating process. Moreover, the students which scored higher on the assessment were active in a process of validating that included a mature sense of the relationship between the proposition and the argument, whereas those who scored lower did not. From an assessment validity standpoint, the assessment from this study genuinely measured students' ability to validate arguments.

In the end, the questions remain: how many students used *struggling to understand* as a placeholder for validating? How many students in our mathematics departments are in the same situation, where they are struggling to understand proof from day to day, never getting to the point of recognizing arguments as proofs? While neither of these questions have good answers, it does point to the need for better instruction concerning proof in terms of validity, as well as future studies to probe at students' self-narratives throughout their validating attempts.

Other limiting factors. I feel the results from the reliability analysis suggest that more work is required to have a working tool. At this point, the product of this study is an assessment with considerable potential, but which requires further refinement. As was

stated in the results section, each form was well below what is required for accurate parameter estimation in an IRT-based test development (Şahin & Anil, 2016). IRT was the process selected for analysis of this small pilot, because it is the process I plan to use to analyze a larger scale trial.

Before such a trial can be performed, it is important to make adjustments not only to the testlets as needed, but to select the best set of testlets as possible. One unforeseen issue which arose in the piloting of the assessment was the effect that having two issues in a single argument had on validating. As students, and even as mathematicians, once one finds a validity issue, they typically do not go looking for another. Once students saw one issue existed in testlet T2 and T7, they were likely disincentivized to list a second issue had occurred. While this did not seem to be as much the case for T2, it certainly seemed to cause issues with T7. I hoped the structure of the assessment, and the way I chose to grade, would help overcome this tendency, but it didn't seem to work with T7. Therefore, before including this item in any further versions, this issue needs to be considered and rectified, whether by adjusting the scoring or changing the argument.

The future. While the assessment is imperfect and requires some refinement, in its current state, it is sufficiently strong enough to point to a deficiency in students' validating abilities. Afterall, I complied eleven arguments with more than 90% agreeance on their validity and the students only identified slightly more than 50% of these arguments' validity correctly. From a teaching standpoint, this means instructors in all university proof-based courses would do well to overtly include ideas like the CVI framework into their courses. It is important students know how to identify genuine proofs from faulty arguments, and, more importantly, use that knowledge to analyze their

own proofs as they create them and after they have written them. If we accept that comprehension and validation are in fact related, then as students learn this, in the classroom will more readily comprehend the arguments their teachers present. This is an important step, as fighting to comprehend an argument within the classroom might limit a students' time to actually gain conviction in the proofs they are presented in class (see Mejía-Ramos & Tall, 2005; Segal 2000; Raman 2002; Weber & Mejía-Ramos, 2015).

Some notational issues existed where a few students struggled with expression like $18|x$, and some of the variations could be attributed to these issues. Of the 6 students that were interviewed, this was the only notational or linguistic issue students had. While the assessment did include some difficult or obscure mathematical ideas novel to the ITP setting, like equivalence classes and symmetric differences, in almost all these cases, clear definitions were included in the assessment, which students reported, they understood well enough from the included definitions.

The assessment needs to undergo some refinement and another process of testing. The next round will include a more simplified assessment with fewer arguments, but still employ the testlet structure. I felt the time required to complete the assessment, between 20 and 45 minutes, was acceptable to most students. If another anchored version of the assessment is run, the aim is to include only two forms, where seven of the eight items are anchored and only two items are interchanged. After this trial, an additional set of interviews will be held and done so in greater number, and face to face when possible.

References

Alcock, L., Bailey, T., Inglis, M., & Docherty, P. (2014). The ability to reject invalid logical inferences predicts proof comprehension and mathematics performance. In *Proceedings of the 17th Conference on Research in Undergraduate Mathematics Education*, Denver, CO: SIGMAA on RUME.

Alcock, L., & Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *The Journal of Mathematical Behavior*, *24*(2), 125-134.

Baker, F. B. (2001). *The Basics of item response theory, second edition.* College Park, Maryland: ERIC Clearinghouse on Assessment and Evaluation.

Balacheff, N. (1988). Aspects of proof in pupils' practice of school mathematics. In D. Pimm (Ed.), *Mathematics, teachers and children* (pp. 216-230). London: Hodder & Stoughton.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction, 28*(2), 113–145. doi:10.1080/07370001003676587

Chartrand, G., Polimeni, A. D., & Zhang, P. (2008). *Mathematical Proofs: A transition to advanced mathematics*. Boston.

Conradie, J., & Frith, J. (2000). Comprehension tests in mathematics. *Educational Studies in Mathematics*, *42*(3), 225-235.

Corbett, E. P., & Connors, R. J. (1999). A survey of rhetoric. *Classical rhetoric for the modern student*, *3*, 539-578.

David and Zazkis (2017). Characterizing the nature of introduction to proof courses: A survey of R1 and R2 institutions across the US. In *Proceedings of the 20th Conference on Research in Undergraduate Mathematics Education*, San Diego, CA: SIGMAA on RUME

Davis, P. J., & Hersh, R. (1981). *The mathematical experience*. New York: Viking Penguin.

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. D. (1990). The role and function of proof in mathematics. *Pythagoras*, 24, 7

Dreyfus, T. (1991). Advanced mathematical thinking processes. In D. Tall (Ed.), *Advanced mathematical thinking* (pp. 25-41). Dordrecht, The Netherlands: Kluwer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fagan, J. B., & Melhuish, K. (2018). Proof norms in introduction to proof textbooks. In *Proceedings of the 21st Conference on Research in Undergraduate Mathematics Education*, San Diego, CA: SIGMAA on RUME.

Fischbein, E. (1983). Intuition and Analytical Thinking in Mathematics Education. *International Reviews on Mathematical Education*, *15*(2), 68-74.

Mallery, P., & George, D. (2003). SPSS for Windows step by step: a simple guide and reference. *Allyn, Bacon, Boston.*

Halliday, M. (1978). Language as social semiotic: The social interpretation of language and meaning. Baltimore, MD: University Press

Hammack, R. H. (2013). *Book of proof*. Richard Hammack.

Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange*, *21*(1), 6-13

Hanna, G. (2000). Proof, explanation and exploration: An overview. *Educational studies in mathematics*, *44*(1), 5-23.

Harel, G., & Sowder, L. (1998). Students' proof schemes. *Research on Collegiate Mathematics Education*, Vol. III. In E. Dubinsky, A. Schoenfeld, & J. Kaput (Eds.), AMS, 234-283.

Harel, G., & Sowder, L (2007). Toward a comprehensive perspective on proof, In F. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning*, National Council of Teachers of Mathematics.

Hazzan, O., & Leron, U. (1996). Students' use and misuse of mathematical theorems: The case of Lagrange's theorem. *For the Learning of Mathematics*, *16*(1), 23-26.

Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. Journal for Research in Mathematics Education, 31, 396-428.

Heinze, A. (2010). Mathematicians' individual criteria for accepting theorems as proofs: An empirical approach. In G. Hanna, H. N. Jahnke, & H. Pulte (Eds.), *Explanation and proof in mathematics: Philosophical and educational perspectives* (pp. 101–111). New York: Springer.

Herbst, P. G. (2002). Engaging students in proving: A double bind on the teacher. *Journal for Research in Mathematics Education*, 33(3), 176-203.

Hersh, R. (1993). Proving is convincing and explaining. *Educational Studies in Mathematics, 24*, 389-399.

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, *30*(3), 159–166. doi:10.1119/1.2343498

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141.

Hill, H., Ball, D., & Schilling, S. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

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*(2), 270-282.

Knuth, E. J. (2002). School mathematics teachers' conceptions of proof. *Journal for Research in Mathematics Education, 33*(5) 379-405.

Ko, Y. Y., & Knuth, E. J. (2013). Validating proofs and counterexamples across content domains: Practices of importance for mathematics majors. *The Journal of Mathematical Behavior, 32*(1), 20-35.

Lai, Y., & Weber, K. (2014). Factors mathematicians profess to consider when presenting pedagogical proofs. *Educational Studies in Mathematics*, *85*(1), 93-108.

Lai, Y., Weber, K., & Mejía-Ramos, J. P. (2012). Mathematicians' perspectives on features of a good pedagogical proof. *Cognition and Instruction*, *30*(2), 146-169.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25-47.

Lew, K., Fukawa-Connelly, T. P., Mejia-Ramos, J. P., & Weber, K. (2016). Lectures in advanced mathematics: Why students might not understand what the mathematics professor is trying to convey. *Journal for Research in Mathematics Education*, *47*(2), 162-198.

Lindell, R. S., Peak, E., & Foster, T. M. (2007). Are they all created equal? A comparison of different concept inventory development methodologies. *AIP Conference Proceedings*, *883*(1), 14–17. doi:10.1063/1.2508680

Mariotti, M. A. (2000). Introduction to proof: The mediation of a dynamic software environment. *Educational studies in mathematics*, *44*(1), 25-53.

Martin, G.W. and Harel, G. (1989). Proof frames of preservice elementary teachers. *Journal for Research in Mathematics Education, 20,* 41–51.

Mejia-Ramos, J. P., & Inglis, M. (2009). Argumentative and proving activities in mathematics education research. In F.-L. Lin, F.-J. Hsieh, G. Hanna, & M. de Villiers (Eds.), Proceedings of the ICMI Study 19 conference: Proof and Proving in Mathematics Education (Vol. 2, pp. 88–93). Taipei, Taiwan.

Mejia-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, *79*(1), 3-18.

Mejía-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (in press). Developing and validating proof comprehension tests in undergraduate mathematics. To appear in *Research in Mathematics Education*.

Mejia-Ramos, J. P., & Weber, K. (2014). Why and how mathematicians read proofs: Further evidence from a survey study. *Educational Studies in Mathematics*, *85*(2), 161-173.

Melhuish, K. M. (2015). *The Design and Validation of a Group Theory Concept Inventory* (Doctoral dissertation).

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3-62.

Miyazaki, M., Fujita, T., & Jones, K. (2017). Students' understanding of the structure of deductive proof. *Educational Studies in Mathematics*, *94*(2), 223-239.

Morgan D. L. (1988). *Focus groups as qualitative research*. London: Sage.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in mathematics*, *27*(3), 249-266.

Moore, R. C. (2016). Mathematics Professors' Evaluation of Students' Proofs: A Complex Teaching Practice. *International Journal of Research in Undergraduate Mathematics Education, 2*(2), 246-278.

Morris, A. K. (2007). Factors affecting pre-service teachers' evaluations of the validity of students' mathematical arguments in classroom contexts. *Cognition and Instruction, 25*(4), 479-522.

Powers, R. A., Craviotto, C., & Grassl, R. M. (2010). Impact of proof validation on proof writing in abstract algebra. *International Journal of Mathematical Education in Science and Technology, 41*(4), 501-514.

Pedemonte, B. (2007). How can the relationship between argumentation and proof be analyzed? *Educational Studies in Mathematics, 66*(1), 23-42.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press, 2102 Constitutions Avenue, NW, Lockbox 285, Washington, DC 20055.

Raman, M. (2002). Coordinating informal and formal aspects of mathematics: Student behavior and textbook messages. *The Journal of Mathematical Behavior*, *21*(2), 135-150.

Rav, Y. (1999). Why do we prove theorems?. *Philosophia mathematica*, *7*(1), 5-41.

Rowland, T. (2002). Generic proofs in number theory. In S. R. Campbell & R. Zazkis (Eds.), *Learning and teaching number theory: Research in cognition and instruction* (pp. 157-183). Westport, CT: Ablex Publishing.

Şahin, A., & Anil, D. (2016). The Effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice, 17*(1n), 321-335.

Samkoff, A., Lai, Y., & Weber, K. (2012). On the different ways that mathematicians use diagrams in proof construction. *Research in Mathematics Education*, *14*(1), 49-67.

Segal, J. (1999). Learning about mathematical proof: Conviction and validity. *The Journal of Mathematical Behavior*, *18*(2), 191-210.

Selden, A., & Selden, J. (1987). Errors and misconceptions in college level theorem proving. In *Proceedings of the second international seminar on misconceptions and educational strategies in science and mathematics* (Vol. 3, pp. 457-470). Cornell University New York.

Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics*, *29*(2), 123-151.

Selden, A., & Selden, J. (2003). Validations of proofs written as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education, 36*, 4-36.

Smith, D., Eggen, M., & Andre, R. S. (2014). *A transition to advanced mathematics*. Nelson Education.

Smithson, J. (2000). Using and analyzing focus groups: limitations and possibilities. *International journal of social research methodology*, *3*(2), 103-119.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of personality assessment*, *80*(3), 217-222.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education, 38*, 289-321.

Stylianides, A. J., & Stylianides, G. J. (2009). Proof constructions and evaluations. *Educational Studies in Mathematics, 72*(2), 237–253.

Stylianides, G. J., Stylianides, A. J., & Weber, K. (2017) Research on the teaching and learning of proof: Taking stock and moving forward. In J. Cai (Eds.), *Compendium for Research in Mathematics Education. National Council of Teachers of Mathematics*: Reston, VA.

Thompson, D. R. (2014). Reasoning-and-proving in the written curriculum: Lessons and implications for teachers, curriculum designers, and researchers. *International Journal of Educational Research, 64*, 141-148.

Toulmin, S. (1964) *The uses of argument*, Cambridge, UK, Cambridge University Press.

Vogt, W. P. (2007). *Quantitative research methods for professionals*. Boston, MA, Pearson.

Weber, K., & Mejia-Ramos, J. P. (2011). Why and how mathematicians read proofs: An exploratory study. *Educational Studies in Mathematics, 76*(3), 329-344.

Weber, K., & Mejia-Ramos, J. P. (2014). Mathematics majors' beliefs about proof reading. *International Journal of Mathematical Education in Science and Technology, 45*(1), 89-103.

Weber, K., & Mejia-Ramos, J. P. (2015). On relative and absolute conviction in mathematics. *For the Learning of Mathematics, 35*(2), 15-21.

Weber, K., Inglis, M., & Mejia-Ramos, J. P. (2014). How mathematicians obtain conviction: Implications for mathematics instruction and research on epistemic cognition. *Educational Psychologist, 49*(1), 36-58.

Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics, 48*(1), 101–119.

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. *The Journal of Mathematical Behavior, 23*(2), 115–133.

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education, 39,* 431–459.

Weber, K. (2010). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning, 12*(4), 306–336.

Weber, K. (2015). Effective proof reading strategies for comprehending mathematical proofs. *International Journal of Research in Undergraduate Mathematics Education*, *1*(3), 289-314.

Weber, K., & Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics, 56*(2–3), 209–234.

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

Wilkinson, S. (1998). Focus group methodology: a review. *International journal of social research methodology*, *1*(3), 181-203.